Trends in Mathematics
Research Perspectives

Uwe Kähler
Michael Reissig
Irene Sabadini
Jasson Vindas
Editors

Analysis, Applications, and Computations

Proceedings of the 13th ISAAC Congress, Ghent, Belgium, 2021

Birkhäuser

Trends in Mathematics

Research Perspectives

Research Perspectives collects core ideas and developments discussed at conferences and workshops in mathematics, as well as their increasingly important applications to other fields. This subseries' rapid publication of extended abstracts, open problems and results of discussions ensures that readers are at the forefront of current research developments.

Uwe Kähler • Michael Reissig • Irene Sabadini • Jasson Vindas
Editors

Analysis, Applications, and Computations

Proceedings of the 13th ISAAC Congress, Ghent, Belgium, 2021

Birkhäuser

Editors
Uwe Kähler
Department of Mathematics
Universidade de Aveiro
Aveiro, Portugal

Michael Reissig
Institut für Angewandte Analysis
TU Bergakademie Frieberg
Freiberg, Germany

Irene Sabadini
Dipartimento di Matematica
Politecnico di Milano
Milano, Italy

Jasson Vindas
Department of Mathematics
Ghent University
Ghent, Belgium

ISSN 2297-0215 ISSN 2297-024X (electronic)
Trends in Mathematics
ISSN 2509-7407 ISSN 2509-7415 (electronic)
Research Perspectives
ISBN 978-3-031-36374-0 ISBN 978-3-031-36375-7 (eBook)
https://doi.org/10.1007/978-3-031-36375-7

This book is published under the imprint Birkhäuser, www.birkhauser-science.com by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

Preface

This volume contains selected contributions by participants of the 13th International ISAAC Congress, which was organized at Ghent University, Belgium, and was held from August 2 to August 6, 2021. The ISAAC congress series is the main bi-annual event of the International Society for Analysis, its Applications and Computation. This edition continued the successful series of meetings previously held in: Delaware (USA, 1997), Fukuoka (Japan, 1999), Berlin (Germany, 2001), Toronto (Canada, 2003), Catania (Italy, 2005), Ankara (Turkey, 2007), London (UK, 2009), Moscow (Russia, 2011), Krakow (Poland 2013), Macau (China, 2015), Växjö (Sweden, 2017), and Aveiro (Portugal, 2019).

The 13th ISAAC Congress was an important scientific event that promoted communication of mathematical advances in mathematical analysis, its applications, and its interactions with computation, encouraging further research progress. Mathematicians from different parts of the world had the opportunity to present their results and new ideas. In total, there were 659 participants from all continents who registered to take part in the conference. There were 374 talks, consisting of 6 plenary lectures and 368 contributed talks, contributing to the 16 congress sessions, some of which were organized by the special interest groups of the society. Following a well-established tradition within society, an award is presented to one or various outstanding young mathematicians. The ISAAC award of the 13th ISAAC Congress was given to *Guido De Philippis* (New York University, USA) for his major contributions to calculus of variations, partial differential equations, and geometric measure theory.

The following sessions contributed to the present volume. The volume also features an article by S. Jaffard et al., originating from Jaffard's plenary lecture *Multivariate Multifractal Analysis* delivered at the congress.

- *Applications of dynamical systems theory in biology*, organized by Torsten Lindström.
- *Challenges in STEM education*, organized by Ján Guncaga and Vladimir Mityushev.

- *Complex analysis and partial differential equations*, organized by Sergei Rogosin, Ahmet Okay Celebi, and Carmen Judith Vanegas.
- *Complex variables and potential theory*, organized by Tahir Aliyev Azeroglu, Massimo Lanza de Cristoforis, Anatoly Golberg, and Sergiy Plaksa.
- *Constructive methods in the theory of composite and porous media*, organized by Vladimir Mityushev, Natalia Rylko, and Piotr Drygaś.
- *Generalized functions and applications*, organized by Michael Kunzinger and Marko Nedeljkov.
- *Harmonic analysis and partial differential equations*, organized by Vladimir Georgiev, Michael Ruzhansky, and Jens Wirth.
- *Partial differential equations on curved spacetimes*, organized by Anahit Galstyan, Makoto Nakamura, and Karen Yagdjian.
- *Recent progress in evolution equations*, organized by Marcello D'Abbicco and Marcelo Rempel Ebert.
- *Wavelet theory and its related topics*, organized by Keiko Fujita and Akira Morimoto.

We would like to thank the organizers of all sessions of the congress for their invaluable work and efforts. They very much supported the congress organization by inviting participants, planning their sessions, and selecting speakers. During the congress itself, they did an excellent job organizing the chairing of their meetings. The session organizers were also responsible for the refereeing process of the contributions to this proceedings volume.

The ISAAC board and the participants of the congress thank Jasson Vindas and his group for the excellent organization of the 13th ISAAC Congress.

Aveiro, Portugal — Uwe Kähler
Freiberg, Germany — Michael Reissig
Milan, Italy — Irene Sabadini
Ghent, Belgium — Jasson Vindas
November 2022

Contents

Part I
Plenary Lecture

A Review of Univariate and Multivariate Multifractal Analysis Illustrated by the Analysis of Marathon Runners Physiological Data

Stéphane Jaffard, Guillaume Saës, Wejdene Ben Nasr, Florent Palacin, and Véronique Billat

Abstract We review the central results concerning wavelet methods in multifractal analysis, which consists in analysis of the pointwise singularities of a signal, and we describe its recent extension to multivariate multifractal analysis, which deals with the joint analysis of several signals; we focus on the mathematical questions that this new techniques motivate. We illustrate these methods by an application to data recorded on marathon runners.

1 Introduction

Everywhere irregular signals are ubiquitous in nature: Classical examples are supplied by natural phenomena (hydrodynamic turbulence [89], geophysics, natural textures [76]), physiological data (medical imaging [12], heartbeat intervals [2], E.E.G [37]); they are also present in human activity and technology (finance [17], internet traffic [5], repartition of population [46, 104], text analysis [85], art [3]).

S. Jaffard (✉) · W. B. Nasr
Univ Paris Est Creteil, Univ Gustave Eiffel, CNRS, LAMA UMR8050, Creteil, France
e-mail: jaffard@u-pec.fr; wejdene.nasr@u-pec.fr

G. Saës
Univ Paris Est Creteil, Univ Gustave Eiffel, CNRS, LAMA UMR8050, Creteil, France

Département de Mathématique, Université de Mons, Mons, Belgium
e-mail: guillaume.saes@u-pec.fr

F. Palacin
Laboratoire de neurophysiologie et de biomécanique du mouvement, Institut des neurosciences de l'Université Libre de Bruxelles, Brussels, Belgium

V. Billat
Université Paris-Saclay, Univ Evry, Evry-Courcouronnes, France
e-mail: veronique.billat@billatraining.com

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_1

The analysis of such phenomena requires the modelling by everywhere irregular functions, and it is therefore natural to use mathematical regularity parameters in order to classify such data, and to study mathematical models which would fit their behavior. Constructing and understanding the properties of such functions has been a major challenge in mathematical analysis for a long time: Shortly after Cauchy gave the proper definition of a continuous function, the question of determining if a continuous function is necessarily differentiable at some points was a major issue for a large part of the nineteenth century; though a first counterexample was found by Bolzano, his construction remained unknown from the mathematical community, and it was only in 1872, with the famous Weierstrass functions

$$\mathcal{W}_{a,\omega}(x) = \sum_{n=0}^{+\infty} \frac{\sin(a^n x)}{a^{\omega n}} \qquad \text{for} \qquad a > 1 \qquad \text{and} \qquad \omega \in (0,1), \tag{1}$$

that the problem was settled. However, such constructions were considered as weird counterexamples, and not representative of what is commonly met, both in mathematics and in applications. In 1893, Charles Hermite wrote to Thomas Stieltjes: *I turn my back with fright and horror to this lamentable plague: continuous functions without derivative*. The first statement that smooth or piecewise smooth functions were not adequate for modelling natural phenomena but were rather exceptional came from physicists, see e.g. the introduction of the famous book of Jean Perrin "Les atomes", published in 1913. On the mathematical side, the evolution was slow: In 1931, Mazurkiewicz and Banach showed that most continuous functions are nowhere differentiable ("most" meaning here that such functions form a residual set in the sense of Baire categories). This spectacular result changed the perspective: Functions which were considered as exceptional and rather pathological actually were the common rule, and smooth functions turn out to be exceptional.

A first purpose of multifractal analysis is to supply mathematical notions which allow to quantify the irregularity of functions, and therefore yield quantitative tools that can be applied to real life data in order to determine if they fit a given model, and, if it is the case, to determine the correct parameters of the model. One can also be more ambitious and wonder which "types" on singularities are present in the data, which may yield an important information of the nature of the signal; a typical example is supplied by *chirps* which are singularities which behave like

$$g(x) = |x - x_0|^{\alpha} \cos\left(\frac{1}{|x - x_0|^{\beta}}\right), \tag{2}$$

displaying fast oscillations near the singularity at x_0. Such singularities are e.g. predicted by some models of turbulence and therefore determining if they can be found in the recorded data in wind tunnels is an important issue in the understanding of the physical nature of turbulence.

A first step in this program was performed by A. Kolmogorov in 1941 [80]. Let $f : \mathbb{R}^d \to \mathbb{R}$. The *Kolmogorov scaling function* of f is the function $\eta_f(p)$

implicitly defined by

$$\forall p > 0, \qquad \int |f(x+h) - f(x)|^p dx \quad \sim \quad |h|^{\eta_f(p)}, \tag{3}$$

the symbol $\sim$ meaning that

$$\eta_f(p) = \liminf_{|h|\to 0} \frac{\log\left(\int |f(x+h) - f(x)|^p dx\right)}{\log|h|}. \tag{4}$$

Note that, if f is smooth, then one has to use differences of order 2 or more in order to define correctly the scaling function. Kolmogorov proposed to use this tool as a way to determine if some simple stochastic processes are fitted to model the velocity of turbulent fluids at small scales, and a first success of this approach was that fractional Brownian motions (see Sect. 2.2) do not yield correct models (their scaling functions are linear, whereas the one measured on turbulent flows are significatively concave [9]).

An important interpretation of the Kolmogorov scaling function can be given in terms of *global smoothness* indices in families of functions spaces: the spaces $\mathrm{Lip}(s, L^p(\mathbb{R}^d))$ defined as follows. Let $s \in (0,1)$, and $p \in [1,\infty]$; $f \in \mathrm{Lip}(s, L^p(\mathbb{R}^d))$ if $f \in L^p(\mathbb{R}^d)$ and

$$\exists C > 0, \quad \forall h > 0, \qquad \int |f(x+h) - f(x)|^p dx \leq C|h|^{sp} \tag{5}$$

(here also, larger smoothness indices s are reached by replacing the first-order difference $|f(x+h) - f(x)|$ by higher order differences). It follows from (3) and (5) that,

$$\forall p \geq 1, \qquad \eta_f(p) = p \cdot \sup\{s : f \in \mathrm{Lip}(s, L^p(\mathbb{R}^d))\}. \tag{6}$$

An alternative formulation of the scaling function can be given in terms of global regularity indices supplied by Sobolev spaces, the definition of which we now recall.

Definition 1 Let $s \in \mathbb{R}$ and $p \geq 1$. A function f belongs to the Sobolev space $L^{p,s}(\mathbb{R}^d)$ if $(Id - \Delta)^{s/2} f \in L^p$, where $g = (Id - \Delta)^{s/2} f$ is defined through its Fourier transform as

$$\hat{g}(\xi) = (1 + |\xi|^2)^{s/2} \hat{f}(\xi).$$

This definition amounts to state that the fractional derivative of f of order s belongs to L^p. The classical embeddings between the Sobolev and the $\mathrm{Lip}(s, L^p)$

spaces imply that

$$\forall p \geq 1, \qquad \eta_f(p) = p \cdot \sup\{s : f \in L^{p,s}(\mathbb{R}^d)\}. \tag{7}$$

In other words, the scaling function tells, for each p, the order of (fractional) derivation of f up to which $f^{(s)}$ belongs to L^p.

A limitation of the use of the Kolmogorov scaling function for classification purposes is that many models display almost identical scaling functions (a typical example is supplied by the velocity of fully developed turbulence, see e.g. [82, 98]); the next challenge therefore is to construct alternative scaling functions which would allow to draw distinctions between such models. A major advance in this direction was reached in 1985 when Uriel Frisch and Giorgio Parisi proposed another interpretation of the scaling function in terms of *pointwise singularities* of the data [99]. In order to state their assertion, we first need the recall the most commonly used notion of pointwise regularity.

Definition 2 Let $f : \mathbb{R}^d \to \mathbb{R}$ be a locally bounded function, $x_0 \in \mathbb{R}^d$ and let $\gamma \geq 0$; f belongs to $C^\gamma(x_0)$ if there exist $C > 0$, $R > 0$ and a polynomial P of degree less than γ such that

$$\text{if } |x - x_0| \leq R, \quad \text{then} \qquad |f(x) - P(x - x_0)| \leq C|x - x_0|^\gamma.$$

The Hölder exponent of f at x_0 is

$$h_f(x_0) = \sup\left\{\gamma : \quad f \quad \text{is} \quad C^\gamma(x_0)\right\}. \tag{8}$$

Some functions have a very simple Hölder exponent. For instance, the Hölder exponent of the Weierstrass functions $\mathcal{W}_{a,\omega}$ is constant and equal to ω at every point (such functions are referred to as *monohölder functions*); since $\omega < 1$ we thus recover the fact that $\mathcal{W}_{a,\omega}$ is nowhere differentiable. However, the Hölder exponent of other functions turn out to be extremely irregular, and U. Frisch and G. Parisi introduced the *multifractal spectrum* $\mathcal{D}_f$ as a new quantity which allows to quantify some of its properties: $\mathcal{D}_f(H)$ denotes the fractional dimension of the *isoregularity sets*, i.e. the sets

$$\{x : \quad h_f(x) = H\}. \tag{9}$$

Based on statistical physics arguments, they proposed the following relationship between the scaling function and $\mathcal{D}_f(H)$:

$$\mathcal{D}_f(H) = \inf_p \left(d + Hp - \eta_f(p)\right), \tag{10}$$

which is referred to as the *multifractal formalism*, see [99] (we will discuss in Sect. 2.1 the "right" notion of fractional dimension needed here). Though the remarkable intuition which lies behind this formula proved extremely fruitful, it

needs to be improved in order to be completely effective; indeed many natural processes used in signal or image modelling do not follow this formula if one tries to extend it to negative values of p, see [81]; additionally, the only mathematical result relating the spectrum of singularities and the Kolmogorov scaling function in all generality is very partial, see [55, 60]. In Sect. 2.2 we will discuss (10), and see how it needs to be reformulated in terms of wavelet expansions in order to reach a fairly general level of validity. In Sect. 2.3 we will discuss the relevance of the Hölder exponent (8) and introduce alternative exponents which are better fitted to the analysis of large classes of real-life data. Their characterization requires the introduction of orthonormal wavelet bases. This tool and its relevance for global regularity is recalled in Sect. 2.4 and the characterizations of pointwise regularity which they allow are performed in Sect. 2.5. This leads to a classification of pointwise singularities which yields a precise description of the oscillations of the function in the neighbourhood of its singularities which is developed in Sect. 2.6. This implications of this classification on the different formulations of the multifractal formalism are developed in Sect. 2.7. The tools thus developed are applied to marathon runners physiological data (heart rate, acceleration, cadence, i.e. number of steps per minute) in Sect. 2.9; thus showing that they lead to a sharper analysis of the physiological modifications during the race. The numerical results derived on real-life data have been obtained using the Wavelet p-Leader and Bootstrap based MultiFractal analysis (PLBMF) toolbox available on-line at https://www.irit.fr/~Herwig.Wendt/software.html.

The explosion of data sciences recently made available collections of signals the singularities of which are expected to be related in some way; typical examples are supplied by EEG collected at different areas of the brain, or by collections of stock exchange prizes. The purpose of Sect. 3 is to address the extension of multifractal analysis to the multivariate setting, i.e. to several functions. In such situations, a pointwise regularity exponent $h_i(x)$ is associated with each signal $f_i(x)$ and the challenge is to recover the *joint multivariate spectrum* of the f_i which is defined as the fractional dimension of the sets of points x where each of the exponents $h_i(x)$ takes a given value: If m signals are available, we define

$$E_{f_1,\dots,f_m}(H_1,\dots,H_m) = \{x : \quad h_1(x) = H_1,\dots,h_m(x) = H_m\}, \tag{11}$$

and the *joint multifractal spectrum* is

$$D_{f_1,\dots,f_m}(H_1,\dots,H_m) = \dim(E_{f_1,\dots,f_m}(H_1,\dots,H_m)). \tag{12}$$

These notions were introduced by C. Meneveau et al. in the seminal paper [95] which addressed the joint analysis of the dissipation rate of kinetic energy and passive scalar fluctuations for fully developed turbulence, and a general abstract setting was proposed by J. Peyrière in [100]; In Sect. 3.1, we introduce the mathematical concepts which are relevant to this study. In Sect. 3.2 we give a probabilistic interpretation of the scaling functions introduced in Sect. 2, and we show how they naturally lead to a 2-variable extension in terms of correlations.

The initial formulation of the multifractal formalisms based on extensions of the Kolmogorov scaling function suffers from the same drawbacks as in the univariate case. This leads naturally to a reformulation of the multifractal formalism which is examined in Sect. 3.3, where we also investigate the additional advantages supplied by multivariate multifractal analysis for singularity classifications. In order to investigate its relevance, we study a toy-example which is supplied by Brownian motions in multifractal time in Sect. 3.4. In Sect. 3.5, we illustrate the mathematical results thus collected by applications to the joint analysis of heartbeat, cadence and acceleration of marathon runners.

2 Univariate Multifractal Analysis

2.1 The Multifractal Spectrum

In order to illustrate the motivations of multifractal analysis, let us come back to the initial problem we mentioned: How badly can a continuous function behave? We mentioned the surprising result of Mazurkiewicz and Banach stating that a generic continuous function is nowhere differentiable, and the Weierstrass functions yield examples of continuous functions which may have an arbitrarily small (and constant) Hölder exponent. This can actually be improved: A generic continuous function satisfies

$$\forall x \in \mathbb{R}, \qquad h_f(x) = 0, \tag{13}$$

see [115]: At every point the Hölder exponent of f is as bad as possible. An example of such a continuous function is supplied by a slight variant of Weierstrass functions:

$$f(x) = \sum_{j=1}^{\infty} \frac{1}{j^2} \sin(2^j x).$$

Let us now consider a different functional setting: Let $f : [0, 1] \longrightarrow [0, 1]$ be an increasing function. At any given point $x \in [0, 1]$ f can have a discontinuity at x, in which case $h_f(x) = 0$. Nonetheless, this worse possible behavior cannot be met everywhere: An important theorem of Lebesgue states that f is almost everywhere differentiable and therefore satisfies

$$\text{for almost every } x \in [0, 1], \qquad h_f(x) \geq 1.$$

The global regularity assumption (the fact that f is increasing implies that its derivative in the sense of distributions is a bounded Radon measure) implies that, in sharp contradistinction with generic continuous functions, the set of points such that $h_f(x) < 1$ is "small" (its Lebesgue measure vanishes). On other hand, the

set of points where it is discontinuous can be an arbitrary countable set (but one easily checks that it cannot be larger). What can we say about the size of the sets of points with intermediate regularity (i.e. having Hölder exponents between 0 and 1), beyond the fact that they have a vanishing Lebesgue measure? Answering this problem requires to use some appropriate notion of "size" which allows to draw differences between sets of vanishing Lebesgue measure. The right mathematical notion fitted to this problem can be guessed using the following argument. Let

$$E_f^\alpha = \{x : f \notin C^\alpha(x)\}.$$

Clearly, if $x \in E_f^\alpha$, then there exists a sequence of dyadic intervals

$$\lambda_{j,k} = \left[\frac{k}{2^j}, \frac{k+1}{2^j}\right] \tag{14}$$

such that

- x belongs either to $\lambda_{j,k}$ or to one of its two closest neighbours of the same width,
- the increment of f on $\lambda_{j,k}$ is larger than $2^{-\alpha j} = |\lambda_{j,k}|^\alpha$ (where $|A|$ stands for the diameter of the set A).

Let $\varepsilon > 0$, and consider the *maximal* dyadic intervals of this type of width less than $\varepsilon/3$, for all possible $x \in E_f^\alpha$, and denote this set by $\Lambda_\alpha^\varepsilon$. These intervals are disjoint (indeed two dyadic intervals are either disjoint or one is included in the other); and, since f is increasing, the increment of f on $[0, 1]$ is bounded by the sum of the increments on these intervals. Therefore

$$\sum_{\lambda \in \Lambda_\alpha^\varepsilon} |\lambda|^\alpha \leq f(1) - f(0).$$

The intervals 3λ (which consists in the dyadic interval λ and its two closest neighbours of the same length) for $\lambda \in \Lambda_\alpha^\varepsilon$ form an *ε-covering* of E_f^α (i.e. a covering by intervals of length at most ε), and this ε-covering satisfies

$$\sum_{\lambda \in \Lambda_\alpha^\varepsilon} |3\lambda|^\alpha = 3^\alpha \sum_{\lambda \in \Lambda_\alpha^\varepsilon} |\lambda|^\alpha \leq 3^\alpha (f(1) - f(0)).$$

This property can be interpreted as stating that the *α-dimensional Hausdorff measure* of E_f^α is finite; we now give a precise definition of this notion.

Definition 3 Let A be a subset of $\mathbb{R}^d$. If $\varepsilon > 0$ and $\delta \in [0, d]$, let

$$M_\varepsilon^\delta = \inf_R \left(\sum_i |A_i|^\delta\right),$$

where R is an ε-covering of A, i.e. a covering of A by bounded sets $\{A_i\}_{i\in\mathbb{N}}$ of diameters $|A_i| \leq \varepsilon$ (the infimum is therefore taken on all ε-coverings). For any $\delta \in [0, d]$, the δ-dimensional Hausdorff measure of A is

$$mes_\delta(A) = \lim_{\varepsilon\to 0} M_\varepsilon^\delta.$$

One can show that there exists $\delta_0 \in [0, d]$ such that

$$\begin{cases} \forall \delta < \delta_0, & mes_\delta(A) = +\infty \\ \forall \delta > \delta_0, & mes_\delta(A) = 0. \end{cases}$$

This critical δ_0 is called the *Hausdorff dimension* of A, and is denoted by $\dim(A)$ (and an important convention is that, if A is empty, then $\dim(\emptyset) = -\infty$).

The example we just worked out shows that a global regularity information on a function yields information on the Hausdorff dimensions of its sets of Hölder singularities. This indicates that the Hausdorff dimension is the natural choice in (10), and motivates the following definition.

Definition 4 Let $f : \mathbb{R}^d \to \mathbb{R}$ be a locally bounded function. The multifractal Hölder spectrum of f is the function

$$\mathcal{D}_f(H) = \dim(\{x : \quad h_f(x) = H\}),$$

where dim denotes the Hausdorff dimension.

This definition justifies the denomination of *multifractal functions*: One typically considers functions f that have non-empty isoregularity sets (9) for H taking all values in an interval of positive length, and therefore one deals with an infinite number of fractal sets $E_f(H)$. The result we obtained thus implies that, if f is an increasing function, then

$$\mathcal{D}_f(H) \leq H. \tag{15}$$

This can be reformulated in a function space setting which puts in light the sharp contrast with (13): Indeed, recall that any function of bounded variation is the difference of an increasing and a decreasing function; we have thus obtained the following result.

Proposition 1 *Let $f : \mathbb{R} \to \mathbb{R}$ be a function of bounded variation. Then its multifractal spectrum satisfies*

$$\forall H, \qquad \mathcal{D}_f(H) \leq H.$$

Remark 1 This result does not extend to several variables functions of bounded variation which, in general, are not locally bounded, in which case their Hölder exponent is not even well defined.

2.2 Alternative Formulations of the Multifractal Formalism

We mentioned that (10) yields a poor estimate of the multifractal spectrum. A typical example is supplied by sample paths of *fractional Brownian motion* (referred to as fBm), a family of stochastic processes introduced by Kolmogorov [79], the importance of which was put in light for modeling by Mandelbrot and Van Ness [91]. This family is indexed by a parameter $\alpha \in (0, 1)$, and generalizes Brownian motion (which corresponds to the case $\alpha = 1/2$); fBm of index α is the only centered Gaussian random process B^α defined on $\mathbb{R}^+$ which satisfies

$$\forall x, y \geq 0 \qquad \mathbb{E}(|B^\alpha(x) - B^\alpha(y)|^2) = |x - y|^{2\alpha}.$$

FBm plays an important role in signal processing because it supplies the most simple one parameter family of stochastic processes with stationary increments. Its sample paths are monohölder and satisfy

$$\text{a.s. } \forall x, \qquad h_{B^\alpha}(x) = \alpha,$$

(see [77] and [41] for a recent sharp analysis of the pointwise regularity of their sample paths) so that their multifractal spectrum is

$$\text{a.s.} \quad \forall H, \qquad \begin{cases} \mathcal{D}_{B^\alpha}(H) = 1 & \text{if } H = \alpha \\ \qquad\quad = -\infty & \text{else.} \end{cases}$$

However, the right hand-side of (10) yields a different value for $H \in (\alpha, \alpha + 1]$: It coincides almost surely with the function defined by

$$\begin{cases} \mathcal{L}_{B^\alpha}(H) = \alpha + 1 - H & \text{if } H \in [\alpha, \alpha + 1] \\ \qquad\quad = -\infty & \text{else,} \end{cases}$$

see [5, 69, 70]. This is due to the fact that the decreasing part of the spectrum is recovered from negative values of p in (10), and the corresponding integral is not well defined for negative ps, and may even diverge. It follows that sharper estimates of the multifractal spectrum require a renormalization procedure which would yield a numerically robust output for negative ps. Several methods have been proposed to solve this deadlock. They are all based on a modification of the Kolmogorov scaling function in order to incorporate the underlying intuition that it should include some pointwise regularity information. A consequence will be that they provide an extension of the scaling function to negative ps. This extra range of parameters plays a crucial role in several applications where it is required for classifications, see e.g. [83, 98] where the validation of turbulence models is considered, and for which the key values of the scaling function which are needed to draw significative differences between these models are obtained for $p < 0$.

A first method is based on the *continuous wavelet transform*, which is defined as follows. Let ψ be a *wavelet*, i.e. a well localized, smooth function with, at least, one vanishing moment. The continuous wavelet transform of a one-variable function f is

$$C_{a,b}(f) = \frac{1}{a}\int_{\mathbb{R}} f(t)\psi\left(\frac{t-b}{a}\right)dt \qquad (a>0, \quad b\in\mathbb{R}); \tag{16}$$

Alain Arneodo, Emmanuel Bacry and Jean-François Muzy proposed to replace, in the integral (3), the increments $|f(x+\delta)-f(x)|$ at scale δ by the continuous wavelet transform $C_{a,b}(f)$ for $a=\delta$ and $b=x$. This choice follows the heuristic that the continuous wavelet transform satisfies $|C_{a,b}(f)| \sim a^{h_f(x)}$ when a is small enough and $|b-x| \sim a$. Note that it is not valid in all generality, but typically fails for *oscillating singularities*, such as the chirps (2). Nonetheless Yves Meyer showed that this heuristic actually characterizes another pointwise regularity exponent, the *weak scaling exponent*, see [97]. Assuming that the data do not include oscillating singularities, the integral (3) is discretized and replaced by the more meaningful values of the continuous wavelet transform i.e. at its local maxima [8]; if we denote by b_k the points where these extrema are reached at the scale a, the integral (3) is thus replaced by the sum

$$\sum_{b_k} |C_{a,b_k}(f)|^p \sim a^{\zeta_f(p)} \quad \text{when } a\to 0, \tag{17}$$

This reformulations using the *multiresolution quantities* $|C_{a,b_k}(f)|$ yields better numerical results than when using the increments $|f(x+\delta)-f(x)|$; above all, the restriction to the local suprema is a way to bypass the small values of the increments which were the cause of the divergence of the integral (3) when p is negative. Numerical experiments consistently show that the multifractal formalism based on these quantities yields the correct spectrum for the fBm, and also for large collections of mutifractal models, see [11].

Another way to obtain a numerically robust procedure in order to perform multifractal analysis is supplied by *Detrended Fluctuation Analysis* (DFA): From the definition of the Hölder exponent, Kantelhardt et al. [78] proposed the following multiresolution quantity based on the following local L^2 norms

$$T_{mfd}(a,k) = \left(\frac{1}{a}\sum_{i=1}^{a} |f(ak+i) - P_{k,a,N_P}(i)|^2\right)^{\frac{1}{2}}, k=1,\dots,n/a, \tag{18}$$

where n denotes the number of available samples and P_{t,a,N_P} is a polynomial of degree N_P obtained by local fit to f on portions of length proportional to a. The

integral (3) is now replaced by

$$S_{mfd}(a,q) = \frac{a}{n}\sum_{k}^{n/a} T_{mfd}(a,k)^q \sim a^{\zeta_{mfd}(q)},$$

and the multifractal spectrum is obtained as usual through a Legendre transform of this new scaling function ζ_{mfd}, thus yielding the *multifractal detrended fluctuation analysis* (MFDFA). Note that, here again, we cannot expect the multifractal formalism based on such a formula to be fitted to the Hölder exponent: The choice of an L^2 norm in (18) is rather adapted to an alternative pointwise exponent, the 2-exponent, which is defined through local L^2-norms, see Definition 5 (and [84] for an explanation of this interpretation). The MFDFA formalism performs satisfactorily and is commonly used in applications (cf., e.g., [50, 111]).

The methods we mentioned meet the following limitations: They cannot be taylored to a particular pointwise exponent: We saw that the WTMM is fitted to the weak-scaling exponent, and the MFDFA to the 2-exponent. They lack of theoretical foundation, and therefore the estimates that they yield on the multifractal spectrum are not backed by mathematical results. In practice, they are difficult to extend to data in two or more variables (for MFDFA, the computation of local best fit polynomials is an intricate issue). The obtention of an alternative formulation of the multifractal formalism which brings an answer to these two problems requires a detour through the notions of pointwise exponents, and their characterizations.

2.3 Pointwise Exponents

At this point we need to discuss the different notions of pointwise regularity. One of the reasons is that, though Hölder regularity is by far the one which is most used in mathematics and in applications, it suffers a major limitation: Definition 2 requires f to be locally bounded. In applications, this limitation makes the Hölder exponent unfitted in many settings where modelling data by locally bounded functions is inadequate; in Sect. 2.4 we will give a numerically simple criterium which allows to verify if this assumption is valid, and we will see that the physiological data we analyse are typical examples for which it is not satisfied. On the mathematical side too, this notion often is not relevant. A typical example is supplied by the Riemann series defined as

$$\forall x \in \mathbb{R}, \qquad \mathcal{R}_s(x) = \sum_{n=1}^{\infty} \frac{\sin(n^2 x)}{n^s}, \tag{19}$$

which, for $s > 1$, are locally bounded and turn out to be multifractal (in which case their multifractal analysis can be performed using the Hölder exponent [32, 54]),

but it is no more the case if $s < 1$, in which case an alternative analysis is developed in [108] (using the p-exponent for $p = 2$, see Definition 5 below).

There exist two ways to deal with such situations. The first one consists in first regularizing the data, and then analyzing the new data thus obtained. Mathematically, this means that a *fractional integral* is performed on the data. Recall that, if f is a tempered distribution defined on $\mathbb{R}$, then the fractional integral of order t of f, denoted by $f^{(-t)}$ is defined as follows: Let $(Id - \Delta)^{-t/2}$ be the convolution operator which amounts to multiplying the Fourier transform of f with $(1 + |\xi|^2)^{-t/2}$. The fractional integral of order t of f is the function

$$f^{(-t)} = (Id - \Delta)^{-t/2}(f).$$

If f is large enough, then $f^{(-t)}$ is a locally bounded function, and one can consider the Hölder exponent of t (the exact condition under which this is true is that t has to be larger than the exponent h_f^{min} defined below by (25) or equivalently by (26). This procedure presents the obvious disadvantage of not yielding a direct analysis of the data but of a smoothed version of them.

The other alternative available in order to characterize the pointwise regularity of non-locally bounded functions consists in using a weaker notion of pointwise regularity, the p-exponent, which we now recall. We define $B(x_0, r)$ as the ball of center x_0 and radius r.

Definition 5 Let $p \geq 1$ and assume that $f \in L^p_{loc}(\mathbb{R}^d)$. Let $\alpha \in \mathbb{R}$; f belongs to $T^p_\alpha(x_0)$ if there exists a constant C and a polynomial P_{x_0} of degree less than α such that, for r small enough,

$$\left(\frac{1}{r^d}\int_{B(x_0,r}|f(x) - P_{x_0}(x)|^p dx\right)^{1/p} \leq Cr^\alpha. \tag{20}$$

The p**-exponent** of f at x_0 is

$$h^p_f(x_0) = \sup\{\alpha : f \in T^p_\alpha(x_0)\} \tag{21}$$

(the case $p = +\infty$ corresponds to the Hölder exponent).

This definition was introduced by Calderón and Zygmund in 1961 in order to obtain pointwise regularity results for the solutions of elliptic PDEs, see [35]. For our concern, it has the important property of being well defined under the assumption that $f \in L^p_{loc}$. For instance, in the case of the Riemann series (19), an immediate computation yields that they belong to L^2 if $s > 1/2$ so that, if $1/2 < s < 1$, p-exponents with $p \leq 2$ are relevant to study their regularity, in contradistinction with the Hölder exponent which won't be defined. Another example of multifractal function which is not locally bounded is supplied by Brjuno's function, which plays an important role in holomorphic dynamical systems, see [92]. Though its is nowhere locally bounded, it belongs to all L^p

spaces and its multifractal analysis using p-exponents has been performed in [66]. Note that p-exponents can take values down to $-d/p$, see [73]. Therefore, they allow the use of *negative regularity exponents*, such as singularities of the form $f(x) = 1/|x - x_0|^\alpha$ for $\alpha < d/p$.

The general framework supplied by multifractal analysis now is ubiquitous in mathematical analysis and has been successfully used in a large variety of mathematical situations, using diverse notion of pointwise exponents such as pointwise regularity of probability measures [33], rates of convergence or divergence of series of functions (either trigonometric [13, 25] or wavelet [13, 65]) order of magnitude of ergodic averages [43, 44], to mention but a few.

2.4 Orthonormal Wavelet Decompositions

Methods based on the use of orthonormal wavelet bases follow the same motivations we previously developed, namely to construct alternative scaling functions based on multiresolution quantities which "incorporate" some pointwise regularity information. However, we will see that they allow to turn some of the limitations met by the previously listed methods, and they enjoy the following additional properties:

- numerical simplicity,
- explicit links with pointwise exponents (which, as we saw, may differ from the Hölder exponent),
- no need to construct local polynomial approximations (which is the case for DFA methods now in use),
- mathematical results hold concerning either the validity of the multifractal formalism supplied by (10) or of some appropriate extensions; such results can be valid for all functions, or for "generic" functions, in the sense of Baire categories, or for other notions of genericity.

Let us however mention an alternative technique which was proposed in [4] where multiresolution quantities based on local oscillations, such as

$$d_\lambda = \sup_{3\lambda} f(x) - \inf_{3\lambda} f(x),$$

or higher order differences such as

$$d_\lambda = \sup_{x,y\in 3\lambda} \left| f(x) + f(y) - 2f\left(\frac{x+y}{2}\right)\right|,$$

and which wouldn't present the third problem that we mention. However, as far as we know, they haven't been tested numerically.

One of the reasons for these remarkable properties is that (in contradistinction with other expansions, such as e.g. Fourier series) wavelet analysis allows to

characterize both global and pointwise regularity by simple conditions on the moduli of the wavelet coefficients; as already mentioned, the multifractal formalism raises the question of how global and pointwise regularity are interconnected; wavelet analysis therefore is a natural tool in order to investigate this question and this explains why it was at the origin of major advances in multifractal analysis both in theory and applications.

We now recall the definition of orthonormal wavelet bases. For the sake of notational simplicity, we assume in all the remaining of Sect. 2 that $d = 1$, i.e. the functions we consider are defined on $\mathbb{R}$, extensions in several variables being straightforward. Let $\varphi(x)$ denote a smooth function with fast decay, and good joint time-frequency localization, referred to as the *scaling function*, and let $\psi(x)$ denote an oscillating function (with N first vanishing moments), with fast decay, and good joint time-frequency localization, referred to as the *wavelet*. These functions can be chosen such that the

$$\varphi(x-k), \qquad \text{for,} \quad k \in \mathbb{Z} \tag{22}$$

and

$$2^{j/2}\psi(2^j x - k), \qquad \text{for,} \quad j \geq 0, \quad k \in \mathbb{Z} \tag{23}$$

form an orthonormal basis of $L^2(\mathbb{R})$ [96]. The wavelet coefficients of a function f are defined as

$$c_k = \int_{\mathbb{R}} f(x)\ \varphi(x-k)\,dx \qquad \text{and} \qquad c_{j,k} = 2^j \int_{\mathbb{R}} f(x)\ \psi(2^j x - k)\,dx \tag{24}$$

Note the use of an L^1 normalization for the wavelet coefficients that better fits local regularity analysis.

As stated above, the Hölder exponent can be used as a measurement of pointwise regularity in the locally bounded functions setting only, see [70]. Whether empirical data can be well-modelled by locally bounded functions or not can be determined numerically through the computation of the *uniform Hölder exponent* h_f^{min}, which, as for the scaling function, enjoys a function space characterization

$$h_f^{min} = \sup\{\alpha : f \in C^\alpha(\mathbb{R})\}, \tag{25}$$

where $C^\alpha(\mathbb{R})$ denotes the usual Hölder spaces. Assuming that φ and ψ are smooth enough and that ψ has enough vanishing moments, then the exponent h_f^{min} has the following simple wavelet characterization:

$$h_f^{min} = \liminf_{j \to +\infty} \frac{\log\left(\sup_k |c_{j,k}|\right)}{\log(2^{-j})}. \tag{26}$$

It follows that, if $h_f^{min} > 0$, then f is a continuous function, whereas, if $h_f^{min} < 0$, then f is not a locally bounded function, see [5, 71].

In numerous real world applications the restriction $h_f^{min} > 0$ constitutes a severe limitation; we will meet such examples in the case of physiological data (see also [5] for other examples). From a practical point of view, the regularity of the wavelets should be larger than h_f^{min} in order to compute the estimation of h_f^{min}. In the applications that we will see later, we took Daubechies compactly supported wavelets of increasing regularity and we stopped as soon as we found a threshold beyond which there is no more modification of the results. In our case, we stopped at order 3. In applications, the role of h_f^{min} is twofold: It can be used as a classification parameter and it tells whether a multifractal analysis based on the Hölder exponent is licit. Unlike other multifractality parameters that will be introduced in the following, its computation does not require a priori assumptions: It can be defined in the widest possible setting of tempered distributions.

We represent these two types of data on Fig. 1 for a marathon runner. The race is composed of several stages including a warm-up at the beginning, a recovery at the end of the marathon, and several moments of small breaks during the marathon. The signal was cleaned by removing the data that did not correspond to the actual race period (warm-ups, recoveries and breaks) and by making continuous connections to keep only the homogeneous parts. This type of connection is suitable for regularity exponents lower than 1 as in the case of our applications.

If $h_f^{min} < 0$, then a multifractal analysis based on the Hölder exponent cannot be developed, and the question whether a multifractal analysis based on the p-exponent can be raised see Fig. 2. Wavelet coefficients can also be used to determine whether f locally belongs to L^p or not (which is the a priori requirement needed in order to use the corresponding p-exponent), see [4, 5, 71]: Indeed, a simple wavelet criterium can be applied to check this assumption, through the computation of the *wavelet structure function*. Let

$$S_c(j, p) = 2^{-j} \sum_k |c_{j,k}|^p. \tag{27}$$

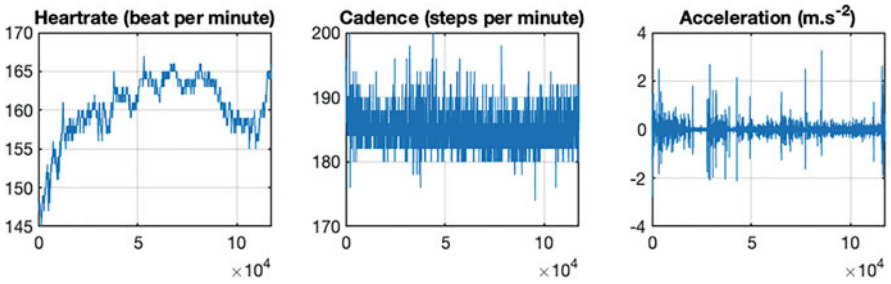


Fig. 1 Representation of data: heart rate (left) in beats per minute, cadence (middle) in steps per minute and acceleration (top) in meters per second squared. The time scale is in 0.1 s

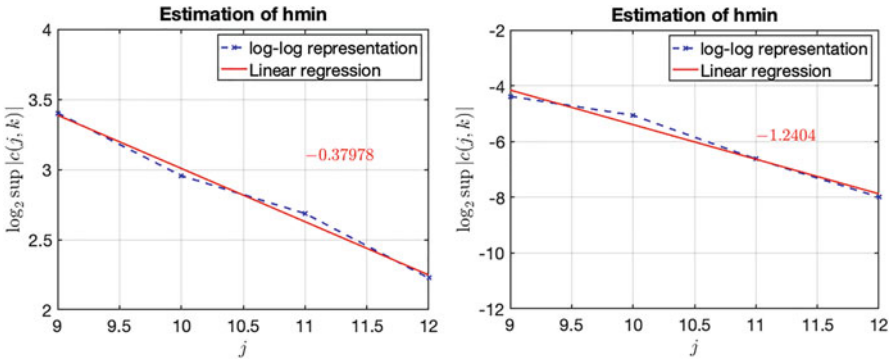


Fig. 2 Estimation by log-log regression of the h_{min} of a heart rate (left) and an acceleration (right). The points of the regression line up successfully along a close to straight line thus showing that the values of h_{min}, are precisely estimated and are negative. It follows that a multifractal analysis based on Hölder exponent cannot be performed on these data

The *wavelet scaling function* is defined as

$$\forall p > 0, \qquad \eta_f(p) = \liminf_{j \to +\infty} \frac{\log\left(S_c(j,\, p)\right)}{\log(2^{-j})}; \tag{28}$$

one can show that it coincides with the Kolmogorov scaling function if $p > 1$, see [55]. The following simple criterion can be applied in order to check if data locally belong to L^p [73]:

$$\left.\begin{array}{l} \text{if } \eta_f(p) > 0 \text{ then } f \in L^p_{loc}, \\ \text{if } \eta_f(p) < 0 \text{ then } f \notin L^p_{loc}. \end{array}\right\} \tag{29}$$

Remark 2 The wavelet scaling function enjoys the same property as h_f^{min}: Its computation does not require some a priori assumptions on the data, and it can be defined in the general setting of tempered distributions. Note that it is also defined for $p \in (0, 1]$; in that case the Sobolev space interpretation of the scaling function has to be slightly modified: In Definition 1 the Lebesgue space L^p has to be replaced by the real Hardy spaces H^p, see [96] for the notion of Hardy spaces and their wavelet characterization. Note that these function space interpretations imply that the wavelet scaling function does not depend on the specific (smooth enough) wavelet basis which is used; it also implies that it is unaltered by the addition of a smooth function, or by a smooth change of variables, see [4] and ref. therein (Fig. 3). For the same reasons, these properties also hold for the exponent h_f^{min}; they are required in order to derive intrinsic parameters for signal or image classification. In the following, we shall refer to them as *robustness properties*. In applications (28)

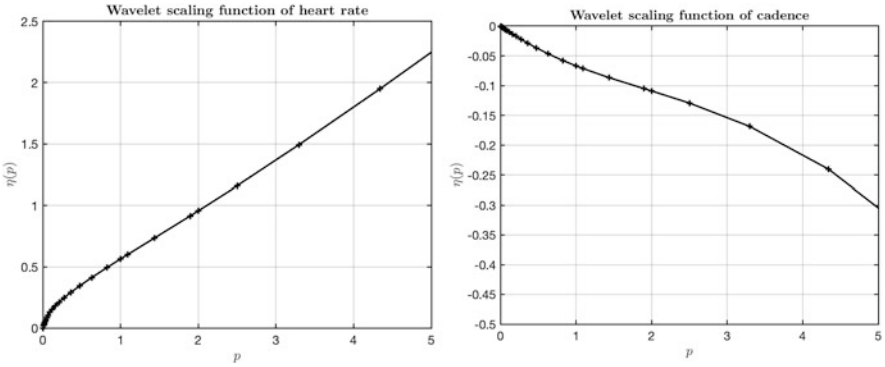


Fig. 3 Wavelet scaling function of heart rate (left) and cadence (right) of a marathon runner. It allows to determine the values of p such that $\eta_f(p) > 0$. We conclude that a multifractal analysis based on p-exponents is directly possible for heart rate data, but not for the cadence, where the analysis will have to be carried out on a fractional integral of the data

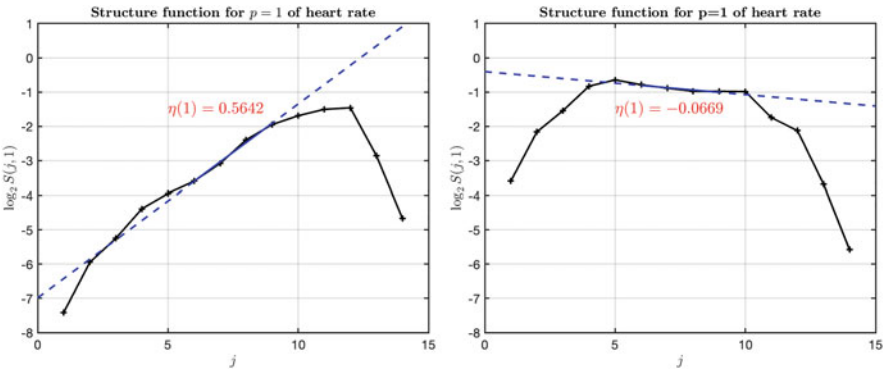


Fig. 4 Estimation by log-log regression of the wavelet scaling function of heart rate (left) and cadence (right) for $p = 1$. The slope of the regression is positive for heart rate and negative for cadence. These regressions, estimated for a sufficiently large number of values of p allow to plot the wavelet scaling functions, as shown in Fig. 3

can be used only if $\eta_f(p)$ can be determined by a log-log plot regression, i.e. when the lim inf actually is a limit, see e.g. Fig. 4. This means that the structure functions (27) satisfy $S_c(j, p) \sim 2^{-\eta_f(p)j}$ in the limit of small scales, a phenomenon coined *scale invariance*. The practical relevance of the wavelet scaling function (and other multifractal parameters that we will meet later), comes from the fact that it can be used for classification of signals and images without assuming that the data follow an a priori model.

2.5 Wavelet Pointwise Regularity Characterizations

One advantage of orthonormal wavelet based methods is that they allow to construct a multifractal analysis which is taylored for a given p-exponent, which is not the case of the alternative methods we mentioned. We shall see in Sects. 2.6 and 2.9 the benefits of this extra flexibility. For this purpose, we have to construct multiresolution quantities (i.e., in this context, a non-negative function defined on the collection of dyadic cubes) which are fitted to p-exponents. We start by introducing more adapted notations for wavelets and wavelet coefficients; instead of the two indices (j, k), we will use dyadic intervals (14) and, accordingly, $c_\lambda = c_{j,k}$, and $\psi_\lambda = \psi_{j,k}$. The wavelet characterization of p-exponents requires the definition of p-leaders. If $f \in L^p_{loc}(\mathbb{R})$, the wavelet p- *leaders* of f are defined as

$$\ell^{(p)}_{j,k} \equiv \ell^{(p)}_\lambda = \left(\sum_{\lambda' \subset 3\lambda} |c_{\lambda'}|^p \; 2^{j-j'} \right)^{1/p}, \tag{30}$$

where $j' \geq j$ is the scale associated with the sub-cube λ' included in 3λ (i.e. λ' has width $2^{-j'}$). Note that, when $p = +\infty$ (and thus $f \in L^\infty_{loc}(\mathbb{R})$), p-leaders boil down to *wavelet leaders*

$$\ell_\lambda = \sup_{\lambda' \subset 3\lambda} |c_{\lambda'}|,$$

[61, 112].

Let us indicate where such quantities come from. They are motivated by constructing quantities based on simple conditions on wavelet coefficients and which well approximate the local L^p norm of Definition 5. For that purpose we use the wavelet characterization of the Besov space $B^{0,p}_p$ which is "close" to L^p (indeed the classical embeddings between Besov and L^p spaces imply that $B^{0,1}_p \hookrightarrow L^p \hookrightarrow B^{0,\infty}_p$); with the normalization we chose for wavelet coefficients, the wavelet characterization of $B^{0,p}_p$ is given by

$$f \in B^{0,p}_p \quad \text{if} \quad \sum_k |c_k|^p < \infty \qquad \text{and} \qquad \sum_{j,k} 2^{(sp-1)j} |c_{j,k}|^p < \infty,$$

see [96] and, because of the localization of the wavelets, the restriction of the second sum to the dyadic cubes $\lambda' \subset 3\lambda$ yields an approximation of the local L^p norm of $f - P$ around the interval λ (the substraction of the polynomial P comes from the fact that the wavelets have vanishing moments so that P is reconstructed by the first sum in (22), and the wavelet coefficients $c_{j,k}$ of f and $f - P$ coincide). Actually, the uniform regularity assumption $\eta_f(p) > 0$ (which we will make) implies that the quantities (30) are finite.

Denote by $\lambda_{j,k}(x)$ the unique dyadic interval of length 2^{-j} which includes x; a key result is that both the Hölder exponent and the p-exponent can be recovered from, respectively, wavelet leaders and p-leaders, according to the following formula.

Definition 6 Let $h(x)$ be a pointwise exponent and (d_λ) a multiresolution quantity indexed by the dyadic cubes. The exponent h is derived from the (d_λ) if

$$\forall x, \qquad h(x) = \liminf_{j \to +\infty} \frac{\log\left(d_{\lambda_{j,k}(x)}\right)}{\log(2^{-j})}. \tag{31}$$

It is proved in [64, 67, 71] that if $\eta_f(p) > 0$, then the p-exponent is derived from p-leaders, and, if $h_f^{min} > 0$, then the Hölder exponent is derived from wavelet leaders. Note that the notion of p-exponent can be extended to values of p smaller that 1, see [63]; this extension requires the use of "good" substitutes of the L^p spaces for $p < 1$ which are supplied by the real Hardy spaces H^p. The important practical result is that the p-leaders associated with this notion also are given by (30).

In applications, one first computes the exponent h_f^{min} and the function $\eta_f(p)$. If $h_f^{min} > 0$, then one has the choice of using either p-leaders or wavelet leaders as multiresolution quantities. Though leaders are often preferred because of the simple interpretation that they yield in terms of the most commonly used (Hölder) exponent, it has been remarked that p-leaders constitute a quantity which displays better statistical properties, because it is based on averages of wavelet coefficients, instead of a supremum, i.e. a unique extremal value, see [7] and ref. therein. If both $h_f^{min} < 0$ and $\eta_f(p) < 0$ for all ps, then one cannot use directly these techniques and one performs a (fractional) integration on the data first. If one wants to use wavelet leaders, the order of integration s has to satisfy $s > -h_f^{min}$ since $h_{f^{(-s)}}^{min} = h_f^{min} + s$. Similarly, in the case of p-leaders it follows immediately from the Sobolev interpretation (7) of the wavelet scaling function that

$$\eta_{f^{(-s)}}(p) = ps + \eta_f(p).$$

Thus, if $\eta_f(p) < 0$, then an analysis based on p-leaders will be valid if the order of fractional integration s applied to f satisfies $s > -\eta_f(p)/p$. In practice, one does not perform a fractional integration on the data, but one simply replaces the wavelet coefficients $c_{j,k}$ by $2^{-sj}c_{j,k}$, which leads to the same scaling functions [5], and has the advantage of being performed at no extra computational cost.

2.6 Towards a Classification of Pointwise Singularities

In Sect. 2.3 we motivated the introduction of alternative pointwise regularity exponents by the requirement of having a tool available for non locally bounded

functions, which allows to deal directly with the data without having recourse to a smoothing procedure first; but this variety of exponents can also serve another purpose: By comparing them, one can draw differences between several types of singularities. This answers an important challenge in several areas of science; for example, in fully developed turbulence, some models predict the existence of extremely oscillating structures such as (2) and the key signal processing problem for the detection of gravitational waves also involves the detection of pointwise singularities similar to (2) in extremely noisy data [45].

Let us start with a simple example: Among the functions which satisfy $h_f(x_0) = \alpha$, the most simple pointwise singularities are supplied by *cusps* singularities, i.e. by functions which "behave" like

$$C_\alpha(x) = |x - x_0|^\alpha \qquad (\text{if } \alpha > 0 \text{ and } \alpha \notin 2\mathbb{N}). \tag{32}$$

How can we "model" such a behavior? A simple answer consists in remarking that the primitive of (32) is of the same form, and so on if we iterate integrations. Since the mapping $t \to h_{f^{(-t)}}(0)$ is concave [10], it follows that (32) satisfies

$$\forall t > 0, \qquad h_{C_\alpha^{(-t)}}(t_0) = \alpha + t.$$

For cusp singularities, the pointwise Hölder exponent is exactly shifted by the order of integration. This is in sharp contrast with the chirps (2), for which a simple integration by parts yields that the Hölder exponent of its n-th iterated primitive is

$$\forall n \in \mathbb{N}, \qquad h_{C_{\alpha,\beta}^{(-n)}}(t_0) = \alpha + (1+\beta)n,$$

from which it easily follows that the fractional primitives of the chirp satisfy

$$\forall t > 0, \qquad h_{C_{\alpha,\beta}^{(-t)}}(t_0) = \alpha + (1+\beta)t,$$

[10]. We conclude from these two typical examples that inspecting simultaneously the Hölder exponents of f and its primitives, or its fractional integrals, allows to put in light that oscillating behaviour of f in the neighbourhood of its singularities which is typical of (2) (see [107] for an in-depth study of the information revealed by the mapping $t \to h_{f^{(-t)}}(t_0)$). To that end, the following definition was proposed, which encapsulates the relevant "oscillatory" information contained in this function, using a single parameter.

Definition 7 Let $f : \mathbb{R}^d \to \mathbb{R}$ be such that $f \in L^p_{loc}$. If $h^p_f(x_0) \neq +\infty$, then the oscillation exponent of f at x_0 is

$$O_f(x_0) = \left(\frac{\partial}{\partial t} h^p_{f^{(-t)}}(x_0) \right)_{t=0^+} - 1. \tag{33}$$

Remark 3 In theory, a dependency in p should appear in the notation since f belongs to several L^p spaces. However, in practice, a given p is fixed, and this inaccuracy does not pose problems.

The choice of taking the derivative at $t = 0^+$ is motivated by a robustness argument: The exponent should not be perturbed when adding to f a smoother term, i.e. a term that would be a $O(|x-x_0|^h)$ for an $h > h_f(x_0)$; it is a consequence of the following lemma, which we state in the setting of Hölder exponents (i.e. we take $p = +\infty$ in Definition 7).

Lemma 1 *Let f be such that $h_f(x_0) < +\infty$ and $O_f(x_0) < +\infty$; let $g \in C^\alpha(x_0)$ for an $\alpha > h_f(x_0)$. Then, for s small enough, the Hölder exponents of $(f+g)^{(-s)}$ and of $f^{(-s)}$ coincide.*

Proof By the concavity of the mapping $s \to h_{f^{(-s)}}(x_0)$, see [6, 72], it follows that

$$h_{f^{(-s)}}(x_0) \leq h_f(x_0) + (1 + O_f(x_0))s;$$

but one also has $h_{g^{(-s)}}(x_0) \leq \alpha + s$; so that, for s small enough, $h_{g^{(-s)}}(x_0) > h_{f^{(-s)}}(x_0)$, and it follows that $h_{(f+g)^{(-s)}}(x_0) = h_{f^{(-s)}}(x_0)$. □

The oscillation exponent takes the value β for a chirp; it is the first of *second generation exponents* that do not measure a regularity, but yield additional information, paving the way to a richer description of singularities. In order to go further in this direction, we consider another example: *Lacunary combs*, which were first considered in [6, 72] (we actually deal here with a slight variant). Let $\phi = \mathbb{1}_{[0,1]}$.

Definition 8 Let $\alpha \in \mathbb{R}$ and $\gamma > \omega > 0$. The lacunary comb $F^\alpha_{\omega,\gamma}$, is

$$F^\alpha_{\omega,\gamma}(x) = \sum_{j=1}^{\infty} 2^{-\alpha j} \phi\left(2^{\gamma j}(x - 2^{-\omega j})\right). \tag{34}$$

We consider its behaviour near the singularity at $x_0 = 0$: if $\alpha > -\gamma$, then $F^\alpha_{\omega,\gamma} \in L^1(\mathbb{R})$ and it is locally bounded if and only if $\alpha \geq 0$. In that case, one easily checks that

$$h_{F^\alpha_{\omega,\gamma}}(0) = \frac{\alpha}{\omega}, \quad \text{and} \quad h_{F^\alpha_{\omega,\gamma}{}^{(-1)}}(0) = \frac{\alpha+\gamma}{\omega} \tag{35}$$

and one obtains (see [6]) that $O_{F^\alpha_{\omega,\gamma}}(0) = \frac{\gamma}{\omega} - 1$.

We conclude that chirps and lacunary combs are two examples of oscillating singularities. They are, however, of different nature: In the second case, oscillation is due to the fact that this function vanishes on larger and larger proportions of small balls centered at the origin (this is detailed in [72], where this phenomenon is precisely quantified through the use of *accessibility exponent* of a set at a point). On the other hand, chirps are oscillating singularities for a different reason: It is

due to very fast oscillations, and compensations of signs. This can be checked by verifying that the oscillation exponent of $|C_{\alpha,\beta}|$ at 0 vanishes.

We will now see that this difference can be put in evidence by considering the variations of the p-exponent. Comparing the p-exponents of chirps and lacunary combs allows to draw a distinction between their singularities; indeed, for $p \geq 1$, see [73],

$$h^p_{F^\alpha_{\omega,\gamma}}(0) = \alpha + \frac{1}{p}\left(\frac{\gamma}{\omega} - 1\right) \tag{36}$$

whereas a straightforward computation yields that

$$\forall p, \qquad h^p_{C_{\alpha,\beta}}(0) = \alpha.$$

We conclude that the p-exponent of $F^\alpha_{\omega,\gamma}$ varies with p, whereas the one of $C_{\alpha,\beta}$ does not. We will introduce another pointwise exponent which captures the lacunarity of the combs; it requires first the following notion: If $f \in L^p_{loc}$ in a neighborhood of x_0 for $p > 1$, the *critical Lebesgue index* of f at x_0 is

$$p_f(x_0) = \sup\{p : f \in L^p_{loc}(\mathbb{R}) \text{ in a neighborhood of } x_0\}. \tag{37}$$

The p-exponent at x_0 is defined on the interval $[1, p_f(x_0)]$ or $[1, p_f(x_0))$. We denote: $q_f(x_0) = 1/p_f(x_0)$. Note that $p_f(x_0)$ can take the value $+\infty$. An additional pointwise exponent, which, in the case of lacunary combs, quantifies the sparsity of the "teeth" of the comb, can be defined as follows see [72]. Its advantage is that it quantifies the "lacunarity information" using a single parameter instead of the whole function $p \to h^{(p)}_f(x_0)$.

Definition 9 Let $f \in L^p_{loc}$ in a neighborhood of x_0 for a $p > 1$. The lacunarity exponent of f at x_0 is

$$L_f(x_0) = \frac{\partial}{\partial q}\left(h^{(1/q)}_f(x_0)\right)_{q=q_f(x_0)^+}. \tag{38}$$

This quantity may have to be understood as a limit when $q \to q_f(x_0)$, since $h^{1/q}_f(x_0)$ is not necessarily defined for $q = q_f(x_0)$. This limit always exists as a consequence of the concavity of the mapping $q \to h^{1/q}_f(x_0)$, and it is nonnegative (because this mapping is increasing).

The lacunarity exponent of $F^\alpha_{\omega,\gamma}$ at 0 is $\frac{\gamma}{\omega} - 1$, which puts into light the fact that this exponent allows to measure how $F^\alpha_{\omega,\gamma}$ vanishes on "large sets" in the neighborhood of 0 (see [72] for a precise statement). Furthermore the oscillation exponent of $F^\alpha_{\omega,\gamma}$ at 0 is $\frac{\gamma}{\omega} - 1$, so that it coincides with the lacunarity exponent. The oscillation exponent is always larger than the lacunarity exponent. A way to distinguish between the effect due to lacunarity and the one due to cancellations is

to introduce a third exponent, the *cancellation exponent*

$$C_f(x_0) = O_f(x_0) - L_f(x_0).$$

The lacunarity and the cancellation exponents lead to the following classification of pointwise singularities see [6].

Definition 10 Let f be a tempered distribution on $\mathbb{R}$:

- f has a **canonical singularity** at x_0 if $O_f(x_0) = 0$.
- f has a **balanced singularity** at x_0 if $L_f(x_0) = 0$ and $C_f(x_0) \neq 0$.
- f has a **lacunary singularity** at x_0 if $C_f(x_0) = 0$ and $L_f(x_0) \neq 0$.

Cusps are typical examples of canonical singularities, chirps are typical examples of balanced singularities and lacunary combs are typical examples of lacunary singularities.

Many probabilistic models display lacunary singularities: It is the case e.g. for random wavelet series [6, 72], some Lévy processes, see [18] or fractal sums of pulses [103]. Note that our comprehension of this phenomenon is very partial: For instance, in the case of Lévy processes, the precise determination of the conditions that a Lévy measure should satisfy in order to guarantee the existence of lacunary singularities has not been worked out: in [18], P. Balanca proved that some self-similar Lévy processes with even Lévy measure display oscillating singularities, which actually turn out to be lacunary singularities and also that Lévy processes which have only positive jumps do not display such singularities; and, even in these cases, only a lower bound on their Hausdorff dimensions has been obtained. In other words, for Lévy processes, a joint multifractal analysis of the Hölder and the lacunarity exponent remains to be worked out. Note also that there exists much less examples of functions with balanced singularities: In a deterministic setting it is the case for the Riemann function [68] at certain rational points. However, to our knowledge, stochastic processes with balanced singularities have not been met up to now.

Another important question is to find numerically robust ways to determine if a signal has points where it displays balanced or lacunary singularities. This question is important in several areas of physics; for instance, in hydrodynamic turbulence, proving the presence of oscillating singularities would validate certain vortex stretching mechanisms which have been proposed, see [49]. Another motivation is methodological: if a signal only has canonical singularities, then its p-multifractal spectrum does not depend on p and its singularity spectrum is translated by t after a fractional integral of order, so that all methods that can be used to estimate its multifractal spectrum yield the same result (up to a known shift in the case of a fractional integration). An important questions related with the multifractal formalism is to determine if some of its variants allow to throw some light on these problems. Motivated by applications to physiological data, we shall come back to this question in Sects. 2.9 and 3.3.

Note that the choice of three exponents to characterize the "behaviour" of a function in the neighbourhood of one of its singularities may seem arbitrary; indeed, one could use the very complete information supplied by the following two variables function: If f is a tempered distribution, then the *fractional exponent* of f at x_0 is the two variable function

$$\mathcal{H}_{f,x_0}(q,t) = h^{1/q}_{f^{(-t)}}(x_0) - t,$$

see [6] where this notion is introduced and its properties are investigated. However, storing the pointwise regularity behaviour through the use of a two-variables function defined at every point is unrealistic, hence the choice to store only the information supplied by the three parameters we described. This choice is motivated by two conflicting requirements: On one hand, one wishes to introduce mathematical tools which are sophisticated enough to describe several "natural" behaviours that can show up in the data, such as those supplied by cusps, chirps, and lacunary combs. On other hand, at the end, classification has to bear on as little parameters as possible in order to be of practical use in applications; the goal here is to introduce a multivariate multifractal analysis based on a single function f, but applied to several pointwise exponents associated with f (say two or three among a regularity, a lacunarity and a cancellation exponent).

Our theoretical comprehension of which functions can be pointwise exponents is extremely partial, see [106] for a survey on this topic: It has been known for a long time that a pointwise Hölder exponent $h_f(x)$ can be any nonnegative function of x which can be written as a liminf of a sequence of continuous functions, see [16, 39, 53], but the same question for p-exponents is open (at least in the case where it takes negative values). Similarly, which couples of functions $(h(x), O(x))$ can be the joint Hölder and oscillation exponents of a function also is an open question (see [57] for partial results), and it is the same if we just consider the oscillation exponent, or couples including the lacunarity exponent. One meets similar limitations for multifractal spectra: In the univariate setting supplied by the multifractal Hölder spectrum, the general form of functions which can be multifractal spectra is still open; nonetheless a partial result is available: functions which can be written as infima of a sequence of continuous functions are multifractal spectra [52]; additionally, as soon as two exponents are involved, extremely few results are available. For instance, if f is a locally bounded function, define its *bivariate oscillation spectrum* as

$$\mathcal{D}_f(H,\beta) = \dim\{h : h_f(x) = H \quad \text{and} \quad O_f(x) = \beta\}.$$

Which functions of two variables $D(H,\beta)$ can be bivariate oscillation spectra is a completely open problem.

2.7 Mathematical Results Concerning the Multifractal Formalism

We now consider a general setting where $h : \mathbb{R} \to \mathbb{R}$ is a pointwise exponent derived from a multiresolution quantity $d_\lambda (= d_{j,k})$ according to Definition 6, and defined in space dimension d. The associated multifractal spectrum $\mathcal{D}$ is

$$\mathcal{D}(H) = \dim(\{x : \quad h(x) = H\}).$$

The *support* of the spectrum is the image of the mapping $x \to h(x)$, i.e. the collection of values of H such that

$$\{x \in \mathbb{R} : h(x) = H\} \neq \emptyset$$

(note that this denomination, though commonly used, is misleading, since it may not coincide with the mathematical notion of support of a function).

The *leader scaling function* associated with the multiresolution quantities $(d_{j,k})$ is

$$\forall q \in \mathbb{R}, \qquad \zeta_f(q) = \liminf_{j \to +\infty} \frac{\log\left(2^{-j} \sum_k |d_{j,k}|^q .\right)}{\log(2^{-j})}. \tag{39}$$

Note that, in contradistinction with the wavelet scaling function, it is also defined for $p < 0$. Referring to "leaders" in the name of the scaling function does not mean that the $d_{j,k}$ are necessarily obtained as wavelet leaders or wavelet p-leaders, but only to prevent any confusion with the wavelet scaling function. The *Legendre spectrum* is

$$\mathcal{L}(H) := \inf_{q \in \mathbb{R}} (1 + qH - \zeta_f(q)). \tag{40}$$

As soon as relationships such as (31) hold, then the following upper bound is valid

$$\forall H, \qquad \mathcal{D}(H) \leq \mathcal{L}(H) \tag{41}$$

(see [61] for particular occurrences of this statement, and [4] for the general setting). However, for a number of synthetic processes with known $\mathcal{D}(H)$ (and for a proper choice of the multiresolution quantity), this inequality turns out to be an equality, in which case, we will say that the *multifractal formalism holds*. The leader scaling functions obtained using wavelet leaders or p-leaders can be shown to enjoy the same robustness properties as listed at the end of Sect. 2.4, see [4] (it is therefore also the case for the Legendre spectrum). It follows from their mathematical and numerical properties that wavelet leader based techniques form the state of the art for real-life signals multifractal analysis (Fig. 5).

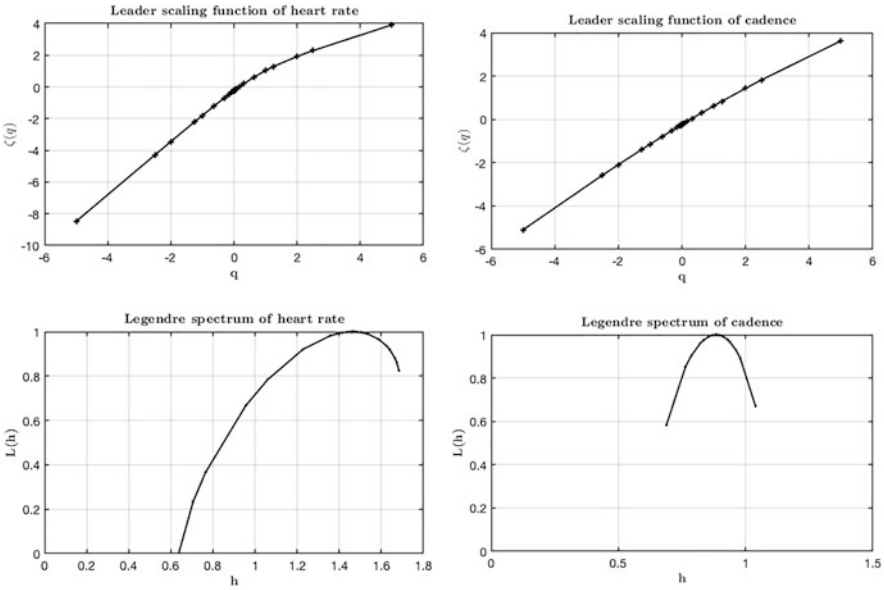


Fig. 5 Representation of scale function and the univariate Hölder Legendre spectra of the primitives of heart beat frequency (left) and cadence (right) of one marathon runner during the entire race. The multiresolution quantities used in these derivation are the wavelet leaders of the primitive of the data

In applications, one cannot have access to the regularity exponent at every point in a numerically stable way, and thus $\mathcal{D}(H)$ is unaccessible; this explains why, in practice, $\mathcal{L}(H)$ is the only computationally available spectrum, and it is used as such in applications. However, information on the pointwise exponent may be inferred from the Legendre spectrum. Such results are collected in the following theorem, where they are stated in decreasing order of generality.

Theorem 1 *Let $h : \mathbb{R} \to \mathbb{R}$) be a pointwise exponent, and assume that it is derived from multiresolution quantities $d_{j,k}$ according to Definition 6. The following results on h hold:*

- *Let*

$$h^{min} = \liminf_{j \to +\infty} \frac{\log\left(\sup_k d_{j,k}\right)}{\log(2^{-j})} \quad \text{and} \quad h^{max} = \liminf_{j \to +\infty} \frac{\log\left(\inf_k d_{j,k}\right)}{\log(2^{-j})} \tag{42}$$

then

$$\forall x \in \mathbb{R} \qquad h^{min} \leq h(x) \leq h^{max}. \tag{43}$$

- *If the Legendre spectrum has a unique maximum for $H = c_1$, then*

$$\text{for almost every } x, \qquad h(x) = c_1; \tag{44}$$

- *If the leader scaling function (39) associated with the $d_{j,k}$ is affine, then f is a monohölder function, i.e.*

$$\exists H_0 : \quad \forall x, \qquad h(x) = H_0,$$

where H_0 is the slope of the leader scaling function.

Remark 4 The last statement asserts that, if h is a pointwise exponent associated with a function f, then f is a monohölder function. This result has important implications in modeling since it yields a numerically simple test, based on global quantities associated with the signal, and which yields the pointwise exponent everywhere. This is in strong contradistinction with the standard pointwise regularity estimators, see e.g. [19] and ref. therein, which are based on local estimates, and therefore on few data thus showing strong statistical variabilities, and additionally often assume that the data follow some a priori models.

Proof We first prove the upper bound in (43). Let $\alpha > h^{max}$; there exists a sequence $j_n \to +\infty$ such that

$$\log\left(\inf_k d_{j_n,k}\right) \geq \log(2^{-\alpha j_n}),$$

so that at the scales j_n all d_λ are larger than $2^{-\alpha j_n}$. It follows from (31) that

$$\forall x, \qquad h(x) \leq \alpha,$$

and the upper bound follows. The proof of the lower bound is similar (see e.g. [70]). □

The second statement is direct consequence of the following upper bounds for the dimensions of the sets

$$E_H^+ = \{h(x) \geq H\} \quad \text{and} \quad E_H^- = \{h(x) \leq H\} \tag{45}$$

which are a slight improvement of (41), see [70]:

Proposition 2 *Let h be a pointwise exponent derived from the multiresolution quantity $(d_{j,k})$. Then the following bounds hold:*

$$\dim(E_H^-) \leq \inf_{q>0}(1 + qH - \zeta_f(q)) \quad \text{and} \quad \dim(E_H^+) \leq \inf_{q<0}(1 + qH - \zeta_f(q)) \tag{46}$$

Let us check how (44) follows from this result. Note that the first (partial) Legendre transform yields the increasing part of $\mathcal{L}(H)$ for $H \leq c_1$ and the second one yields the decreasing part for $H \geq c_1$. If $\mathcal{L}$ has a unique maximum for $H = c_1$, it follows from (46) that

$$\forall n, \qquad \dim(E^-_{c_1-1/n}) < 1 \quad \text{and} \quad \dim(E^-_{c_1+1/n}) < 1.$$

All of these sets therefore have a vanishing Lebesgue measure, which is also the case of their union. But this union is $\{x : h(x) \neq c_1\}$. It follows that almost every x satisfies $h(x) = c_1$.

Finally, if the leader scaling function is affine, then its Legendre transform is supported by a point H_0 and takes the value $-\infty$ elsewhere. The upper bound (41) implies that, if $H \neq H_0$ the corresponding isoregularity set is empty. In other words, H_0 is the only value taken by the pointwise exponent, and f is a monohölder function.

Remark 5 If $h^{min} = h^{max}$, the conclusion of the first and last statement are the same. However, one can check that the condition $h^{min} = h^{max}$ is slightly less restrictive than requiring the leader scaling function to be affine (the two conditions are equivalent if, additionally, the lim inf in (42) is a limit).

The parameter c_1 defined in Theorem 1 can be directly estimated using log-log plot (see [5] and ref. therein), and, in practice it plays an important role in classification as we will see in the next section. When the multiresolution quantity used is the p-leaders of a function f, the associated exponent c_1 may depend on p, and we will mention this dependency and denote this parameter by $c_1(p, f)$. This is in contradistinction with the exponent h^{min} defined by (42), which, in the case of functions with some uniform Hölder regularity, coincides with the exponent h_f^{min} defined by (26) for leaders and p-leaders, as shown by the following lemma; note that it is actually preferable to compute it using (26), which has the advantages of being well defined without any a priori assumption on f.

Lemma 2 *Let $f : \mathbb{R} \to \mathbb{R}$ be such that $h_f^{min} > 0$. Then the h^{min} parameter computed using p-leaders all coincide with the h_f^{min} computed using wavelet coefficients.*

Let us sketch the poof of this result. Suppose that $h_f^{min} > 0$ and let $\alpha > 0$ be such that $\alpha < h_f^{min}$. Then, the wavelet coefficients of f satisfy

$$\exists C, \quad \forall j, k \qquad |c_{j,k}| \leq C 2^{-\alpha j}.$$

Therefore the p-leaders of f satisfy

$$\ell_\lambda^{(p)} \leq \left(\sum_{\lambda' \subset 3\lambda} (2^{-\alpha j'})^p \; 2^{j-j'} \right)^{1/p}$$

$$\leq \left(\sum_{j' \geq j} 2^{-\alpha p j'} \; 2^{j-j'} \right)^{1/p} \leq C 2^{-\alpha j};$$

it follows that the corresponding p-leader is smaller that $|c_{\lambda_n}|$ so that the h^{min} computed using p-leaders is smaller that the one computed using wavelet coefficients. Conversely, by definition of h_f^{min}, there exists a sequence of dyadic intervals c_{λ_n} of width decreasing to 0, and such that

$$|c_{\lambda_n}| \sim 2^{-h_f^{min} j_n},$$

and the corresponding p-leader is larger that $|c_{\lambda_n}|$ so that the h^{min} computed using p-leaders is smaller that the one computed using wavelet coefficients.

The following result yields an important a priori bound on the dimensions of the singularity sets corresponding to negative regularity exponents, see [73].

Proposition 3 *Let $p > 0$, and let $f : \mathbb{R} \to \mathbb{R}$ be a function such that $\eta_f(p) > 0$. Then its p-spectrum satisfies*

$$\forall h, \qquad \mathcal{D}_p(H) \leq 1 + Hp$$

Let us elaborate on the information supplied by the exponent $c_1(p, f)$: A direct consequence of (44) is that, if a signal f satisfies that the exponent $c_1(p, f)$ takes the same value for $p_1 < p_2$, then this implies that the p-exponent satisfies that

$$\text{for almost every } x, \qquad h_f^{p_1}(x) = h_f^{p_2}(x),$$

which implies that the mapping $p \to h_f^{p_1}(x)$ is constant for $p \in [p_1, p_2]$; but, since the mapping $p \to h_f^{1/p}(x_0)$ is concave and increasing, see [6, 72], it follows that this mapping is constant for p small enough; as a consequence, the lacunarity exponent vanishes at x. Similarly, if, for a given p, $c_1(p, f^{(-1)}) - c_1(p, f) = 1$, this implies that

$$\text{for almost every } x, \qquad h_{f^{(-1)}}^{p}(x) = h_f^p(x) + 1,$$

and the same argument as above, see [6, 72], yields the absence of oscillating singularities for almost every point. In other words, the computation of $c_1(p)$ yields a key information on the nature of the singularities a.e. of the signal, which we

summarize in the following statement, which will have implications in the next section for the analysis of marathon runners data.

Proposition 4 *Let $f : \mathbb{R} \to \mathbb{R}$ be a function in L^p.*
If

$$\exists q > p : \qquad c_1(p, f) = c_1(q, f),$$

then for almost every x, f has no lacunary singularity at x.
If f satisfies

$$\exists p : \qquad c_1(p, f^{(-1)}) - c_1(p, f) = 1,$$

then, for almost every x, f has a canonical singularity at x.

These two results are characteristic of signals that only contain canonical singularities, see Sect. 2.6, and they also demonstrate that $c_1(p, f)$, which, in general, depends on the value of p is intrinsic for such data (see a contrario [72] where the exponent $c_1(p, f)$ of lacunary wavelet series is shown to depend on the value of p, and [103] where the same result is shown for random sums of pulses). Note that such results are available in the discrete wavelet approach only; they would not be possible using the WTMM or the MFDFA approaches, which do not allow to draw differences between various pointwise regularity exponents and therefore do not yield spectra fitted to different values of the p-exponent. To summarize, the advantages of the p-leader based multifractal analysis framework are: the capability to estimate negative regularity exponents, better estimation performances, and a refined characterization of the nature of pointwise regularities.

One important argument in favor of multifractal analysis is that it supplies robust classification parameters, in contradistinction with pointwise regularity which can be extremely erratic. Consider for instance the example of a sample path of a Lévy process without Brownian component (we choose this example because such processes now play a key role in statistical modeling): Its Hölder exponent is a random, everywhere discontinuous, function which cannot be numerically estimated or even drawn [56]: In any arbitrary small interval $[a, b]$ it takes all possible values $H \in [0, H^{max}]$. On the opposite, the multifractal spectrum (which coincides with the Legendre spectrum) is extremely simple and robust to estimate numerically: It is a deterministic linear function on the interval $[0, H^{max}]$ (with $D(H^{max}) = 1$). This example is by no means accidental: though one can simply construct stochastic processes with a random multifractal spectrum (consider for instance a Poisson process restricted to an interval of finite length), large classes of classical processes have simple deterministic multifractal spectra (and Legendre spectra), though no simple assumption which would guarantee this results is known. The determination of a kind of "0-1 law" for multifractal spectra, which would guarantee that, under fairly general assumptions, the spectrum almost surely is a deterministic function, is a completely open problem, and its resolution would greatly improve our understanding of the subject. Even in the case of Gaussian processes, though it is

known that such processes can have a random Hölder exponent [15], the possibility of having a random multifractal spectrum still is a open issue.

2.8 Generic Results

Let us come back to the problem raised in Sect. 2.1 of estimating the size of the Hölder singularity sets of increasing functions which led us to the key idea that the Hausdorff dimension is the natural way to estimate this size. One can wonder if the estimate (15) that we found for the multifractal spectrum is optimal. In 1999, Z. Buczolich and J. Nagy answered this question in a very strong way, showing that it is sharp for a *residual* set of continuous increasing functions, see [34]. What does this statement precisely mean? Let E be the set of continuous increasing functions $f : \mathbb{R} \to \mathbb{R}$, endowed with the natural distance supplied by the sup norm. Then equality in (15) holds (at least) on a residual set in the sense of Baire categories, i.e. on a countable intersection of open dense sets.

This first breakthrough opened the way to genericity results in multifractal analysis. They were the consequence of the important remark that scaling functions for $p > 0$ can be interpreted as stating that f belongs to an intersection of Sobolev spaces E_η (in the case of the Kolmogorov scaling function) or of a variant of these spaces, the *oscillation spaces* in the case of the leader scaling function [62]. One easily checks that E_η is a complete metric space, and the Baire property therefore is valid (i.e. a countable intersection of open dense sets is dense). The question formulated by Parisi and Frisch in [99], can be reformulated in this setting: If equality in (41) cannot hold for *every* function in E_η (since e.g. because it contains C^∞ functions), nonetheless it holds on a residual set [59]. This result found many extensions: The first one consists in replacing the genericity notion supplied by Baire's theorem by the more natural notion supplied by *prevalence*, which is an extension, in infinite dimensional function spaces of the notion of "Lebesgue almost everywhere", see [38, 115] for the definition of this notion and its main properties, and [48] for its use in the setting of multifractal analysis. The conclusions drawn in the Baire setting also hold in the prevalence setting, and raise the question of the determination of a stronger notion of genericity, which would imply both Baire and prevalence genericity, and which would be the "right " setting for the validity of the multifractal formalism. A natural candidate is supplied by the notion of *porosity* , see [87], but the very few results concerning multifractal analysis in this setting do not allow to answer this question yet. Note also that Baire and prevalence results have been extended to the p-exponent setting [47], which allows to deal with spaces of functions that are not locally bounded. Another key problem concerning the generic validity of the multifractal formalism concerns the question of taking into account the information supplied by negative values of p in the scaling function (39). The main difficulty here is that the scaling function does not define a function space any longer, and the "right" notion of genericity which should be picked is completely open: Though Baire and prevalence do not really require the setting

supplied by a (linear) function space, nonetheless these notions are not fitted to the setting supplied by a given scaling function which includes negative values of p. In [22] J. Barral and S. Seuret developed an alternative point of view which is less "data driven": They reinterpreted the question in the following way: Given a certain scaling function $\eta(p)$, they considered the problem of constructing an ad hoc function space which is taylored so that generically (for the Baire setting), functions in such a space satisfy the multifractal formalism for the corresponding scaling function, including its values for $p < 0$ (and Legendre spectrum). Another limitation of the mathematical results of genericity at hand is that they are not able to take into account *selfsimilarity* information: In (28), in order to introduce a quantity which is always well-defined, and corresponds to a function space regularity index, the scaling function is defined by a lim inf. But, most of the time, what is actually observed on the data (and what is really needed in order to obtain a numerically robust estimate) is that this lim inf actually is a true limit, which means that the L^p averages of the data display exact power-law behaviours at small scales. Up to now, one has not been able to incorporate this type of information in the function space modeling developed.

2.9 Implications on the Analysis of Marathon Runners Data

The increasing popularity of marathons today among all ages and levels is inherited from the human capacity to run long distances using the aerobic metabolism [86], which led to a rising number of amateur marathon runners who end the 42,195 km between 2h40min and 4h40min. Therefore, even if nowadays, marathon running becomes "commonplace", compared with ultra-distance races, this mythic Olympic race is considered to be the acme of duration and intensity [93]. Running a marathon remains scary and complex due to the famous "hitting the wall" phenomenon, which is the most iconic feature of the marathon [28]. This phenomenon was previously evaluated in a large-scale data analysis of late-race pacing collapse in the marathon [110]; Smyth [109] presented an analysis of 1.7 million recreational runners, focusing on pacing at the start and end of the marathon, two particularly important race stages. They showed how starting or finishing too quickly could result in poorer finish-times, because fast starts tend to be very fast, leading to endurance problems later, while fast finishes suggest overly cautious pacing earlier in the race [109]. Hence, the definition of a single marathon pace is based on the paradigm that a constant pace would be the ideal one. However, in [30], a 3 years study shows that large speed and pace variations are the best way to optimize performance. Marathon performance depends on pacing oscillations between non symmetric extreme values [101]. Heart rate (HR) monitoring, which reflects exercise intensity and environmental factors, is often used for running strategies in marathons. However, it is difficult to obtain appropriate feedback for only the HR value since, as we saw above, the cardiovascular drift occurs during prolonged exercise. Therefore, now we have still to investigate whether this pace

(speed) variation has a fractal behavior and if so, whether this is the case for the runners's heart rate which remains a pacer for the runners who aim to keep their heart rate in a submaximal zone (60–80% of the maximal heart rate) [93]. Here, we hypothesized that marathonians acceleration (speed variation), cadence (number of steps per minute) and heart rate time series follow a multifractal formalism and could be described by a self similar function. Starting in the 1990s, many authors demonstrated the fractal behavior of physiological data such as heart rate, arterial blood pressure, and breath frequency of human beings, see e.g. [2, 51]. In 2005, using the Wavelet Transform Maxima Method, E. Wesfreid, V. L. Billat and Y. Meyer [113] performed the first multifractal analysis of marathonians heartbeats. This study was complemented in 2009 using the DFA (Detrended Fluctuation Analysis) and wavelet leaders applied on a primitive of the signal [29]. Comparing the outputs of these analyses is hasardous; indeed, as already mentioned, these methods are not based on the same regularity exponents: WTMM is adapted to the *weak scaling exponent* [97], DFA to the p-exponent for $p = 2$ [73, 84], and wavelet leaders to the Hölder exponent [61]. In the following, we will propose a method of digital multifractal analysis of signals based on p-leaders, which, in some cases, can avoid performing fractional integrations (or primitives) and thus transform the signal. In [29], it was put in evidence that multifractal parameters associated with heart beat intervals evolve during the race when the runner starts to be deprived of glycogen (which is the major cause of the speed diminution at the end of the race. This study also revealed that fatigue decreases the running speed and affects the regularity properties of the signal which can be related with the feelings of the runner measured by the Rate of Perception of Exhaustion (RPE), according to the psychophysiological scale of Borg (mainly felt through the breathing frequency). In addition, there is a consistent decrease in the relationship between speed, step rate, cardiorespiratory responses (respiratory rate, heart rate, volume of oxygen consumed), and the level of Rate of Perception of Exhaustion (RPE), as measured by Borg's psychophysiological scale. The runner does not feel the drift of his heart rate, in contradistinction with his respiratory rate. These physiological data are not widely available and only heart rate and stride rate are the measures available to all runners for economic reasons. Moreover, these data are generated heartbeat by heartbeat and step by step.

Our purpose in this section is to complement these studies by showing that a direct analysis on the data is possible if using p-leaders (previous studies using the WTMM or the standard leaders had to be applied to a primitive of the signal), and that they lead to a sharper analysis of the physiological modifications during the race. We complement the previous analyses in order to demonstrate the modifications of multifractal parameters during the race, and put in evidence the physiological impact of the intense effort after the 20th km. For that purpose, we will perform a multifractal analysis based on p-leaders.

We analyzed the heartbeat frequency of 8 marathon runners (men in the same age area). Figure 2 shows the determination of exponents h_f^{min} for heartbeat frequency and cadence through a log-log regression; the regression is always performed between the scales j = 8 and j = 11 (i.e. between 26s and 3mn 25s), which have

been identified as the pertinent scales for such physiological data, see [2]. For most marathon runners, h_f^{min} is negative, see Table 1, which justifies the use of p-leaders. We then compute the wavelet scaling function in order to determine a common value of p for which all runners satisfy $\eta(p) > 0$, see Fig. 3 where examples of wavelet scaling function are supplied for heartbeat frequency and cadence. In the case of heartbeat frequency, the computation of the 8 wavelet scaling functions yields that $p = 1$ and $p = 1.4$ can be picked. The corresponding p-leaders multifractal analysis is performed for these two values of p, leading to values of $c_1(p)$ which are also collected in Table 1.

In Fig. 6, the value of the couple $(h_f^{min}, c_1(p))$ is plotted (where we denote by $c_1(p)$ the value of H for which the maximum of the p-spectrum is reached). The values of $c_1(p)$ are very close to 0.4 whereas the values of h_f^{min} notably differ, and are clearly related with the level of practice of the runners. Thus M8 is the only trail runner and improved his personal record on that occasion; he practices more and developed a very uneven way of running. Table 1 shows that the values of $c_1(p)$ do not notably differ for different values of p and, when computed on a primitive of the signal, are shifted by 1. We are in the situation described in Proposition 4 and we conclude in the absence of oscillating singularities at almost every point. This result also shows that $c_1(p)$, which may depend on the value of p (see [72] where it is shown that it is the case for lacunary wavelet series), is intrinsic for such data. We will see in Sect. 3.5 that a bivariate analysis allows to investigate further in the nature of the pointwise singularities of the data.

We now consider the evolution of the multifractality parameters during a marathon: at about the 25th km (circa 60% of the race) runners feel an increased penibility on the RPE Borg scale (Fig. 7). Therefore we expect to find two regimes with different parameters before and after this moment. This is put in evidence by Fig. 8 which shows the evolution of the multifractality parameters during the first half and the last fourth of the marathon thus putting in evidence the different physiological reactions at about the 28th km. From the evolution of the multifractal parameters between the beginning and the end of the marathon race, we can distinguish between the less experimented marathon runners, whichever their level of fitness, and those who know how to self pace their race. Indeed, according to the evolution of the couple $(h_f^{min}, c_1(p))$, the less experimented (R 7) loosed the regularity of his heart rate variation. This shows that the marathon running experience allows to feel how to modulate the speed for a conservative heart rate variability. From the evolution of the multifractal parameters between the beginning and the end of the marathon race, we can distinguish between the less experimented marathon runners, whichever their level of fitness and those who know how to self pace their race. In [101] its was shown that the best marathon performance was achieved with a speed variation between extreme values. Furthermore, a physiological steady state (heart rate and other cardiorespiratory variables), are obtained with pace variation [31]. This conclusion is in opposition with the less experimented runners beliefs that the constant pace is the best, following the mainstream non scientific basis recommendations currently available on internet.

Table 1 Multifractal analysis of heartbeat frequency of marathon runners (Pr.: primitive)

	H_{min}	H_{min} of the Pr.	c_1 for $p = 1$	c_1 for $p = 1.4$	c_1 of the Pr. for $p = 1$	c_1 of the Pr. for $p = 1.4$
R1	−0.2768	0.7232	0.8099	0.8064	1.8242	1.8213
R2	−0.0063	0.9937	0.4564	0.4043	1.3926	1.3509
R3	−0.0039	0.9961	0.6856	0.6625	1.6942	1.6351
R4	−0.1633	0.8367	0.6938	0.6785	1.6653	1.6636
R5	−0.2434	0.7566	0.5835	0.5689	1.5401	1.5224
R6	−0.3296	0.6704	0.5809	0.5636	1.5644	1.5500
R7	0.1099	1.1099	0.5652	0.5483	1.4754	1.4379
R8	−0.5380	0.4620	0.3382	0.2977	1.2588	1.2086

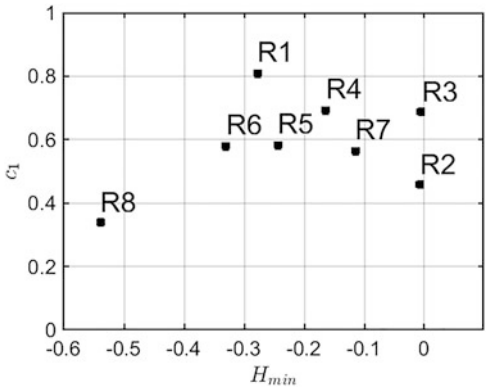


Fig. 6 Representation of the pair $(H_{min}, c_1(p))$ with $p = 1$ deduced from the 1-spectrum of heart rate and computed for the entire race; H_{min} appears as the most relevant classification parameter. The isolated point on the left corresponds to R8, the most trained runner

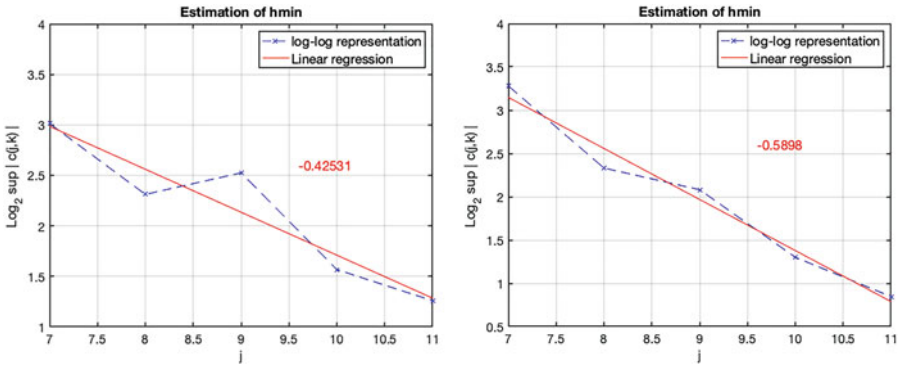


Fig. 7 Estimation of h_f^{min} by log-log regression for the heart rate of a marathon runner at the beginning (50% first part of the race) on the left and the end (25% last part of the race) on the right. The clear difference of the values obtained shows that the exponent h_f^{min} is well fitted to characterize the evolution of physiological rythms during the race. These data, together with the evolution of the parameter $c_1(p)$, are collected in Fig. 8 with $p = 1$

In Sect. 3.5 we will investigate the additional information which is revealed by the joint analysis of several physiological data.

3 Multivariate Multifractal Analysis

Up to now, in most applications, multifractal analysis was performed in univariate settings, (see a contrario [88]), which was mostly due to a lack of theoretical foundations and practical analysis tools. Our purpose in this section is to provide a comprehensive survey of the recent works that started to provide these foundations, and to emphasize the mathematical questions which they open. In particular, multivariate spectra also encode on specific data construction mechanisms. Mul-

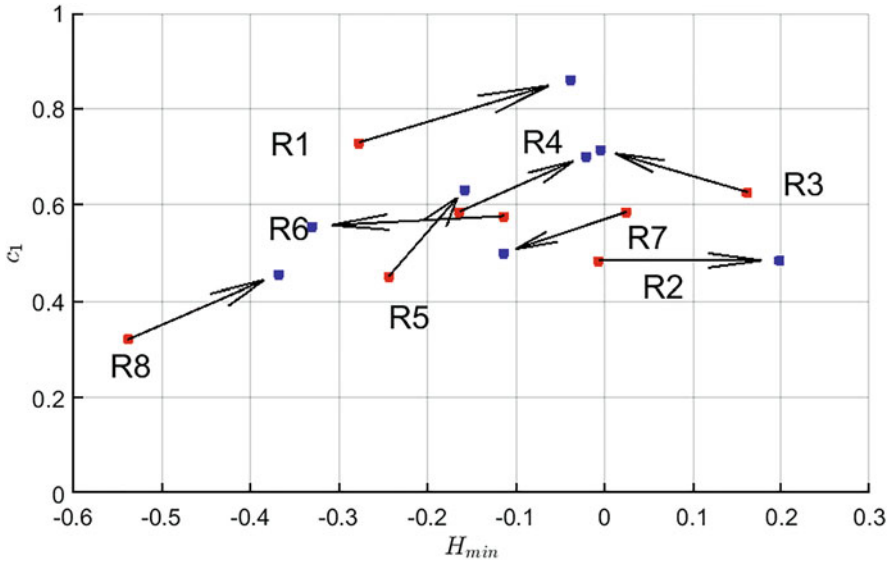


Fig. 8 Evolution of the couple $(H_{min}, c_1(p))$ with $p = 1$ deduced from the 1-spectrum of the heart rate between the beginning (in blue) and the end (in red) of the marathon: the evolutions are similar except for three runners: R3 and R6 who had great difficulties and R7 who is the least experienced runner with a much longer running time

tivariate multifractal analysis deals with the joint multifractal analysis of several functions. For notational simplicity, we assume in the following that we deal with two functions f_1 and f_2 defined on $\mathbb{R}^d$ and that, to each function is associated a pointwise regularity exponent $h_1(x)$ and $h_2(x)$ (which need not be the same).

3.1 Multivariate Spectrum

On the mathematical side, the main issue is to understand how the isoregularity sets

$$E_{f_1}(H_1) = \{x : h_1(x) = H_1\} \quad \text{and} \quad E_{f_2}(H_2) = \{x : h_2(x) = H_2\}$$

of each function are "related". A natural way to translate this loose question into a precise mathematical problem is to ask for the determination of the *multivariate multifractal spectrum* defined as the two-variables function

$$\mathcal{D}_{(f_1,f_2)}(H_1, H_2) = \dim(\{x : h_1(x) = H_1 \text{ and } h_2(x) = H_2\}). \tag{47}$$

this means that we want to determine the dimension of the intersection of the two isoregularity sets $E_{f_1}(H_1)$ and $E_{f_2}(H_2)$. The determination of the dimension of the intersection of two fractal sets usually is a difficult mathematical question, with no general results available, and it follows that few multivariate spectra have been determined mathematically, see e.g. [23, 24] for a joint analysis of invariant measures of dynamical systems. One can also mention correlated and anticorrelated binomial cascades, see Sect. 3.4 for the definition of these cascades, and [74] for the determination of bivariate spectra when two of these cascades are considered jointly.

On the mathematical side, two types of results often show up. A first category follows from the intuition supplied by intersections of smooth manifolds: In general, two surfaces in $\mathbb{R}^3$ intersect along a curve and, more generally, in $\mathbb{R}^d$, manifolds intersect *generically* according to the *sum of codimensions rule*:

$$\dim(A \cap B) = \min(\dim A + \dim B - d, -\infty)$$

(i.e. the "codimensions" $d - \dim A$ and $d - \dim B$ add up except if the output is negative, in which case we obtain the emptyset). This formula is actually valid for numerous examples of fractal sets, in particular when the Hausdorff and Packing dimensions of one of the sets A or B coincide (e.g. for general Cantor sets) [94]; in that case "generically" has to be understood in the following sense: For a subset of positive measure among all rigid motions σ, $\dim(A \cap \sigma(B)) = \min(\dim A + \dim B - d, -\infty)$. However the coincidence of Hausdorff and Packing dimensions needs not be satisfied by isoregularity sets, so that such results cannot be directly applied for many mathematical models. The only result that holds in all generality is the following: if A and B are two Borel subsets of $\mathbb{R}^d$, then, for a generic set of rigid motions σ, $\dim(A \cap \sigma(B)) \geq \dim A + \dim B - d$. This leads to a first rule of thumb for multivariate multifractal spectra: When two functions are randomly shifted, then their singularity sets will be in "generic" position with respect to each other, yielding

$$\mathcal{D}_{(f_1,f_2)}(H_1, H_2) \geq \mathcal{D}_{f_1}(H_1) + \mathcal{D}_{f_2}(H_2) - d.$$

In practice, this result suffers from two limitations: the first one is that, usually, one is not interested in randomly shifted signals but on the opposite for particular configurations where we expect the conjunction of singularity sets to carry relevant information. Additionally, for large classes of fractal sets, the *sets with large intersection*, the codimension formula is not optimal as they satisfy

$$\dim(A \cap B) = \min(\dim A,\ \dim B).$$

While this alternative formula may seem counterintuitive, general frameworks where it holds were uncovered, cf. e.g., [20, 40, 42] and references therein. This is notably commonly met by *limsup sets*, obtained as follows: There exists a collection of sets A_n such that A is the set of points that belong to an infinite number of the A_n.

This is particularly relevant for multifractal analysis where the singularity sets E_H^- defined in (45) often turn out to be of this type: It is the case for Lévy processes or random wavelet series, see e.g. [14, 56, 58]). For multivariate multifractal spectra, this leads to an alternative formula

$$\mathcal{D}_{(f_1,f_2)} = \min(\mathcal{D}_{f_1}(H_1), \mathcal{D}_{f_2}(H_2)) \tag{48}$$

expected to hold in competition with the codimension formula, at least for the sets E_H^-. The existence of two well motivated formulas in competition makes it hard to expect that general mathematical results could hold under fairly reasonable assumptions. Therefore, we now turn towards the construction of multifractal formalisms adapted to a multivariate setting, first in order to inspect if this approach can yield more intuition on the determination of multivariate spectra and, second, in order to derive new multifractality parameters which could be used for model selection and identification, and also in order to get some understanding on the ways that singularity sets of several functions are correlated.

In order to get some intuition in that direction, it is useful to start with a probabilistic interpretation of the multifractal quantities that were introduced in the univariate setting.

3.2 Probabilistic Interpretation of Scaling Functions

We consider the following probabilistic toy-model: We assume that, for a given j, the wavelet coefficients $(c_{j,k})_{k\in\mathbb{Z}}$ of the signal considered share a common law X_j and display short range memory, i.e. become quickly decorrelated when the wavelets $\psi_{j,k}$ and $\psi_{j,k'}$ are located far away (i.e. when $k - k'$ gets large); then, the wavelet structure functions (27) can be interpreted as an empirical estimation of $\mathbb{E}(|X_j|^p)$, i.e. the moments of the random variables X_j, and the wavelet scaling function characterizes the power law behaviour of these moments (as a function of the scale 2^{-j}). This interpretation is classically acknowledged for signals which display some stationarity, and the vanishing moments of the wavelets reinforce this decorrelation even if the initial process displays long range correlations, see e.g. the studies performed on classical models such as fBm ([1] and ref. therein). We will not discuss the relevance of this model; we just note that his interpretation has the advantage of pointing towards probabilistic tools when one shifts from one to several signals, and these tools will allow to introduce natural classification parameters which can then be used even when the probabilistic assumptions which led to their introduction have no reason to hold.

From now on, we consider two signals f_1 and f_2 defined on $\mathbb{R}$ (each one satisfying the above assumptions) with wavelet coefficients respectively $c^1_{j,k}$ and $c^2_{j,k}$. The "covariance" of the wavelet coefficients at scale j is estimated by the

empirical correlations

$$\text{for } m, n = 1, 2, \qquad S_{m,n}(j) = 2^{-j} \sum_k c^m_{j,k} c^n_{j,k}. \tag{49}$$

Log-log regressions of these quantities (as a function of $\log(2^{-j})$ allow to determine if some power-law behaviour of these auto-correlations (if $m = n$) and cross-correlations (if $m \neq n$) can be put in evidence: When these correlations are found to be significantly non-negative, one defines the *scaling exponents* $H_{m,n}$ implicitly by

$$S_{m,n}(j) \sim 2^{-H_{m,n} j}$$

in the limit of small scales. Note that, if $m = n$, the exponent associated with the auto-correlation simply is $\eta_f(2)$ and is referred to as the *Hurst exponent* of the data.

Additionally, the *wavelet coherence function* is defined as

$$C_{1,2}(j) = \frac{S_{1,2}(j)}{\sqrt{S_{1,1}(j) S_{2,2}(j)}}.$$

It ranges within the interval $[-1, 1]$ and quantifies, as a scale-dependent correlation coefficient, which scales are involved in the correlation of the two signals, see [7, 114].

Note that probabilistic denominations such as "auto-correlation", "cross-correlations" and "coherence function" are used even if no probabilistic model is assumed, and used in order to derive scaling parameters obtained by log-log plot regression which can prove powerful as classification tools.

As an illustration, we estimated these crosscorrelations concerning the following couples of data recorded on marathon runners: heart-beat frequency vs. cadence, and cadence vs. acceleration, see Fig. 9. In both cases, no correlation between

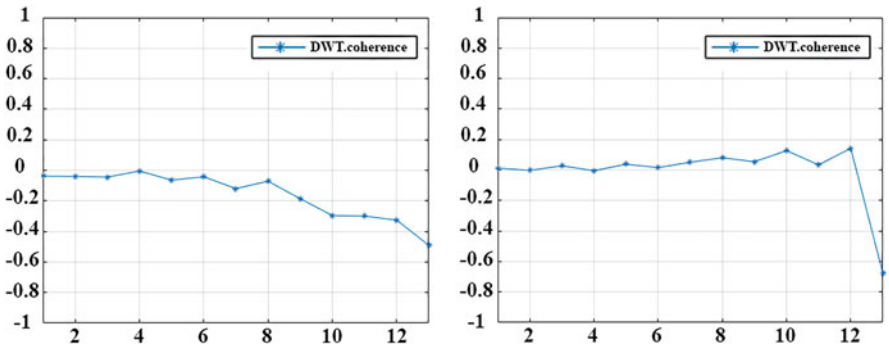


Fig. 9 Wavelet coherence between heart-beat frequency and cadence (left) and between acceleration and cadence (right)

the wavelet coefficients at a given scale is put in evidence. Therefore, this is a situation where the additional bonus brought by measuring multifractal correlations is needed. Indeed, if the cross-correlations of the signals do not carry substantial information, this does not imply that the singularity sets of each signal are not related (as shown by the example supplied by *Brownian motions in multifractal time*, see below in Sect. 3.4). In that case, a natural idea is to look for correlations that would be revealed by the multiscale quantities associated with pointwise exponents rather than by wavelet coefficients.

3.3 Multivariate Multifractal Formalism

The idea that leads to a multivariate multifractal formalism is quite similar as the one which led us from wavelet scaling functions to leaders and p-leaders scaling functions: One should incorporate in the cross-correlations the multiscale quantities which allow to characterize pointwise regularity, i.e. replace wavelet coefficients by wavelet leaders in (49).

Suppose that two pointwise regularity exponents h_1 and h_2 defined on $\mathbb{R}$ are given. We assume that each of these exponents can be derived from corresponding multiresolution quantities $d^1_{j,k}$, and $d^2_{j,k}$ according to (31). A *grandcanonical multifractal formalism* allows to estimate the joint spectrum $\mathcal{D}(H_1, H_2)$ of the couple of exponents (h_1, h_2) as proposed in [95]. In the general setting provided by multiresolution quantities, it is derived as follows: The *multivariate structure functions* associated with the couple $(d^1_{j,k}, d^2_{j,k})$ are defined by

$$\forall r = (r_1,\ r_2) \in \mathbb{R}^2, \qquad S(r,\, j) = 2^{-j} \sum_k (d^1_{j,k})^{r_1} (d^2_{j,k})^{r_2}, \tag{50}$$

see [6, 27] for the seminal idea of proposing such multivariate multiresolution quantities as building blocks of a *grandcanonical formalism*. Note that they are defined as a cross-correlation, which would be based on the quantities $d^1_{j,k}$ and $d^2_{j,k}$, with the extra flexibility of raising them to arbitrary powers, as is the case for univariate structure functions. The corresponding *bivariate scaling function* is

$$\zeta(r) = \liminf_{j \to +\infty} \frac{\log\left(S(r,\, j)\right)}{\log(2^{-j})}. \tag{51}$$

The *bivariate Legendre spectrum* is obtained through a 2-variable Legendre transform

$$\forall H = (H_1,\, H_2) \in \mathbb{R}^2, \qquad \mathcal{L}(H) = \inf_{r \in \mathbb{R}^2} (1 - \zeta(r) + H \cdot r), \tag{52}$$

where $H \cdot r$ denotes the usual scalar product in $\mathbb{R}^2$. Apart from [95], this formalism has been investigated in a wavelet framework for joint Hölder and oscillation exponents in [10], in an abstract general framework in [100], and on wavelet leader and p-leader based quantities in [6, 72].

Remark 6 The setting supplied by orthonormal wavelet bases is well fitted to be extended to the multivariate setting, because the multiresolution quantities d_λ are defined on a preexisting (dyadic) grid, which is shared by both quantities. Note that this is not the case for the WTMM, where the multiresolution quantities are defined at the local maxima of the continuous wavelet transform (see (17)), and these local maxima differ for different signals; thus, defining multivariate structure functions in this setting would lead to the complicated questions of matching these local maxima correctly in order to construct bivariate structure functions similar to (50).

The multivariate multifractal formalism is backed by only few mathematical results. A first reason is that, as already mentioned, the Legendre spectrum does not yield in general an upper bound for the multifractal spectrum, and this property is of key importance in the univariate setting. Another drawback is that, in contradistinction with the univariate case, the scaling function (51) has no function space interpretation. It follows that there exists no proper setting for genericity results except if one defines a priori this function space setting (as in [26, 27] where generic results are obtained in couples of function spaces endowed with the natural norm on a product space). We meet here once again the problem of finding a "proper" genericity setting that would be fitted to the quantities supplied by scaling functions. We now list several positive results concerning multivariate Legendre spectra.

The following result of [75] shows how to recover the univariate Legendre spectra from the bivariate one.

Proposition 5 *Let $d^1_{j,k}$, and $d^2_{j,k}$ be two multiresolution quantities associated with two pointwise exponents $h_1(x)$ and $h_2(x)$. The associated uni- and bi-variate Legendre spectra are related as follows:*

$$\mathcal{L}_1(H_1) = \sup_{H_2} \mathcal{L}(H_1, H_2) \quad \text{and} \quad \mathcal{L}_2(H_2) = \sup_{H_1} \mathcal{L}(H_1, H_2).$$

This property implies that results similar to Theorem 1 hold in the multivariate setting.

Corollary 1 *Let $d^1_{j,k}$ and $d^2_{j,k}$ be two multiresolution quantities associated with two pointwise exponents $h_1(x)$ and $h_2(x)$. The following results on the couple $(h_1(x), h_2(x))$ hold:*

- *If the bivariate Legendre spectrum has a unique maximum for $(H_1, H_2) = (c_1, c_2)$, then*

$$\text{for almost every } x, \qquad h_1(x) = c_1 \quad \text{and} \quad h_2(x) = c_2. \tag{53}$$

- *If the leader scaling function is affine then*

$$\exists (c_1, c_2), \quad \forall x, \qquad h_1(x) = c_1 \quad \textit{and} \quad h_2(x) = c_2.$$

Note that the fact that the leader scaling function is affine is equivalent to the fact that the bivariate Legendre spectrum is supported by a point. In that case, if the exponents h_1 and h_2 are associated with the functions f_1 and f_2, then they are monohölder functions.

Proof The first point holds because, if the bivariate Legendre spectrum has a unique maximum, then, its projections on the H_1 and the H_2 axes also have a unique maximum at respectively $H_1 = c_1$ and $H_2 = c_2$ and Proposition 5 together with Theorem 1 imply (53).

As regards the second statement, one can use Proposition 5: If the bivariate scaling function is affine, then $\mathcal{L}(H_1, H_2)$ is supported by a point, so that Proposition 5 implies that it is also the case for univariate spectra $\mathcal{L}(H_1)$ and $\mathcal{L}(H_2)$, and Theorem 1 then implies that h_1 is constant and the same holds for h_2. □

Recall that, in general, the bivariate Legendre spectrum does not yield an upper bound for the multifractal spectrum (in contradistinction with the univariate case), see [74] where a counterexample is constructed; this limitation raises many open questions: Is there another way to construct a Legendre spectrum which would yield an upper bound for $\mathcal{D}(H_1, H_2)$? which information can actually be derived from the Legendre spectrum? A first positive result was put in light in [74], where a notion of "compatibility" between exponents is put in light and is shown to hold for several models: When this property holds, then the upper bound property is satisfied. It is not clear that there exists a general way to check directly on the data if it is satisfied; however, an important case where it is the case is when the exponents derived are the Hölder exponent and one of the "second generation exponents" that we mentioned, see [6, 72]. In that case, the upper bound property holds, and it allows to conclude that the signal does not display e.g. oscillating singularities, an important issue both theoretical and practical. Let us mention a situation where this question shows up: In [18], P. Balanca showed the existence of oscillating singularities in the sample of some Lévy processes and also showed that they are absent in others (depending on the Lévy measure which is picked in the construction); however, he only worked out several examples, and settling the general case is an important issue; numerical estimations of such bivariate spectra could help to make the right conjectures in this case.

The general results listed in Corollary 1 did not require assumptions on correlations between the exponents h_1 and h_2. We now investigate the implications of such correlations on the joint Legendre spectrum. For that purpose, let us come back to the probabilistic interpretation of the structure functions (50) in terms of cross-correlation of the $(d^1_{j,k})^{r_1}$ and $(d^2_{j,k})^{r_2}$. As in the univariate case, if we assume that, for a given j, the multiresolution quantities $d^1_{j,k}$ and $d^2_{j,k}$ respectively share common

laws X_j^1 and X_j^2 and display short range memory, then (50) can be interpreted as an empirical estimation of $\mathbb{E}(|X_j^1|^{r_1}|X_j^2|^{r_2})$. If we additionally assume that the $(d_{j,k}^1)$ and $(d_{j,k}^2)$ are independent, then we obtain

$$S(r, j) = \mathbb{E}(|X_j^1|^{r_1}|X_j^2|^{r_2}) = \mathbb{E}(|X_j^1|^{r_1}) \cdot \mathbb{E}(|X_j^2|^{r_2}),$$

which can be written

$$S(r_1, r_2, j) = S^1(r_1, j)S^2(r_2, j). \tag{54}$$

Assuming that lim inf in (51) actually is a limit, we obtain $S(r_1, r_2, j) \sim 2^{-(\zeta^1(r_1)+\zeta^2(r_2))j}$ yielding $\zeta(r_1, r_2) = \zeta^1(r_1) + \zeta^2(r_2)$. Applying (52), we get

$$\mathcal{L}(H_1, H_2) = \inf_{(r_1,r_2)\in\mathbb{R}^2} (1 - \zeta^1(r_1) + \zeta^2(r_2) + H_1 r_1 + H_2 r_2)$$

$$= \inf_{r_1}(1 - \zeta^1(r_1) + H_1 r_1) + \inf_{r_2}(1 - \zeta^2(r_2) + H_2 r_2) - 1,$$

which leads to

$$\mathcal{L}(H_1, H_2) = \mathcal{L}(H_1) + \mathcal{L}(H_2) - 1. \tag{55}$$

Thus, under stationarity and independence, the codimension rule applies for the multivariate Legendre spectrum. In practice, this means that any departure of the Legendre spectrum from (55), which can be checked on real-life data, indicates that one of the assumptions required to yield (55) (either stationarity or independence) does not hold (Fig. 10).

As a byproduct, we now show that multivariate multifractal analysis can give information on the nature of the singularities of *one* signal, thus complementing results such as Proposition 4 which yielded almost everywhere information of this type. Let us consider the joint multifractal spectrum of a function f and its fractional integral of order s, denoted by $f^{(-s)}$. If f only has canonical singularities, then the Hölder exponent of $f^{(-s)}$ satisfies $\forall x_0$, $h_{f^{(-s)}}(x_0) = h_f(x_0) + s$, so that the joint Legendre spectrum is supported by the line $H_2 = H_1 + s$. In that case, the synchronicity assumption is satisfied and one can conclude that the joint multifractal spectrum is supported by the same segment; a contrario, a joint Legendre spectrum which is not supported by this line is interpreted as the signature of *oscillating singularities* in the data, as shown by the discussion above concerning the cases where the upper bound for bivariate spectra holds. Figs. 11, 12, and 13 illustrate this use of bivariate multifractal analysis: In each case, a signal and its primitive are jointly analyzed: The three signals are collected on the same runner and the whole race is analyzed. Figure 11 shows the analysis of heartbeat, Fig. 12 shows the cadence and Fig. 13 shows the acceleration. In the first case, the analysis is performed directly on the data using a p-exponent with $p = 1$, whereas, for the

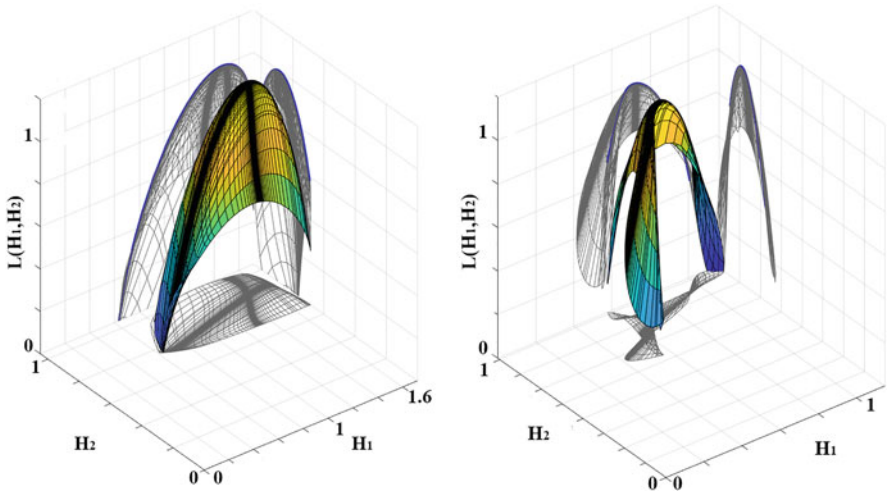


Fig. 10 On the left, the bivariate multifractal spectrum between heart-beat frequency primitive and cadence primitive are shown, and, on the right, the bivariate multifractal spectrum between acceleration and cadence with fractional integral of order 1.5 are shown. This demonstrates the strong correlation between the pointwise singularities of the two data: indeed the bivariate spectra are almost carried by a segment, and a bivariate spectrum carried by a line $H_2 = aH_1 + b$ indicates a perfect match between the pointwise exponents according to the same relationship: $\forall x$, $h_2(x) = ah_1(x) + b$

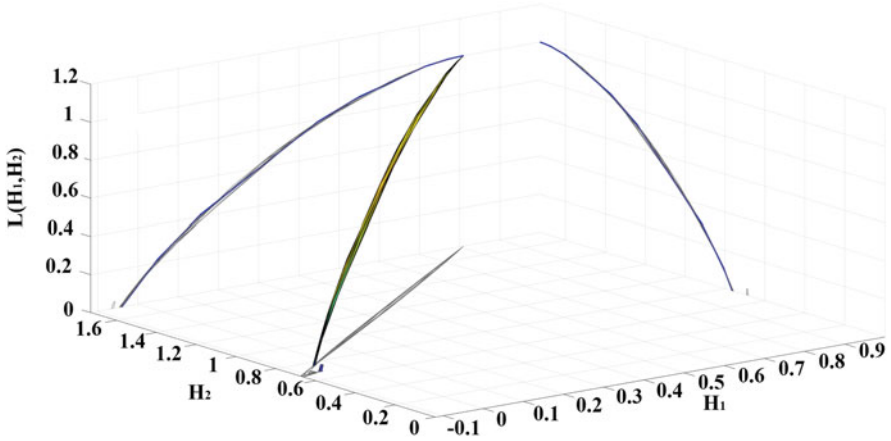


Fig. 11 Bivariate 1-spectrum of heartbeat frequency and its primitive: the bivariate spectrum lines up perfectly along the line $H_2 = H_1 + 1$

two last ones, the analysis is performed on a fractional integral of order $1/2$. In each case, the results yield a bivariate Legendre spectrum supported by the segment $H_2 = H_1 + s$, which confirms the almost everywhere results obtained in Sect. 2.9: The data only contain canonical singularities.

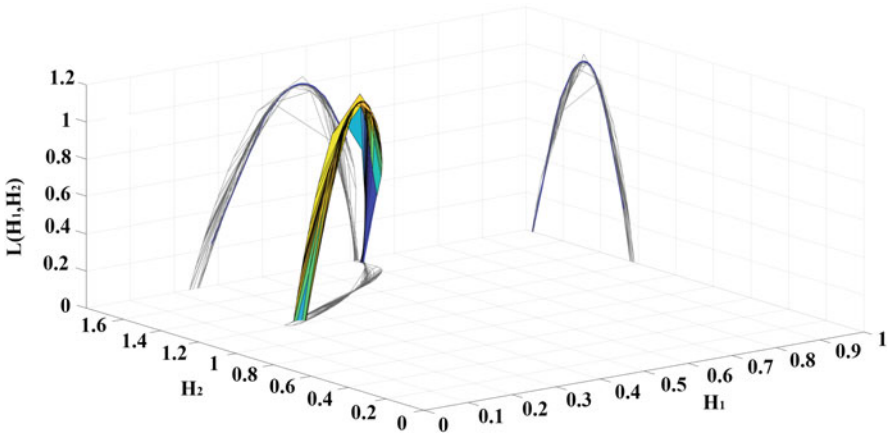


Fig. 12 Bivariate Hölder spectrum of fractional integrals order $1/2$ and $3/2$ of cadence: the bivariate spectrum lines up perfectly along the line $H_2 = H_1 + 1$

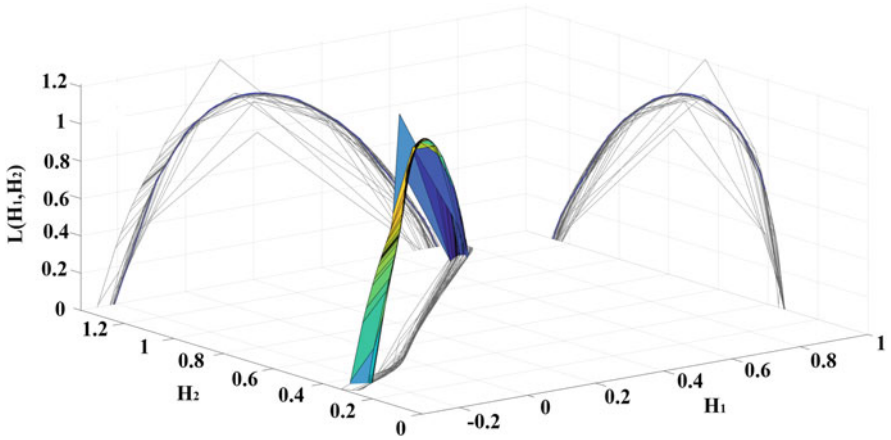


Fig. 13 Bivariate Hölder spectrum of fractional integrals of order $1/2$ and $3/2$ of acceleration: the bivariate spectrum lines up perfectly along the line $H_2 = H_1 + 1$

3.4 Fractional Brownian Motions in Multifractal Time

In order to put in light the additional information between wavelet correlations and bivariate scaling functions (and the associated Legendre spectrum), we consider the model supplied by Brownian motion in multifractal time, which has been proposed by B. Mandelbrot [36, 90] as a simple model for financial time series: Instead of the classical Brownian model $B(t)$, he introduced a time change (sometimes referred to as a *subordinator*)

$$B(f(t)) = (B \circ f)(t)$$

where the irregularities of f model the fluctuations of the intrinsic "economic time", and typically is a multifractal function. In order to be a "reasonable" time change, the function f has to be continuous and strictly increasing; such functions usually are obtained as distribution functions of probability measures $d\mu$ supported on $\mathbb{R}$ (or on an interval), and which have no atoms (i.e. $\forall a \in \mathbb{R}$, $\mu(a) = 0$); typical examples are supplied by deterministic or random cascades, and this is the kind of models that were advocated by B. Mandelbrot in [90]. Such examples will allow to illustrate the different notions that we introduced, and the additional information which is put into light by the bivariate Legendre spectrum and is absent from wavelet correlations.

Let us consider the slightly more general setting of one fBm of Hurst exponent α (the case of Brownian motion corresponds to $\alpha = 1/2$) modified by a time change f. In order to simplify its theoretical multifractal analysis, we take for pointwise regularity exponent the Hölder exponent and we make the following assumptions on f: We assume that it has only canonical singularities and that, if they exist, the non-constant terms of the Taylor polynomial of f vanish at every point even if the Hölder exponent at some points is larger than 1 (this is typically the case for primitives of singular measures). In that case, classical uniform estimates on increments of fBm, see [77] imply that

$$\text{a.s.} \quad \forall t, \qquad h_{B\circ f}(t) = \alpha h_f(t), \tag{56}$$

so that

$$\text{a.s.} \quad \forall H, \qquad \mathcal{D}_{B\circ f}(H) = \mathcal{D}_f(H/\alpha);$$

Note that the simple conclusion (56) may fail if the Taylor polynomial is not constant at every point, as shown by the simple example supplied by $f(x) = x$ on the interval $[0, 1]$.

We now consider $B_1 \circ f$ and $B_2 \circ f$: two independent fBm modified by the *same deterministic time change* f (with the same assumptions as above). It follows from (56) that, with probability 1, the Hölder exponents of $B_1 \circ f$ and $B_2 \circ f$ coincide everywhere, leading to the following multifractal spectrum, which holds almost surely:

$$\begin{cases} \text{if } H_1 = H_2,\ \mathcal{D}_{(B_1\circ f, B_2\circ f)}(H_1, H_2) = \mathcal{D}_f\left(\dfrac{H_1}{\alpha}\right) \\ \\ \text{if } H_1 \neq H_2,\ \mathcal{D}_{(B_1\circ f, B_2\circ f)}(H_1, H_2) = -\infty. \end{cases} \tag{57}$$

Figure 14 gives a numerical backing of this result: The Legendre spectrum numerically obtained corresponds to the theoretical multifractal spectrum. Let us give a non-rigourous argument which backs this result: The absence of oscillating singularities in the data implies that the maxima in the wavelet leaders are attained for a λ' close to λ, so that the wavelet leaders of a given magnitude will be close to coincide for both processes, and therefore the bivariate structure functions (50)

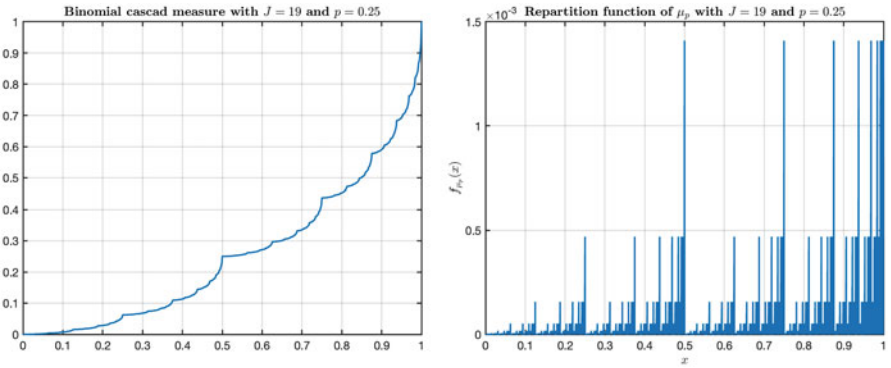


Fig. 14 Binomial measure with $p = 1/4$ (left) and its repartition function (right) which is used as the time change in Fig. 16

satisfy

$$S_f(r, j) = 2^{-j} \sum\nolimits_{\lambda \in \Lambda_j} (d_\lambda^1)^{r_1} (d_\lambda^2)^{r_2} \sim 2^{-dj} \sum\nolimits_{\lambda \in \Lambda_j} (d_\lambda^1)^{r_1 + r_2}$$

so that

$$\text{a.s.}\,, \quad \forall r_1, r_2, \qquad \tilde{\zeta}(r_1, r_2) = \zeta(r_1 + r_2).$$

where $\tilde{\zeta}$ is the bivariate scaling function of the couple $(B_1 \circ f, B_2 \circ f)$ and ζ is the univariate scaling function of $B_1 \circ f$. Taking a Legendre transform yields that the bivariate Legendre spectrum $\mathcal{L}(H_1, H_2)$ also satisfies a similar formula as (57), i.e.

$$\text{a.s.}\,, \quad \forall H_1, H_2, \begin{cases} \text{if } H_1 = H_2, \ \mathcal{L}_{(B_1 \circ f, B_2 \circ f)}(H_1, H_2) = \mathcal{L}_f\left(\dfrac{H_1}{\alpha}\right) \\ \\ \text{if } H_1 \neq H_2, \ \mathcal{L}_f\,(H_1, H_2) = -\infty. \end{cases} \tag{58}$$

Let us now estimate the wavelet cross correlations. Since f is deterministic, the processes $B_1 \circ f$ and $B_2 \circ f$ are two independent centered Gaussian processes. Their wavelet coefficients $c^1_{j,k}$ and $c^2_{j,k}$ therefore are independent centered Gaussians, and, at scale j the quantity

$$S_{m,n}(j) = 2^{-j} \sum_k c^1_{j,k} c^2_{j,k}$$

is an empirical estimation of their covariance, and therefore vanishes (up to small statistical fluctuation). In contradistinction with the bivariate spectrum, the wavelet

cross correlations reveal the decorrelation of the processes but does not yield information of the correlation of the singularity sets.

In order to illustrate these results, we will use for time change the distribution function of a binomial cascade μ_p carried on $[0, 1]$. Let $p \in (0, 1)$; μ_p is the only probability measure on $[0, 1]$ defined by recursion as follows: Let $\lambda \subset [0, 1]$ be a dyadic interval of length 2^{-j}; we denote by λ^+ and λ^- respectively its two "children" of length 2^{-j-1}, λ^+ being on the left and λ^- being on the right. Then, μ_p is the only probability measure carried by $[0, 1]$ and satisfying

$$\mu_p(\lambda^+) = p \cdot \mu_p(\lambda) \quad \text{and} \quad \mu_p(\lambda^-) = (1 - p) \cdot \mu_p(\lambda).$$

Then the corresponding time change is the function

$$\forall x \in [0, 1] \qquad f_{\mu_p}(x) = \mu_p([0, x]).$$

In Fig. 14, we show the binomial cascade $\mu_{1/4}$ and its distribution function, and in Fig. 15 the cross-correlation between two, independant realizations of this process is displayed; and in Fig. 16 we use this time change composed with a fBm of Hurst exponent $\alpha = 0.3$ (Fig. 17).

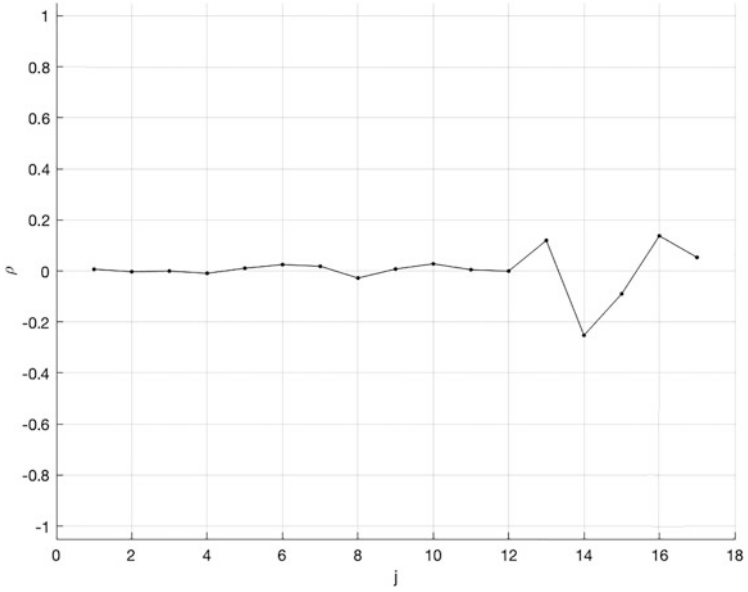


Fig. 15 Cross-correlation of the wavelet coefficients of two independent fBm with the same time change: the distribution function of the binomial measure μ_p with $p = 1/4$. The cross-correlation reflects the independence of the two processes

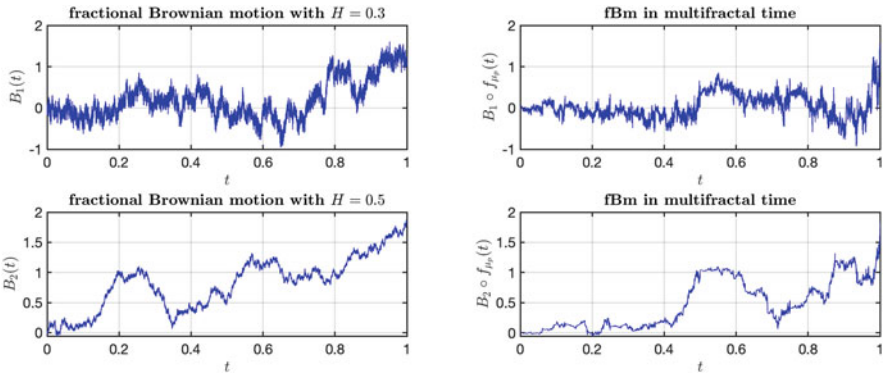


Fig. 16 fBm with $H = 0.3$ and $H = 0.5$ subordinated by the multifractal time change supplied by the distribution function of the binomial measure μ_p with $p = 1/4$

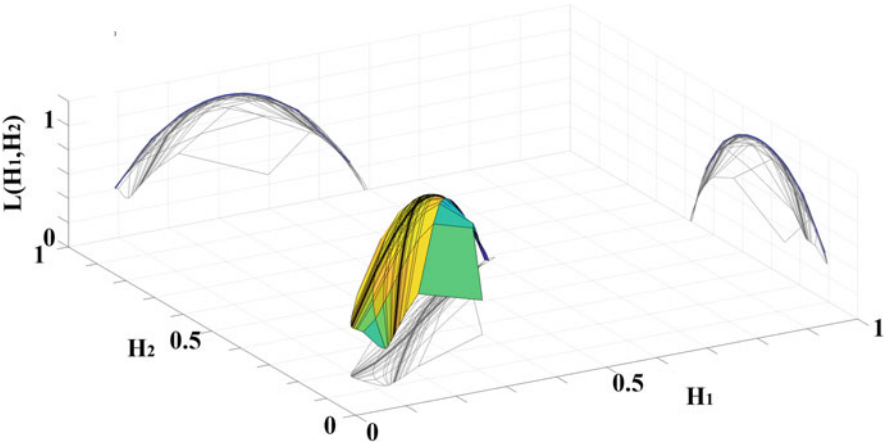


Fig. 17 Bivariate multifractal spectrum of two independent fBm with the same time change: the distribution function of a binomial measure with $p = 1/4$; in contradistinction with the cross-correlation of wavelet coefficients, the wavelet leaders are strongly correlated, leading to a bivariate multifractal Legendre spectrum theoretically supported by the line $H_1 = H_2$, which is close to be the case numerically

Remark 7 The fact that the same time change is performed does not play a particular role for the estimation of the wavelet cross-correlations; the same result would follow for two processes $B_1 \circ f$ and $B_2 \circ g$ with B_1 and B_2 independent, and where f and g are two deterministic time changes. Similarly, B_1 and B_2 can be replaced by two (possibly different) centered Gaussian processes.

Let us mention at this point that the mathematical problem of understanding what is the multifractal spectrum of the composition $f \circ g$ of two multifractal functions f and g, where g is a *time subordinator* i.e. an increasing function, is a largely

open problem (and is posed here in too much generality to find a general answer). This problem was initially raised by B. Mandelbrot and also investigated R. Riedi [102] who worked out several important subcases; see also the article by S. Seuret [105], who determined a criterium under which a function can be written as the composition of a time subordinator and a monohölder function, and [21] where J. Barral and S. Seuret studied the multifractal spectrum of a Lévy process, under a time subordinator given by the repartition function of a multifractal cascade.

3.5 *Multivariate Analysis of Marathon Physiological Data*

Let us consider one of the marathon runners, and denote his heart beat frequency by f_f and his cadence by f_c and by $f_f^{(-1)}$ and $f_c^{(-1)}$ their primitives. We performed the computation of the bivariate scaling function $\zeta_{f_f^{(-1)},f_c^{(-1)}}$ (using wavelet leaders) and we show its Legendre transform $\mathcal{L}_{f_f^{(-1)},f_c^{(-1)}}$ on Fig. 18. This spectrum is widely spread, in strong contradistinction with the bivariate spectra obtained in the previous section; this indicates that no clear correlations between the Hölder singularities of the primitives can be put in evidence. Figure 5 shows the two corresponding univariate spectra (which can be either computed directly, or obtained as projections of the bivariate spectrum).

In order to test possible relationships between the bivariate spectrum and the two corresponding univariate spectra, we compute the difference

$$\mathcal{L}(H_1, H_2) - \mathcal{L}(H_1) - \mathcal{L}(H_2) + 1,$$

which allows to test the validity of (55) and

$$\mathcal{L}_{(f_1,f_2)} - \min(\mathcal{L}_{f_1}(H_1), \mathcal{L}_{f_2}(H_2)),$$

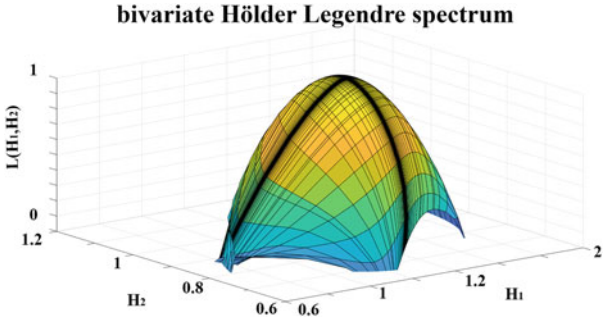


Fig. 18 Representation of the bivariate Hölder Legendre spectrum of the primitives of heart beat frequency and cadence: this bivariate spectrum is derived from the same data that were used to derive the two univariate spectra shown in Fig. 5

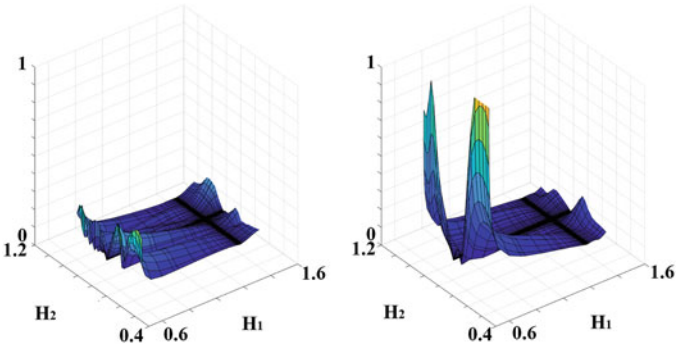


Fig. 19 Representation of the difference of the bivariate spectrum and the two formulas proposed in (54) and (47). The graph on the left is closer to zero, which suggests that the large intersection formula seems more appropriate in this case

which allows to test the validity of (48), they are shown in Fig. 19. This comparison suggests that the large intersection formula is more appropriate than the codimension formula in this case. Keeping in mind the conclusions of Sect. 3.1, these results indicate that an hypothesis of both stationarity and independence for each signals is inappropriate (indeed this would lead to the validity of the codimension formula), and on the opposite, these results are compatible with a pointwise regularity yielded by a limsup set procedure, as explained in Sect. 3.1.

4 Conclusion

Let us give a summary of the conclusions that can be drawn from a bivariate multifractal analysis of data based on the Legendre transform method. This analysis goes beyond the (now standard) technique of estimating correlations of wavelet coefficients; indeed here wavelet coefficients are replaced by wavelet leaders, which leads to new scaling parameters on which classification can be performed. On the mathematical side, even if the relationship between the Legendre and the multifractal spectra is not as clear as in the univariate case, nonetheless, situations have been identified where this technique can either yield information on the nature of the singularities (e.g. the absence of oscillating singularities), or on the type of processes that can be used to model the data (either of additive or of multiplicative type). In the particular case of marathon runners, the present study shows a bivariate spectrum between heart rate and cadence are related by the large intersection formula. In a recent study [31] a multivariate analysis revealed that, for all runners, RPE and respiratory frequency measured on the same runners during the marathon were close (their angle is acute on correlation circle of a principal component analysis) while the speed was closer to the cadence and to the Tidal respiratory volume at each inspiration and expiration). The sampling frequency of

the respiratory parameters did not allow to apply the multifractal analysis which here reveals that the cadence and heart rate could be an additive process such as, possibly a generalization of a Lévy process. Heart rate and cadence are under the autonomic nervous system control and Human beings optimize their cadence according his speed for minimizing his energy cost of running. Therefore, we can conclude that is not recommended to voluntarily change the cadence and this bivariate multifractal analysis mathematically shows that the cadence and heart rate are not only correlated but we can conjecture that they can be modeled by an additive process until the end of the marathon.

Acknowledgments We thank the anonymous referee for many stimulating and insightful remarks and suggestions on the first version of this paper.

References

1. Abry, P., Gonçalvès, P., Flandrin, P.: Wavelets, Spectrum Estimation and $1/f$ Processes, chapter 103. Wavelets and Statistics, Lecture Notes in Statistics. Springer, New York (1995)
2. Abry, P., Wendt, H., Jaffard, S., Helgason, H., Goncalvès, P., Pereira, E., Gharib, C., Gaucherand, P., Doret, M.: Methodology for multifractal analysis of heart rate variability: From lf/hf ratio to wavelet leaders. In: 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology, Buenos Aires, Argentina (2010)
3. Abry, P., Jaffard, S., Wendt, H.: When Van Gogh meets Mandelbrot: Multifractal classification of painting's texture. Signal Process. **93**(3), 554–572 (2013)
4. Abry, P., Jaffard, S., Wendt, H.: A bridge between geometric measure theory and signal processing: Multifractal analysis. In: Gröchenig, K., et al. (eds.) Operator-Related Function Theory and Time-Frequency Analysis, The Abel Symposium 2012, vol. 9, pp. 1–56 (2015)
5. Abry, P., Jaffard, S., Wendt, H.: Irregularities and scaling in signal and image processing: Multifractal analysis. In: Frame, M., Cohen, N. (eds.) Benoit Mandelbrot: A Life in Many Dimensions, pp. 31–116. World scientific publishing (2015)
6. Abry, P., Jaffard, S., Leonarduzzi, R., Melot, C., Wendt, H.: New exponents for pointwise singularity classification. In: Seuret, S., Barral, J. (eds.) Recent Developments in Fractals and Related Fields: Proc. Fractals and Related Fields III, 19–26 September 2015, Porquerolles, France, pp. 1–37 (2017)
7. Abry, P., Wendt, H., Jaffard, S., Didier, G.. Multivariate scale-free temporal dynamics: From spectral (Fourier) to fractal (wavelet) analysis. C. R. Acad. Sci. **20**(5), 489–501 (2019)
8. Arneodo, A., Bacry, E., Muzy, J.F.: The thermodynamics of fractals revisited with wavelets. Physica A **213**(1–2), 232–275 (1995)
9. Arneodo, A., Baudet, C., Belin, F., Benzi, R., Castaing, B., Chabaud, B., Chavarria, R., Ciliberto, S., Camussi, R., Chillà, F., Dubrulle, B., Gagne, Y., Hebral, B., Herweijer, J., Marchand, M., Maurer, J., Muzy, J.F., Naert, A., Noullez, A., Peinke, J., Roux, S.G., Tabeling, P., van der Water, W., Willaime, H.: Structure functions in turbulence, in various flow configurations, at Reynolds number between 30 and 5000, using extended self-similarity. Europhys. Lett. **34**, 411–416 (1996)
10. Arneodo, A., Bacry, E., Jaffard, S., Muzy, J.F.: Singularity spectrum of multifractal functions involving oscillating singularities. J. Fourier Anal. Appl. **4**, 159–174 (1998)
11. Arneodo, A., Audit, B., Decoster, N., Muzy, J.-F., Vaillant, C.: Wavelet-based multifractal formalism: applications to dna sequences, satellite images of the cloud structure and stock market data. In: Bunde, A., Kropp, J., Schellnhuber, H.J. (eds.) The Science of Disasters, pp. 27–102. Springer (2002)

12. Arneodo, A., Decoster, N., Kestener, P., Roux, S.G.: A wavelet-based method for multifractal image analysis: from theoretical concepts to experimental applications. In: Hawkes, P.W., Kazan, B., Mulvey, T. (eds.) Advances in Imaging and Electron Physics, vol.126, pp. 1–98. Academic Press (2003)
13. Aubry, J.-M.: On the rate of pointwise divergence of Fourier and wavelet series in L^p. J. Approx. Theory **538**, 97–111 (2006)
14. Aubry, J.M., Jaffard, S.: Random wavelet series. Commun. Math. Phys. **227**(3), 483–514 (2002)
15. Ayache, A.: On the monofractality of many stationary continuous gaussian fields. J. Funct. Anal. **281**, 109111 (2021)
16. Ayache, A., Jaffard, S.: Hölder exponents of arbitrary functions. Rev. Mat. Iber. **26**, 77–89 (2010)
17. Bacry, E., Kozhemyak, A., Muzy, J.F.: Multifractal models for asset prices. In: Encyclopedia of Quantitative Finance. Wiley (2010)
18. Balanca, P.: Fine regularity of Lévy processes and linear (multi)fractional stable motion. Electron. J. Probab. **101**, 1–37 (2014)
19. Bardet, J.-M.: Statistical study of the wavelet analysis of fractional Brownian motion. EEE Trans. Inform. Theory **48**, 991–999 (2002)
20. Barral, J., Seuret, S.: A heterogeneous ubiquitous systems in R^d and Hausdorff dimensions. Bull. Braz. Math. Soc. **38**(3), 467–515 (2007)
21. Barral, J., Seuret, S.: The singularity spectrum of Lévy processes in multifractal time. Adv. Math. **14**(1), 437–468 (2007)
22. Barral, J., Seuret, S.: Besov spaces in multifractal environment, and the Frisch-Parisi conjecture. Preprint (2021)
23. Barreira, L., Saussol, B.: Variational principles and mixed multivariate spectra. Trans. A. Math. Soc. **353**(10), 3919–3944 (2001)
24. Barreira, L., Saussol, B., Schmeling, J: Higher-dimensional multifractal analysis. J. Math. Pures Appl. **81**, 67–91 (2002)
25. Bayart, F., Heurteaux, Y.: Multifractal analysis of the divergence of Fourier series. Ann. Sci. ENS **45**, 927–946 (2012)
26. Ben Abid, M.: Prevalent mixed Hölder spectra and mixed multifractal formalism in a product of continuous Besov spaces. Nonlinearity **30**, 3332–3348 (2017)
27. Ben Slimane, M.: Baire typical results for mixed Hölder spectra on product of continuous Besov or oscillation spaces. Mediterr. J. Math. **13**, 1513–1533 (2016)
28. Berndsen, J., Lawlor, A., Smyth, B.: Exploring the wall in marathon running. J. Sports Anal. **6**, 173–1860 (1978)
29. Billat, V., Mille-Hamard, L., Meyer, Y., Wesfreid, E.: Detection of changes in the fractal scaling of heart rate and speed in a marathon race. Phys. A 3798–3808 (1997)
30. Billat, V.L., Palacin, F., Correa, M., Pycke, J.R.: Pacing strategy affects the sub-elite marathoner's cardiac drift and performance. Front. Psychol. **10**, 3026 (2020)
31. Billat, V.L., Petot, H., Landrain, M., Meilland, R., Koralsztein, J.-P., Mille-Hamard, L.: Cardiac output and performance during a 571 marathon race in middle-aged recreational runners. Sci. World J. **19**(4), 810–859 (2012)
32. Broucke, F., Vindas, J.: The pointwise behavior of Riemann's function. To appear J. Fract. Geom (2023)
33. Brown, G., Michon, G., Peyrière, J.: On the multifractal analysis of measures. J. Stat. Phys. **66**(3–4), 775–790 (1992)
34. Buczolich, Z., Nagy, J.: Hölder spectrum of typical monotone continuous functions. Real Anal. Exchange **26**(2), 133–156 (2000)
35. Calderón, A.P., Zygmund, A.: Local properties of solutions of elliptic partial differential equations. Stud. Math. **20**, 171–223 (1961)
36. Calvet, L., Fisher, A., Mandelbrot, B.: The multifractal model of asset returns. In: Cowles Foundation Discussion Papers: 1164 (1997)

37. Catrambone, V., Valenza, G., Scilingo, E.P., Vanello, N., Wendt, H., Barbieri, R., Abry, P.: Wavelet p-leader non-gaussian multiscale expansions for eeg series: an exploratory study on cold-pressor test. In: International IEEE EMBS Conference (EMBC), Berlin, Germany, July (2019)
38. Christensen, J.: On sets of Haar measure zero in abelian polish groups. Israel J. Math. **13**(3), 255–260 (1972)
39. Daoudi, K., Lévy-Véhel, J., Meyer, Y.: Construction of continuous functions with prescribed local regularity. Constr. Approx. **14**, 349–385 (1998)
40. Durand, A.: Describability via ubiquity and eutaxy in Diophantine approximation. Ann. Math. Blaise Pascal **22**, 1–149 (2015)
41. Esser, C., Loosveld, L.: Slow, ordinary and rapid points for Gaussian wavelets series and application to fractional brownian motions. Preprint (2021)
42. Falconer, K.: Fractal Geometry: Mathematical Foundations and Applications. John Wiley & Sons, West Sussex (1993)
43. Fan, A.H., Liao, L., Ma, J.-H.: Level sets of multiple ergodic averages. Monatsh. Math. **168**, 17–26 (2012)
44. Fan, A.-H., Liao, L., Wy, M.: Multifractal analysis of some multiple ergodic averages in linear cookie-cutter dynamical systems. Math. Z. **290**, 63–81 (2018)
45. Flandrin, P.: Explorations in Time-Frequency Analysis. Cambridge University Press (2018)
46. Frankhauser, P.: The fractal approach. a new tool for the spatial analysis of urban agglomerations. In: Population: An English Selection, pp. 205–240 (1998)
47. Fraysse, A.: Regularity criteria of almost every function in a Sobolev space. J. Funct. Anal. **258**, 1806–1821 (2010)
48. Fraysse, A., Jaffard, S.: How smooth is almost every function in a Sobolev space? Rev. Mat. Iber. **22**(2), 663–682 (2006)
49. Frisch, U.: Turbulence, the Legacy of A.N. Kolmogorov. Addison-Wesley (1993)
50. Galaska, R., Makowiec, D., Dudkowska, A., Koprowski, A., Chlebus, K., Wdowczyk-Szulc, J., Rynkiewicz, A.: Comparison of wavelet transform modulus maxima and multifractal detrended fluctuation analysis of heart rate in patients with systolic dysfunction of left ventricle. Ann. Noninvasive Electrocardiol. **13**(2), 155–164 (2008)
51. Ivanov, P.C., Nunes Amaral, L.A., Goldberger, A.L., Havlin, S., Rosenblum, M.G., Struzik, Z.R., Stanley, H.E.: Multifractality in human heartbeat dynamics. Nature **399**, 461–465 (1999)
52. Jaffard, S.: Construction de fonctions multifractales ayant un spectre de singularités prescrit. C. R. Acad. Sci. **315**(5), 19–24 (1992)
53. Jaffard, S.: Functions with prescribed Hölder exponent. Appl. Comput. Harmonic Anal. **2**, 400–401 (1995)
54. Jaffard, S.: The spectrum of singularities of Riemann's function,. Rev. Mat. Iber. **12**, 441–460 (1996)
55. Jaffard, S.: Multifractal formalism for functions. SIAM J. Math. Anal. **28**(4), 944–998 (1997)
56. Jaffard, S.: The multifractal nature of Lévy processes. Prob. Theory Related Fields **114**(2), 207–227 (1999)
57. Jaffard, S.: Construction of functions with prescribed Hölder and chirps exponents. Rev. Mat. Iber. **16**(2), 331–349 (2000)
58. Jaffard, S.: On lacunary wavelet series. Ann. Appl. Probab. **10**(1), 313–329 (2000)
59. Jaffard, S.: On the Frisch-Parisi conjecture. J. Math. Pures Appl. **79**(6), 525–552 (2000)
60. Jaffard, S.: On Davenport expansions. In: Fractal Geometry and Applications: A Jubilee of Benoit Mandelbrot - Analysis, Number Theory, and Dynamical Systems, Pt 1, vol. 72, pp. 273–303 (2004)
61. Jaffard, S.: Wavelet techniques in multifractal analysis. In: Lapidus, M., van Frankenhuijsen, M. (eds.) Fractal Geometry and Applications: A Jubilee of Benoît Mandelbrot, Proc. Symp. Pure Math., vol. 72(2), pp. 91–152. AMS (2004)
62. Jaffard, S.: Beyond Besov spaces, part 2: Oscillation spaces. Constr. Approx. **21**(1), 29–61 (2005)

63. Jaffard, S.: Pointwise regularity associated with function spaces and multifractal analysis. In: Figiel, T., Kamont, A. (eds.) Banach Center Pub. Vol. 72 Approximation and Probability, pp. 93–110 (2006)
64. Jaffard, S.: Wavelet techniques for pointwise regularity. Ann. Fac. Sci. Toul. **15**(1), 3–33 (2006)
65. Jaffard, S. Esser, C.: Divergence of wavelet series: a multifractal analysis. Adv. Math. **328**, 928–958 (2018)
66. Jaffard, S., Martin, B.: Multifractal analysis of the Brjuno function. Invent. Math. **212**, 109–132 (2018)
67. Jaffard, S., Melot, C.: Wavelet analysis of fractal boundaries. Commun. Math. Phys. **258**(3), 513–565 (2005)
68. Jaffard, S., Meyer, Y.: Wavelet methods for pointwise regularity and local oscillations of functions. Mem. Am. Math. Soc. **123**, 587 (1972)
69. Jaffard, S., Lashermes, B., Abry, P.: Wavelet leaders in multifractal analysis. In: Qian, T., Vai, M.I., Yuesheng, X. (eds.) Wavelet Analysis and Applications, pp. 219–264. Birkhäuser, Basel (2006)
70. Jaffard, S., Abry, P., Roux, S., Vedel, B., Wendt, H.: The Contribution of Wavelets in Multifractal Analysis, pp. 51–98. Higher Education Press, Series in Contemporary Applied Mathematics, World Scientific Publishing China (2010)
71. Jaffard, S., Abry, P., Roux, S.G.: Function spaces vs. scaling functions: tools for image classification. In: Bergounioux, M. (ed.) Mathematical Image processing (Springer Proceedings in Mathematics) , vol. 5, pp. 1–39 (2011)
72. Jaffard, S., Abry, P., Melot, C., Leonarduzzi, R., Wendt, H.: Multifractal analysis based on p-exponents and lacunarity exponents. In: Bandt, C., et al. (eds.) Fractal Geometry and Stochastics V, Series Progress in Probability, Birkhäuser, vol. 70, pp. 279–313 (2015)
73. Jaffard, S., Melot, C., Leonarduzzi, R., Wendt, H., Roux, S.G., Torres, M.E., Abry, P.: p-exponent and p-leaders, Part I: Negative pointwise regularity. Physica A **448**, 300–318 (2016)
74. Jaffard, S., Seuret, S., Wendt, H., Leonarduzzi, R., Abry, P.: Multifractal formalisms for multivariate analysis. Proc. R. Soc. A **475**, 2229 (2019)
75. Jaffard, S., Seuret, S., Wendt, H., Leonarduzzi, R., Roux, S., Abry, P.: Multivariate multifractal analysis. Appl. Comput. Harmonic Anal. **46**(3), 653–663 (2019)
76. Johnson, C.R., Messier, P., Sethares, W.A., Klein, A.G., Brown, C., Do, A.H., Klausmeyer, P., Abry, P., Jaffard, S., Wendt, H., Roux, S., Pustelnik, N., van Noord, N., van der Maaten, L., Potsma, E., Coddington, J., Daffner, L.A., Murata, H., Wilhelm, H., Wood, S., Messier, M.: Pursuing automated classification of historic photographic papers from raking light photomicrographs. J. Am. Inst. Conserv. **53**(3), 159–170 (2014)
77. Kahane, J.-P.: Some Random Series of Functions. Cambridge University Press (1985)
78. Kantelhardt, J.W., Zschiegner, S.A., Koscielny-Bunde, E., Havlin, S., Bunde, A., Stanley, H.E.: Multifractal detrended fluctuation analysis of nonstationary time series. Physica A **316**(1), 87–114 (2002)
79. Kolmogorov, A.N.: The Wiener spiral and some other interesting curves in Hilbert space (Russian). Dokl. Akad. Nauk SSSR **26**(2), 115–118 (1940)
80. Kolmogorov, A.N.: The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers. C. R. Acad. Sci. De L'Urss **30**, 301–305 (1941)
81. Lashermes, B., Jaffard, S., Abry, P.: Wavelet leader based multifractal analysis. In: 2005 Ieee International Conference on Acoustics, Speech, and Signal Processing, vols. 1–5, pp. 161–164 (2005)
82. Lashermes, B., Roux, S.G., Abry, P., Jaffard, S.: Comprehensive multifractal analysis of turbulent velocity using the wavelet leaders. Eur. Phys. J. B **61**(2), 201–215 (2008)
83. Lashermes, B., Roux, S.G., Abry, P., Jaffard, S.: Comprehensive multifractal analysis of turbulent velocity using the wavelet leaders. Eur. Phys. J. B **61**, 201–215 (2008)
84. Leonarduzzi, R., Wendt, H., Roux, S.G., Torres, M.E., Melot, C., Jaffard, S., Abry, P.: p-exponent and p-leaders, Part II: multifractal analysis. relations to detrended fluctuation analysis. Physica A **448**, 319–339 (2016)

85. Leonarduzzi, R., Abry, P., Jaffard, S., Wendt, H., Gournay, L., Kyriacopoulou, T., Martineau, C., Martinez, C.: P-leader multifractal analysis for text type identification. In: IEEE Int. Conf. Acoust., Speech, and Signal Proces. (ICASSP), New Orleans, USA, March (2017)
86. Lieberman, D.E., Bramble, D.M.: The evolution of marathon running: capabilities in humans. Sports Med. **37**, 288 (2007)
87. Lindenstrauss, J., Benyamini, Y.: Geometric Nonlinear Functional Analysis. Colloquium Publications. American Mathematical Society, Providence (2000)
88. Lux, T.: Higher dimensional multifractal processes: A GMM approach. J. Bus. Econ. Stat. **26**(2), 194—210 (2007)
89. Mandelbrot, B.: Geometry of homogeneous scalar turbulence: iso-surface fractal dimensions $5/2$ and $8/3$. J. Fluid Mech. **72**(2), 401–416 (1975)
90. Mandelbrot, B.: Fractals and scaling in finance. Selected Works of Benoit B. Mandelbrot. Springer, New York (1997). Discontinuity, concentration, risk, Selecta Volume E, With a foreword by R.E. Gomory
91. Mandelbrot, B., van Ness, J.W.: Fractional Brownian motion, fractional noises and applications. SIAM Rev. **10**, 422–437 (1968)
92. Marmi, S., Moussa, P., Yoccoz, J.C.: The Brjuno functions and their regularity properties. Commun. Math. Phys. **186**(2), 265–293 (1997)
93. Maron, M., Horvath, S.M., Wilkerson, J.E., Gliner, J.A.: Oxygen uptake measurements during competitive marathon runnings. J. Appl. Physiol. **10**, 137–150 (1978)
94. Mattila, P.: Geometry of Sets and Measures in Euclidian Spaces. Cambridge University Press (1995)
95. Meneveau, C., Sreenivasan, K.R., Kailasnath, P., Fan, M.S.: Joint multifractal measures - theory and applications to turbulence. Phys. Rev. A **41**(2), 894–913 (1990)
96. Meyer, Y.: Ondelettes et Opérateurs. Hermann, Paris (1990). English translation, *Wavelets and operators*, Cambridge University Press, 1992
97. Meyer, Y.: Wavelets, Vibrations and Scalings. CRM Ser. AMS, vol. 9. Presses de l'Université de Montréal, Paris (1998)
98. Muzy, J.F., Bacry, E., Arneodo, A.: Wavelets and multifractal formalism for singular signals: application to turbulence data. Phys. Rev Lett. **67**, 3515–3518 (1991)
99. Parisi, G., Frisch, U.: Fully developed turbulence and intermittency. In: Ghil, M., Benzi, R., Parisi, G. (eds.) Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics, Proc. of Int. School, p. 84. North-Holland, Amsterdam (1985)
100. Peyrière, J.: A vectorial multifractal formalism. Proc. Symp. Pure Math. **72**(2), 217–230 (2004)
101. Pycke, J.-R., Billat, V.: Marathon performance depends on pacing oscillations between non symmetric extreme values. Int. J. Environ. Res. Public Health **19**(4), 2463 (2022)
102. Riedi, R.H.: Multifractal processes. In Doukhan, P., Oppenheim, G., Taqqu, M.S. (eds.) Theory and Applications of Long Range Dependence, pp. 625–717. Birkhäuser (2003)
103. Saes, G.: Sommes fractales de pulses: Etude dimensionnelle et multifractale des trajectoires et simulations. PhD Thesis of University Paris Est Creteil (2021)
104. Sémécurbe, F., Tannier, C., Roux, S.G.: Spatial distribution of human population in France: exploring the MAUP using multifractal analysis. Geograph. Anal. **48**, 292—313 (2016)
105. Seuret, S.: On multifractality and time subordination for continuous functions. Adv. Math. **220**(3), 936–963 (2009)
106. Seuret, S.: A survey on prescription of multifractal behavior. In: Freiberg, U., Hambly, B., Hinz, M., Winter, S. (eds.) Fractal Geometry and Stochastics VI. Progress in Probability, vol. 76, pp. 47–70. Birkhäuser, Cham (2021)
107. Seuret, S., Lévy-Véhel, J.: The 2-microlocal formalism. In: Fractal Geometry and Applications: A Jubilee of Benoit Mandelbrot - Analysis, Number Theory, and Dynamical Systems, Part 2, **72**, 153–215 (2004)
108. Seuret, S., Ubis, A.: Local L^2-regularity of riemann's fourier series. Ann. Inst. Fourier **67**, 2237–2264 (2017)

109. Smyth, B.: Fast starters and slow finishers: A large-scale data analysis of pacing at the beginning and end of the marathon for 579 recreational runners. J. Sports Anal. **4**, 229–242 (2018)
110. Smyth, B.: How recreational marathon runners hit the wall: A large-scale data analysis of late-race pacing collapse in the 577 marathon. PLoS One **16**, 578 (2022)
111. Wang, H., Xiang, L., Pandey, R.B.: A multifractal detrended fluctuation analysis (MDFA) of the Chinese growth enterprise market (GEM). Physica A **391**(12), 3496–3502 (2012)
112. Wendt, H., Abry, P., Jaffard, S.: Bootstrap for empirical multifractal analysis. IEEE Signal Process. Mag. **24**(4), 38–48 (2007)
113. Wesfreid, E., Billat, V., Meyer, Y.: Multifractal analysis of heartbeat time series in human races. Appl. Comput. Harmon. Anal. 329–335 (2010)
114. Whitcher, B., Guttorp, P., Percival, D.B.: Wavelet analysis of covariance with application to atmospheric time series. J. Geophys. Res. Atmos. **105**, 14941–14962 (2000)
115. Yorke, J. Hunt, B., Sauer, T.: Prevalence: a translation invariance "almost every" on infinite dimensional spaces. Bull. Amer. Math. Soc. **27**(2), 217–238 (1992)

Part II
Applications of Dynamical Systems Theory in Biology

Wavefronts in Forward-Backward Parabolic Equations and Applications to Biased Movements

Diego Berti, Andrea Corli, and Luisa Malaguti

Abstract We consider a discrete biological model concerning the movements of organisms, whose population is formed by isolated and grouped individuals. The movement occurs in a random way in one spatial dimension and the transition probabilities per unit time for a one-step jump are assigned. Differently from other papers on the same subject, we assume that the random walk is biased and so, by passing to the limit, we obtain a parabolic equation which includes a convective term. The noteworthy feature of the equation is that the diffusivity changes sign. We investigate the existence of wavefront solutions for this equation, their qualitative properties and we estimate their admissible speeds; in this way we generalize some recent results concerning the case of unbiased movements. Our discussion makes use of some results obtained by the authors on the existence of wavefront solutions in backward-forward parabolic equations.

1 Introduction

The diffusion-convection reaction parabolic equation

$$u_t + f(u)_x = (D(u)u_x)_x + g(u), \qquad t \geq 0,\ x \in \mathbb{R} \tag{1}$$

D. Berti
Department of Mathematics and Computer Science "Ulisse Dini", University of Florence, Florence, Italy
e-mail: diego.berti@unifi.it

A. Corli
Department of Mathematics and Computer Science, University of Ferrara, Ferrara, Italy
e-mail: andrea.corli@unife.it

L. Malaguti (✉)
Department of Sciences and Methods for Engineering, University of Modena and Reggio Emilia, Reggio Emilia, Italy
e-mail: luisa.malaguti@unimore.it

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_2

is frequently used for modeling and analyzing various problems in different areas; we emphasize here those in biology [17] and crowd dynamics [7]. In these cases the unknown function $u = u(t, x)$ in (1) denotes a normalized density or a concentration and then u is valued in the interval $[0, 1]$. Equation (1) occurs, in the case $f = 0$, in the study of *invasion processes* [13, 14] where it is obtained as a continuum limit of a discrete model. The introduction of the term f is motivated below.

We assume that the convective term f satisfies

(f) $f \in C^1[0, 1]$, $f(0) = 0$,

where the condition $f(0) = 0$ just fixes a representative. We consider in the following a *monostable* reaction term g, i.e.,

(g) $g \in C^1[0, 1]$, $g > 0$ in $(0, 1)$, $g(0) = g(1) = 0$.

The most important condition concerns the diffusivity D: we assume

(D) $D \in C^1[0, 1]$, $D > 0$ in $(0, \alpha) \cup (\beta, 1)$ and $D < 0$ in (α, β),

with $\alpha, \beta \in (0, 1)$ and $\alpha < \beta$. Notice that D may vanish in 0 or 1; moreover, the slopes of g and D in 0 and 1 may also be 0. Condition (D) frequently occurs in models of invasion processes [3, 13, 14], and makes (1) a forward-backward-forward equation. We refer to Fig. 1 for a representation of the assumptions above.

Under conditions (f), (g) and (D), Eq. (1) admits *wavefronts*, i.e. traveling wave solutions $u(t, x) = \varphi(x - ct)$, for some speed $c \in \mathbb{R}$, whose profiles φ reach the equilibria $u \equiv 0$ and $u \equiv 1$ at infinity and are monotone [3]. In particular, if a profile is decreasing, then it satisfies

$$\varphi(-\infty) = 1, \qquad \varphi(\infty) = 0. \tag{2}$$

Wavefronts are in agreement with several experimental data, hence their interest. The study of their existence in reaction-diffusion equations goes back to the seminal papers by Aronson-Weinberger [1] and Fife [9], and was then discussed in various contexts with different techniques (see e.g. [6, 10, 11, 16]). A detailed discussion on wavefronts satisfying (2), when the diffusivity is as in (D), recently appeared in [3]. A result appearing in [3] provides the existence of a continuum of wavefronts for Eq. (1); they are parameterized by their speed c, for c in the half-line $[c^*, \infty)$.

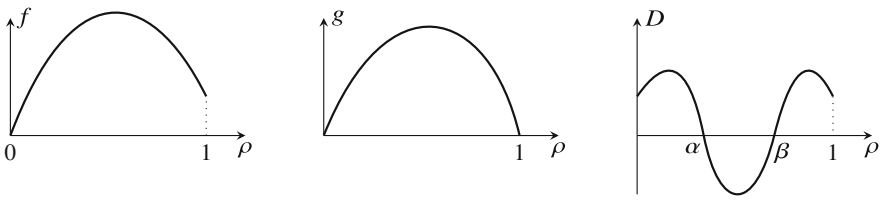


Fig. 1 Typical plots of the functions f, g and D

Estimates on the threshold value c^* are also obtained there. We remark that, in special cases, c^* can be exactly computed.

Consider now the discrete biological model in [13, 14], where the population consists of isolated and grouped individuals moving randomly on a line. However, differently from those papers, assume that the random walk is *biased*. The continuum model obtained by passing to the limit has exactly the form of Eq. (1); since the movement is biased, this limiting equation includes, in particular, the convective term f. This derivation appears in Sect. 2; comparisons with previous similar models are also made there. In Sect. 3 we provide the theoretical tools for the study of this model. In particular, we summarize some general results from [3] (see Theorem 1); we also show that, in the case $f = 0$, the estimates of c^* obtained in [3] lead to an explicit value (see Corollary 1), which coincides with that in [14]. By means of these results, in Sect. 4 we discuss the existence of wavefront solutions to the model proposed in Sect. 2. In this way we extend a result of [14] to the case of a biased movement.

In some cases the reaction term g displays a so-called *bistable* behavior, i.e., $g < 0$ in $(0, \theta)$ and $g > 0$ in $(\theta, 1)$, with $g(0) = g(\theta) = g(1) = 0$ and $\theta \in (0, 1)$. This happens, in particular, in the models of invasion processes under some conditions on their parameters (see Remark 1). Also in this case Eq. (1) admits wavefronts, but the range of their wave speeds now reduces to a bounded closed interval, which can possibly degenerate to a single value. Moreover, the presence of the convective term f allows additional properties on their profiles: for example, the number of profiles for each admissible speed c can be infinite, and profiles may display *plateaus* (i.e., horizontal stretches) at level θ. We refer to [4] for a complete discussion of this case including the presence of a convective term.

2 A Biological Model with Biased Movements

The modeling of the movement of organisms in a biological system by partial differential equations has a long history, and a satisfactory framework was provided in [18–20]. In particular, the continuum model in [19] was obtained by passing to the limit in a discrete model where the transition probabilities per unit time for a one-step jump were assigned; we also refer to [12] for the connections between the discrete and the continuum model. A similar approach, in one space dimension, was more recently proposed in [13] for populations constituted by isolated and grouped individuals, and a focus on the case where the diffusivity becomes negative has been done in [14]; both papers are in particular concerned with wavefront solutions.

In [12, 14] the random movement in the underlying discrete model is assumed to be *unbiased*, in the sense that the transitional probabilities do not depend on the direction of the motion. Then, no convection appears. We provide here a generalization of the model in [13] by introducing a possibly *biased* movement; this leads, in general, to a convective term.

We now introduce the model, referring to [13] for more details. The population is constituted by isolated agents and grouped agents; both classes can reproduce, die and move, with different rates. Agents occupy the sites $j\Delta$, for $j = 0, \pm 1, \pm 2, \ldots$ and $\Delta > 0$. Let c_j be the probability of occupancy in the j site. Its variation during a time-step $\tau > 0$ is provided by the following process:

$$
\begin{aligned}
&\delta c_j = \\
&= \frac{P_m^i}{2}\Big[a^i c_{j-1}(1-c_j)(1-c_{j-2}) + b^i c_{j+1}(1-c_j)(1-c_{j+2}) \\
&\quad - \left(a^i + b^i\right) c_j (1-c_{j-1})(1-c_{j+1})\Big] \\
&\quad + \frac{P_m^g}{2}\left[a^g c_{j-1}(1-c_j) + b^g c_{j+1}(1-c_j) - a^g c_j (1-c_{j+1}) - b^g c_j (1-c_{j-1})\right] \\
&\quad - \frac{P_m^g}{2}\left[a^g c_{j-1}(1-c_j)(1-c_{j-2}) + b^g c_{j+1}(1-c_j)(1-c_{j+2})\right. \\
&\quad \left. -(a^g + b^g) c_j (1-c_{j-1})(1-c_{j+1})\right] + \text{ reaction terms.}
\end{aligned}
\tag{3}
$$

The terms P_m^i and P_m^g are the movement transitional probabilities for isolated and grouped individuals, respectively. If $a^i = b^i = a^g = b^g = 1$, then (3) coincides with [13, (1)]; we have not explicitly written the reaction terms because they are exactly as in [13, (1)]. We introduced instead the positive parameters a^i, b^i and a^g, b^g, which produce a (linearly) biased movement for the isolated and grouped individuals, respectively. We use the notation $a^{i,g}, b^{i,g}$ to denote the two sets of parameters together. By noticing that each bracket is divided by 2, we have

$$
a^i + b^i = a^g + b^g = 2, \tag{4}
$$

because a bias $a^{i,g}$ toward the left implies a complementary bias $b^{i,g} = 2 - a^{i,g}$ toward the right. The continuum model is obtained by understanding c_j as a smooth function $c = c(x,t)$ and expanding c around $x = j\Delta$ at second order. Then, one divides Eq. (3) by τ and passes to the limit for $\Delta \to 0$ and $\tau \to 0$ while keeping Δ^2/τ constant; for simplicity we assume $\Delta^2/\tau = 1$. Analogous assumptions are made in [12, 19] and in [13, 14]. In this procedure, as in [13], one assumes that the following limits are finite:

$$
\lim_{\Delta,\tau\to 0} \frac{P_m^{i,g}}{2} =: D_{i,g}.
$$

Here, we also require, for some $C_{i,g} \in \mathbb{R}$,

$$\lim_{\tau\to 0} a^{i,g}(\tau) = \lim_{\tau\to 0} b^{i,g}(\tau) = 1 \quad \text{and} \quad a^{i,g}(\tau) - b^{i,g}(\tau) \sim C_{i,g}\sqrt{\tau} \text{ for } \tau \to 0. \tag{5}$$

Assumption $(5)_1$, namely, that the limits of $a^{i,g}$ and $b^{i,g}$ coincide, implies that their common value is 1 by (4). Assumption $(5)_2$ is analogous to assumption [13, (3)] for the birth and death probabilities in the reaction terms, which is here assumed as well. Then one finds Eq. (1) with

$$f(u) = -\left(C_i D_i + C_g D_g\right) u(1-u)^2 - C_g D_g u(1-u), \tag{6}$$

$$D(u) = D_i\left(1 - 4u + 3u^2\right) + D_g\left(4u - 3u^2\right), \tag{7}$$

$$g(u) = \lambda_g u(1-u) + \left[\lambda_i - \lambda_g - \left(k_i - k_g\right)\right] u(1-u)^2 - k_g u. \tag{8}$$

Notice that in the case $C_i = 0$ the function f is convex in $[0, 1]$ if $a^g > b^g$, and concave otherwise. The expression of the terms D and g coincide with those in [13, (2)]; the parameters $\lambda_{i,g}$ and $k_{i,g}$ are as in [13, (3)].

3 Wavefronts in a Forward-Backward-Forward Parabolic Model

We briefly report in the following the main results on the existence and main properties of wavefronts for Eq. (1); we denote with φ the wave profile and assume that it is decreasing. We refer to [3, Theorem 6.1] for a comprehensive discussion.

Theorem 1 *Consider Eq.* (1) *and assume conditions* (f), (g) *and* (D). *Then there exists* $c^* \in \mathbb{R}$ *such that* (1) *admits wavefronts satisfying* (2) *if and only if* $c \geq c^*$. *Moreover,*

(i) we have $\varphi' < 0$ *if* $0 < \varphi < 1$ *and* $\varphi \neq \alpha$, $\varphi \neq \beta$;
(ii) $D(\varphi)\varphi' \to 0$ *as* $\xi \to \alpha$ *and* $\xi \to \beta$;
(iii) the threshold value c^* *can be estimated depending on* f, g *and* D.

We now comment on the previous results.

First, for every $c \geq c^*$ the corresponding wavefront is unique up to space shifts, as usual when dealing with wavefronts.

Second, the threshold c^* is defined as the maximum among three sub-thresholds, that we call here c_1^*, c_2^* and c_3^*; they are the thresholds for the existence of wavefronts connecting 0 with α, α with β and β with 1, respectively. In general, and in particular if $f \neq 0$, an explicit value of c^* cannot be given. However, we provide below some

estimates for these thresholds. They can all be obtained by [4, (4,3), (4.6)] and are the following

$$\max\left\{\sup_{(0,\alpha]}\mathcal{A}_0,\, f'(0)+2\sqrt{D(0)g'(0)}\right\}\le c_1^*\le \sup_{(0,\alpha]}\mathcal{A}_0+2\sqrt{\sup_{(0,\alpha]}\mathcal{B}_0}, \tag{9}$$

$$\max\left\{\sup_{[\alpha,\beta)}\mathcal{A}_\beta,\, f'(\beta)+2\sqrt{D'(\beta)g(\beta)}\right\}\le c_2^*\le \sup_{[\alpha,\beta)}\mathcal{A}_\beta+2\sqrt{\sup_{[\alpha,\beta)}\mathcal{B}_\beta}, \tag{10}$$

$$\max\left\{\sup_{(\beta,1]}\mathcal{A}_\beta,\, f'(\beta)+2\sqrt{D'(\beta)g(\beta)}\right\}\le c_3^*\le \sup_{(\beta,1]}\mathcal{A}_\beta+2\sqrt{\sup_{(\beta,1]}\mathcal{B}_\beta}, \tag{11}$$

where, for $v\in[0,1]$, the functions $\mathcal{A}_v=\mathcal{A}_v(u)$ and $\mathcal{B}_v=\mathcal{B}_v(u)$ are defined by, for $u\in[0,1]\setminus\{v\}$,

$$\mathcal{A}_v(u):=\frac{f(u)-f(v)}{u-v}\qquad \mathcal{B}_v(u):=\frac{1}{u-v}\int_v^u\frac{D(s)g(s)}{s-v}\,ds.$$

Third, the C^1 regularity of g is assumed here in order to simplify the discussion; similar results hold true when g is barely continuous in the interval $[0,1]$, see [3].

Fourth, let φ be a profile of a wavefront for (1). The value $\xi_\alpha\in\mathbb{R}$ such that $\varphi(\xi_\alpha)=\alpha$ is easily proved to be unique; moreover, one can deduce that $\varphi'(\xi_\alpha)<0$ if $D'(\alpha)<0$, while it can be $\varphi'(\xi_\alpha)=-\infty$ if $D'(\alpha)=0$. An analogous discussion holds for the point β. Explicit formulas for $\varphi'(\xi_\alpha)$ and $\varphi'(\xi_\beta)$ are provided in [3, (2.9), (2.16)].

Now, we give a rough sketch of the proof of Theorem 1. The main technique is a first-order reduction: since the profiles are strictly monotone φ when $0<\varphi<1$ (and $\varphi\neq\alpha$, $\varphi\neq\beta$), hence the inverse function $\xi(\varphi)$ is well defined and so is

$$z(\varphi):=D(\varphi)\varphi(\xi(\varphi)).$$

It is plain to see that $z(\varphi)$ satisfies the first-order, singular equation

$$\dot z(\varphi)=-c+f'(\varphi)-\frac{D(\varphi)g(\varphi)}{z(\varphi)} \tag{12}$$

for $\dot z=dz/d\varphi$, and also that $z(0)=z(\alpha)=z(\beta)=z(1)=0$. The investigation then proceeds by studying (12) separately in each interval $[0,\alpha]$, $[\alpha,\beta]$ and $[\beta,1]$ by mainly exploiting comparison-type techniques, for which we refer to [2].

In order to check how the estimates on c^* that we obtained are precise, we now focus on the case $f = 0$. We show, under a light further condition, that in this case they provide an explicit value for c^*. More precisely, we assume

$$(Dg)'(u) \le (Dg)'(0) \quad \text{for every } u \in [0, 1]. \tag{13}$$

Assumption (13) is motivated by the fact that when it holds and also $f'(u) \le f'(0)$ for $u \in [0, \alpha]$, then (9) clearly reduces to

$$c_1^* = f'(0) + 2\sqrt{D(0)g'(0)}.$$

The following simple result is new. It generalizes to the case of general g and D a result in [14].

Corollary 1 *Consider Eq.* (1) *with* $f = 0$ *and assume conditions* (g), (D) *and* (13). *Then*

$$c^* = 2\sqrt{D(0)g'(0)}. \tag{14}$$

Proof By (9) we have

$$2\sqrt{D(0)g'(0)} \le c_1^* \le 2\sqrt{\sup_{u\in(0,\alpha]} \frac{D(u)g(u)}{u}}.$$

Since for every $u \in (0, \alpha]$ we have

$$D(u)g(u) = (Dg)'(\theta_u)u \quad \text{for some } \theta_u \in (0, u),$$

then

$$\sup_{u\in(0,\alpha]} \frac{D(u)g(u)}{u} \le \max_{[0,\alpha]} (Dg)' = D(0)g'(0),$$

because of (13), from which it follows $c_1^* = 2\sqrt{D(0)g'(0)}$. It remains to prove that $c_2^*, c_3^* \le c_1^*$. To this end, we make use of the right-hand side of (10) and (11) as follows.

As above, we have

$$\frac{1}{u-\beta}\int_\beta^u \frac{D(s)g(s)}{s-\beta}\,ds = (Dg)'(\sigma_u) \quad \text{for some } \sigma_u \in (\alpha, 1).$$

where σ_u is contained either in (u, β) or in (β, u), according to $u < \beta$ or $u > \beta$.

Since $(Dg)' = (Dg)'(u)$ has its maximum at $u = 0$ because of Dg, then, from the right-side of (10) and (11), we deduce

$$c_2^*, c_3^* \leq c_1^*.$$

This concludes the proof since c^* is the maximum among c_1^*, c_2^* and c_3^*. □

4 Wavefronts in a Biological Model with Biased Movements

We now apply Theorem 1 and Corollary 1 to the model proposed in Sect. 2. We are then looking for wavefronts for Eq. (1) when we assume (6)–(8). As in [13, 14], we also assume

$$D_i > 4D_g > 0 \quad \text{and} \quad \lambda_g = \lambda_i = \lambda > 0, k_i = k_g = 0. \tag{15}$$

With these choices, D has two sign-changes as prescribed by (D), occurring at some $\alpha \in (1/3, 2/3)$ and $\beta \in (2/3, 1)$ (see [14, (7)] for their expressions), while g simply takes the shape

$$g(u) = \lambda u(1 - u). \tag{16}$$

As far as f is concerned, i.e. the choice of C_i and C_g, we just assume that $C_i, C_g \in \mathbb{R}$. The case $f = 0$, i.e. $C_i = C_g = 0$, was treated in [13, 14].

Theorem 1 applies directly, without additional computations. In particular, under (15) there exists $c^* \in \mathbb{R}$ such that this model admits wavefronts with speed $c \in \mathbb{R}$, connecting 1 to 0, if and only if $c \geq c^*$. Moreover, the profile φ is unique up to space shifts for every admissible c.

Furthermore, we claim that condition (13) is satisfied. Indeed, we have

$$(Dg)'(u) = -12\lambda \left(D_i - D_g\right) u^3 + 21\lambda \left(D_i - D_g\right) u^2 - \lambda \left(10D_i - 8D_g\right) u + \lambda D_i.$$

First, we observe that $(Dg)'(0) = \lambda D_i$. Then, since $-12u^2 + 21u - 8 < 2$, we deduce

$$-12 \left(D_i - D_g\right) u^3 + 21 \left(D_i - D_g\right) u^2 - 8 \left(D_i - D_g\right) u < 2D_i u - 2D_g u,$$

and hence, because of $D_g > 0$, we obtain

$$(Dg)'(u) < \lambda \left(2D_i u - 2D_g u - 2D_i u + D_i\right) < \lambda D_i = (Dg)'(0).$$

Thus, in case $f = 0$, by Corollary 1, we obtain the result in [14, (11)], that is

$$c^* = 2\sqrt{\lambda D_i}.$$

We stress that in [14] the authors claimed the result as an application of the *geometric singular perturbation theory* [8]. Instead, Corollary 1 is a consequence of the upper and lower solution techniques that we adopted to investigate wavefronts.

Now, consider the case $f \neq 0$, and then at least one between C_i and C_g does not vanish. If we denote

$$H(u) := \frac{f(u)}{u} = -(C_i D_i + C_g D_g)(1-u)^2 - C_g D_g (1-u),$$

then (9) becomes

$$\max\left\{\sup_{u\in(0,\alpha]} H(u),\, f'(0) + 2\sqrt{\lambda D_i}\right\} \leq c_1^* \leq \sup_{u\in(0,\alpha]} H(u) + 2\sqrt{\lambda D_i},$$

because (13) is satisfied. Since f is given in (6), then $H(u)$ can be explicitly computed and its supremum can be evaluated depending on the choices of C_i and C_g. This provides explicit bounds for c_1^*, and analogously for c_2^* and c_3^*.

Remark 1 Assumption (15) implies that the reaction term g in (16) is *monostable*. A different choice of the parameters $D_{i,g}, \lambda_{i,g}, k_{i,g}$ clearly gives rise to other models. For instance, we can modulate the values of $\lambda_{i,g}$ and $k_{i,g}$ to obtain a *bistable* reaction term. This happens if

$$k_g = 0 \ \text{ and } \ k_i > \lambda_i \geq 0 \ \text{ and } \ \lambda_g > 0,$$

because then g in (8) becomes

$$g(u) = \theta u(1-u)(-1+\gamma u),$$

with

$$\theta = k_i - \lambda_i \ \text{ and } \ \gamma := \frac{\lambda_g + \theta}{\theta} > 1.$$

For models with unbiased movements, i.e. when $f = 0$, this case was discussed in [13, 15]. In particular, numerical simulations are proposed in [15] to suggest the presence of shock-fronted wavefronts. To the best of our knowledge the existence of wavefronts in biased models has never been investigated. Such a discussion can be led by means of the results and techniques developed in [4], where the equation includes a convective term; we refer to [5].

Acknowledgments 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) and acknowledge financial support from this institution.

References

1. Aronson, D.G., Weinberger, H.F.: Multidimensional nonlinear diffusion arising in population genetics. Adv. Math. **30**, 33–76 (1978)
2. Berti, D., Corli, A., Malaguti, L.: Uniqueness and nonuniqueness of fronts for degenerate diffusion-convection reaction equations. Electron. J. Qual. Theory Diff. Equ. **66**, 1–34 (2020)
3. Berti, D., Corli, A., Malaguti, L.: Wavefronts for degenerate diffusion-convection reaction equations with sign-changing diffusivity. Discrete Contin. Dyn. Syst. **41**(12), 6023–6046 (2021)
4. Berti, D., Corli, A., Malaguti, L.: Diffusion-convection reaction equations with sign-changing diffusivity and bistable reaction term. Nonlinear Anal. Real World Appl. **67**, 29 pp. (2022). Paper No. 103579
5. Berti, D., Corli, A., Malaguti, L.: The role of convection in the existence of wavefronts for biased movements, Submitted (2023) arXiv:2304.02305v1
6. Campos, J., Corli, A., Malaguti, L.: Saturated fronts in crowds dynamics. Adv. Nonlinear Stud. **21**(2), 303–326 (2021)
7. Corli, A., Malaguti, L.: Semi-wavefront solutions in models of collective movements with density dependent diffusivity. Dyn. Partial Differ. Equ. **13**(4), 297–331 (2016)
8. Fenichel, N.: Geometric singular perturbation theory for ordinary differential equations. J. Differ. Equ. **31**(1), 53–98 (1979)
9. Fife, P.C.: Mathematical Aspects of Reacting and Diffusing Systems. Lecture Notes in Biomathematics, vol. 28. Springer, Berlin (1979)
10. Garrione, M., Strani, M.: Monotone wave fronts for (p,q)-Laplacian driven reaction-diffusion equations. Discrete Contin. Dyn. Syst. Ser. S **12**(1), 91–103 (2019)
11. Gilding, B.H., Kersner, R.: Travelling Waves in Nonlinear Diffusion-Convection Reaction. Birkhäuser Verlag, Basel (2004)
12. Horstmann, D., Painter, K.J., Othmer, H.G.: Aggregation under local reinforcement: from lattice to continuum. Eur. J. Appl. Math. **15**(5), 546–576 (2004)
13. Johnston, S.T., Baker, R.E., McElwain, S.D., Simpson, M.J.: Co-operation, competition and crowding: a discrete framework linking Allee kinetic, nonlinear diffusion, shocks and sharp-fronted travelling waves. Sci. Rep. **7**, 42134 (2017)
14. Li, Y., van Heijster, P., Marangell, R., Simpson, M.J.: Travelling wave solutions in a negative nonlinear diffusion-reaction model. J. Math. Biol. **81**(6–7), 1495–1522 (2020)
15. Li, Y., van Heijster, P., Simpson, M.J., Wechselberger, M.: Shock-fronted travelling waves in a reaction-diffusion model with nonlinear forward-backward-forward diffusion. Physica D **423**, 14 pp. (2021). Paper No. 132916
16. Malaguti, L., Marcelli, C., Matucci, S.: Continuous dependence in front propagation of convective reaction-diffusion equations. Commun. Pure Appl. Anal. **9**(4), 1083–1098 (2010)
17. Murray, J.D.: Mathematical Biology. Springer, Berlin (1993)
18. Okubo, A., Levin, S.A.: Diffusion and Ecological Problems: Modern Perspectives. Springer, Berlin (2001)
19. Othmer, H.G., Stevens, A.: Aggregation, blowup, and collapse: the ABCs of taxis in reinforced random walks. SIAM J. Appl. Math. **57**(4), 1044–1081 (1997)
20. Patlak, C.S.: Random walk with persistence and external bias. Bull. Math. Biophys. **15**, 311–338 (1953)

Bohr-Levitan Almost Periodic and Almost Automorphic Solutions of Equation $x'(t) = f(t-1, x(t-1)) - f(t, x(t))$

David Cheban

Abstract This paper is dedicated to the problem of almost periodicity of solutions for functional differential equations $x' = h(t, x_t)(*)$ when the right hand side is monotone with respect to spacial variable. The main results are established in the framework of general non-autonomous (cocycle) dynamical systems. We apply our general results to a class of delay differential equations $x'(t) = f(t-1, x(t-1)) - f(t, x(t))$ (**). In particular, we prove that every solution of equation (**) with Bohr-Levitan right hand side is asymptotically Bohr-Levitan almost periodic.

1 Introduction

This paper is dedicated to the problem of almost periodic solutions and structure of compact global attractor (Levinson center) of functional differential equations

$$x' = h(t, x_t). \tag{1}$$

We apply our general results for a class of delay differential equations

$$x'(t) = f(t-1, x(t-1)) - f(t, x(t)). \tag{2}$$

Equation (2) may be viewed as the nonautonomous form of a model growth processes and gonorrhea epidemics introduced by K. Cooke and J. Yorke [1, 2, 11].

The writing of this paper was motivated by the works O. Arino [1], J. Jiang and X. Zhao [13] as well as a series of works by the author [6–10]. The aim of this paper is to study the problem of the existence of Levitan/Bohr almost periodic (respectively, almost automorphic, recurrent and Poisson stable) solutions for dissipative functional differential equation (1) when the right hand side is

D. Cheban (✉)
Moldova State University, Chisinau, Moldova
e-mail: david.ceban@usm.md

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_3

monotone with respect to spacial variable. We establish the main results in the framework of general non-autonomous (cocycle) dynamical systems. We apply our general results for a class equations (2) which may be viewed as the nonautonomous form of a model growth processes and gonorrhea epidemics. In particular, we prove that every solution of equation (2) with Bohr-Levitan right hand side is asymptotically Bohr-Levitan almost periodic. For Bohr almost periodic (respectively, almost automorphic) equations our result improve and/or refine the results of O. Arino [1] (respectively, J. Jiang and X. Zhao [13]).

2 Non-autonomous (Cocycle) Dynamical Systems

In this section we collect some notions and facts from the autonomous and non-autonomous dynamical systems [3] (see also, [5, Ch.IX]) which we will use below.

2.1 Cocycles

Let Y be a complete metric space, let $\mathbb{R} := (-\infty, +\infty)$, $\mathbb{Z} := \{0, \pm 1, \pm 2, \ldots\}$, $\mathbb{T} = \mathbb{R}$ or $\mathbb{Z}$, $\mathbb{T}_+ = \{t \in \mathbb{T} | \quad t \geq 0\}$ and $\mathbb{T}_- = \{t \in \mathbb{T} | \quad t \leq 0\}$. Let $(Y, \mathbb{T}, \sigma)$ be an autonomous two-sided dynamical system on Y and E be a real or complex Banach space with the norm $|\cdot|$.

Definition 1 (Cocycle on the State Space E with the Base $(Y, \mathbb{T}, \sigma)$) The triplet $\langle E, \phi, (Y, \mathbb{T}, \sigma)\rangle$(or briefly ϕ) is said to be a cocycle (see, for example, [5] and [15]) on the state space E with the base $(Y, \mathbb{T}, \sigma)$ if the mapping $\phi : \mathbb{T}_+ \times Y \times E \to E$ satisfies the following conditions:

1. $\phi(0, y, u) = u$ for all $u \in E$ and $y \in Y$;
2. $\phi(t + \tau, y, u) = \phi(t, \phi(\tau, u, y), \sigma(\tau, y))$ for all $t, \tau \in \mathbb{T}_+, u \in E$ and $y \in Y$;
3. the mapping ϕ is continuous.

Definition 2 (Skew-Product Dynamical System) Let $\langle E, \phi, (Y, \mathbb{T}, \sigma)\rangle$ be a cocycle on $E, X := E \times Y$ and π be a mapping from $\mathbb{T}_+ \times X$ to X defined by equality $\pi = (\phi, \sigma)$, i.e., $\pi(t, (u, y)) = (\phi(t, \omega, u), \sigma(t, y))$ for all $t \in \mathbb{T}_+$ and $(u, y) \in E \times Y$. The triplet $(X, \mathbb{T}_+, \pi)$ is an autonomous dynamical system and it is called [15] a skew-product dynamical system.

2.2 Bebutov's Dynamical Systems

Let X, W be two metric spaces. Denote by $C(\mathbb{T} \times W, X)$ the space of all continuous mappings $f : \mathbb{T} \times W \mapsto X$ equipped with the compact-open topology and let σ

be the mapping from $\mathbb{T} \times C(\mathbb{T} \times W, X)$ into $C(\mathbb{T} \times W, X)$ defined by the equality $\sigma(\tau, f) := f_\tau$ for all $\tau \in \mathbb{T}$ and $f \in C(\mathbb{T} \times W, X)$, where f_τ is the τ-translation (shift) of f with respect to variable t, i.e., $f_\tau(t, x) = f(t+\tau, x)$ for all $(t, x) \in \mathbb{T} \times W$. Then [5, Ch.I] the triplet $(C(\mathbb{T} \times W, X), \mathbb{T}, \sigma)$ is a dynamical system on $C(\mathbb{T} \times W, X)$ which is called a *shift dynamical system* (*dynamical system of translations* or *Bebutov's dynamical system*).

Recall that the function $\varphi \in C(\mathbb{T}, X)$ (respectively, $f \in C(\mathbb{T} \times W, X)$) possesses the property (A), if the motion $\sigma(\cdot, \varphi)$ (respectively, $\sigma(\cdot, f)$) generated by the function φ (respectively, f) possesses this property in the dynamical system $(C(\mathbb{T}, X), \mathbb{T}, \sigma)$ (respectively, $(C(\mathbb{T} \times W, X), \mathbb{T}, \sigma)$).

As the quality of the property (A) there can stand the Lagrange stability (st. L), uniform Lyapunov stability (un. st. $\mathcal{L}^+$), periodicity, almost periodicity, asymptotical almost periodicity and so on.

For example, a function $f \in C(\mathbb{T} \times W, X)$ is called almost periodic (respectively, recurrent etc.) in $t \in \mathbb{R}$ uniformly with respect to (w.r.t.) w on every compact subset from W, if the motion $\sigma(\cdot, f)$ is almost periodic (respectively, recurrent) in the dynamical system $(C(\mathbb{T} \times W, X), \mathbb{T}, \sigma)$.

2.3 Bohr-Levitan Almost Periodic, Almost Automorphic and Recurrent Motions

Definition 3 A number $\tau \in \mathbb{S}$ is called an $\varepsilon > 0$ shift of x (respectively, almost period of x), if $\rho(\pi(\tau, x), x) < \varepsilon$ (respectively, $\rho(\pi(\tau + t, x), \pi(t, x)) < \varepsilon$ for all $t \in \mathbb{S}$).

Definition 4 A point $x \in X$ is called almost recurrent (respectively, Bohr almost periodic), if for any $\varepsilon > 0$ there exists a positive number l such that at any segment of length l there is an ε shift (respectively, almost period) of point $x \in X$.

Definition 5 If the point $x \in X$ is almost recurrent and the set $H(x) := \overline{\{\pi(t, x) \mid t \in \mathbb{T}\}}$ is compact, then x is called recurrent.

Definition 6 A point $x \in X$ of the dynamical system $(X, \mathbb{T}, \pi)$ is called Levitan almost periodic [14] (see also [6, Ch.I]), if there exists a dynamical system $(Y, \mathbb{T}, \sigma)$ and a Bohr almost periodic point $y \in Y$ such that $\mathfrak{N}_y \subseteq \mathfrak{N}_x$, where $\mathfrak{N}_x := \{\{t_k\} | \, \pi(t_k, x) \to x \text{ as } k \to \infty\}$.

Definition 7 A point $x \in X$ is called Lagrange stable, if its trajectory $\Sigma_x := \{\pi(t, x) : t \in \mathbb{T}\}$ is relatively compact.

Definition 8 A point $x \in X$ is called almost automorphic in the dynamical system $(X, \mathbb{T}, \pi)$, if it is Lagrange stable and Levitan almost periodic.

2.4 B. A. Shcherbakov's Principle of Comparability of Motions by Their Character of Recurrence

In this subsection we will present some notions and results stated and proved by B. A. Shcherbakov [16, 17] (see also [6, Ch.I]).

Let $(X, \mathbb{T}_1, \pi)$ and $(Y, \mathbb{T}_2, \sigma)$ be two dynamical systems.

Definition 9 A point $x \in X$ is said to be comparable (respectively, uniformly comparable) with $y \in Y$ by the character of recurrence, if the mapping $h : \Sigma_y \to \Sigma_x$, defined by equality

$$h(\sigma(t, y)) = \pi(t, x)$$

for any $t \in \mathbb{T}$, is continuous (respectively, uniformly continuous).

Theorem 1 *Let x be comparable with $y \in Y$. If the point $y \in Y$ is stationary (respectively, τ–periodic, Levitan almost periodic, almost recurrent), then the point $x \in X$ is also so.*

Denote by $\mathfrak{M}_x := \{\{t_n\} \subset \mathbb{R} : \text{ such that } \{\pi(t_n, x)\} \text{ converges }\}$.

Definition 10 A point $x \in X$ is said [4, ChII] to be strongly comparable by character of recurrence with the point $y \in Y$, if $\mathfrak{M}_y \subseteq \mathfrak{M}_x$.

Theorem 2 *Let X be a complete metric space. If the point x uniformly comparable by character of recurrence with y, then $\mathfrak{M}_y \subseteq \mathfrak{M}_x$.*

Theorem 3 *Let y be stable in the sense of Lagrange. The inclusion $\mathfrak{M}_y \subseteq \mathfrak{M}_x$ takes place, if and only if the point x is stable in the sense of Lagrange and the point x uniformly comparable by character of recurrence with y.*

Theorem 4 *Let X and Y be two complete metric spaces, the point x be uniformly comparable with $y \in Y$ by the character of recurrence. If the point $y \in Y$ is stationary (respectively, τ-periodic, quasi-periodic with the frequency basis $\nu_1, \nu_2, \ldots, \nu_m$, Bohr almost periodic, almost automorphic, almost recurrent, recurrent), then so is the point $x \in X$.*

2.5 Monotone Non-autonomous Dynamical Systems

Assume that E is an ordered Banach space [18]. A subset U of E is called lower-bounded (respectively, upper-bounded) if there exists an element $a \in E$ such that $a \leq U$ (respectively, $a \geq U$). Such an a is said to be a lower bound (respectively, upper bound) for U. A lower bound α is said to be the *greatest lower bound* (g.l.b.) or *infimum*, if any other lower bound a satisfies $a \leq \alpha$. Similarly, we can define the *least upper bound* (l.u.b.) or *supremum*.

Let V be an order convex subset of E.

Definition 11 Let $\langle E, \varphi, (Y, \mathbb{T}, \sigma)\rangle$ be a cocycle and $\langle (X, \mathbb{T}_+, \pi), (Y, \mathbb{T}, \sigma), h\rangle$ be a non-autonomous dynamical system associated by cocycle φ (i.e., $X := E \times Y$, $\pi = (\varphi, \sigma)$ and $h := pr_2 : X \to Y$). The cocycle φ is said to be monotone if $u_1 \leq u_2$ implies $\varphi(t, u_1, y) \leq \varphi(t, u_2, y)$ for any $t > 0$ and $y \in Y$.

Recall that

1. a forward orbit $\{\pi(t, x_0)|\ t \geq 0\}$ ($x_0 = (u_0, y_0)$) of skew-product dynamical systems $\langle (X, \mathbb{T}_+, \pi)$ ($X = E \times Y$, $\pi = (\varphi, \sigma)$) is said to be uniformly stable if for any $\varepsilon > 0$, there is a $\delta = \delta(\varepsilon) > 0$ such that $\rho(\varphi(t_0, u, y_0), \varphi(t_0, u_0, y_0)) < \delta$ implies $\rho(\varphi(t, u, y_0), \varphi(t, u_0, y_0)) < \varepsilon$ for every $t \geq t_0$;
2. a continuous mapping $\gamma : \mathbb{S} \to X$ is said to be an entire (full) trajectory of the skew-product dynamical system $(X, \mathbb{T}_+, \pi)$ if $\gamma(t + \tau) = \pi(t, \gamma(\tau))$ for any $t \in \mathbb{T}_+$ and $\tau \in \mathbb{S}$.

Below we will use the following assumptions:

(**C1**) every compact subset K of V has both infimum $\alpha(K)$ and supremum $\beta(K)$;
(**C2**) for every $x \in V \times Y$, the set $\varphi(\mathbb{T}_+, u, y)$ is pre-compact and positively uniformly Lyapunov stable;
(**C3**) the cocycle $\langle E, \varphi, (Y, \mathbb{T}, \sigma)\rangle$ is monotone;
(**C4**) for any two bounded full trajectories γ_j ($j = 1, 2$) with $\gamma_1(t) \leq \gamma_2(t)$ for any $t \in \mathbb{T}$, there exists $t_0 > 0$ such that whenever $\gamma_1(s) < \gamma_2(s)$ holds for some $s \in \mathbb{T}$, then $\gamma_1(t) \ll \gamma_2(t)$ for all $t \geq s + t_0$.

Denote by

$$\omega_x := \bigcap_{t \geq 0} \overline{\bigcup_{\tau \geq t} \pi(\tau, x)}$$

ω-limit set of the point $x \in X$.

Theorem 5 ([8]) *Suppose that (**C1**)–(**C3**) and (**C4**) are fulfilled. Assume that $(Y, \mathbb{T}, \sigma)$ is almost recurrent, i.e., there exists an almost recurrent point $y_0 \in Y$ such that $Y = H(y_0)$.*

Then for any $(u_0, y_0) \in V \times Y$ the following statements hold:

1. *for any $q \in Y$ the set*

$$\omega_{(u_0, y_0)} \bigcap X_q$$

consists of a single point $\{(u_q, q)\}$;
2. *the point (x_q, q) is strongly comparable by character of recurrence with the point $q \in Y$;*
3.

$$\lim_{t \to +\infty} \rho(\varphi(t, u_0, y_0), \varphi(t, u_{y_0}, y_0)) = 0.$$

3 Functional-Differential Equations with Finite Delay

Let us first recall some notions and notations from [12]. Let $r > 0$, $C([a, b], \mathbb{R}^n)$ be the Banach space of all continuous functions $\varphi : [a, b] \to \mathbb{R}^n$ equipped with the sup–norm. If $[a, b] = [-r, 0]$, then we set $C := C([-r, 0], \mathbb{R}^n)$. Let $\sigma \in \mathbb{R}$, $A \geq 0$ and $u \in C([\sigma - r, \sigma + A], \mathbb{R}^n)$. We will define $u_t \in C$ for any $t \in [\sigma, \sigma + A]$ by the equality $u_t(\theta) := u(t + \theta)$, $-r \leq \theta \leq 0$. Consider a functional differential equation

$$\dot{u} = f(t, u_t), \tag{3}$$

where $f : \mathbb{R} \times C \to \mathbb{R}^n$ is continuous.

Denote by $C(\mathbb{R} \times C, \mathbb{R}^n)$ the space of all continuous mappings $f : \mathbb{R} \times C \to \mathbb{R}^n$ equipped with the compact-open topology. On the space $C(\mathbb{R} \times C, \mathbb{R}^n)$ there is defined (see, e.g. [5, ChI] and [16, ChI]) a shift dynamical system $(C(\mathbb{R} \times C, \mathbb{R}^n), \mathbb{R}, \sigma)$, where σ is a mapping from $\mathbb{R} \times C(\mathbb{R} \times C, \mathbb{R}^n)$ to $C(\mathbb{R} \times C, \mathbb{R}^n)$ defined by equality $\sigma(\tau, f) := f_\tau$ for any $f \in C(\mathbb{R} \times C, \mathbb{R}^n)$ and $\tau \in \mathbb{R}$ and f_τ is τ-translation of f, i.e. $f_\tau(t, \phi) := f(t + \tau, \phi)$ for any $(t, \phi) \in \mathbb{R} \times C$. Let us set $H(f) := \overline{\{f_\tau : \tau \in \mathbb{R}\}}$.

Along with Eq. (3) let us consider the family of equations

$$\dot{v} = g(t, v_t), \tag{4}$$

where $g \in H(f)$.

Condition (F1). In this Section, we suppose that Eq. (3) is regular, i.e., the conditions of existence, uniqueness and extendability on $\mathbb{R}_+$ for any Eq. (4) are fulfilled.

Remark 1 Denote by $\tilde{\varphi}(t, u, f)$ the solution of equation (3) defined on $[-r, +\infty)$ with the initial condition $u \in C$. By $\varphi(t, u, f)$ we will denote below the trajectory of Eq. (3), corresponding to the solution $\tilde{\varphi}(t, u, f)$, i.e. a mapping from $\mathbb{R}_+$ into C, defined by $\varphi(t, u, f)(s) := \tilde{\varphi}(t + s, u, f)$ for any $t \in \mathbb{R}_+$ and $s \in [-r, 0]$. Below we will use the notions of "solution" and "trajectory" for Eq. (3) as synonymous concepts.

It is well-known [15] that the mapping $\varphi : \mathbb{R}_+ \times C \times H(f) \to C$ possesses the following properties:

1. $\varphi(0, v, g) = v$ for any $v \in C$ and $g \in H(f)$;
2. $\varphi(t + \tau, v, g) = \varphi(t, \varphi(\tau, v, g), \sigma(\tau, g))$ for any $t, \tau \in \mathbb{R}_+$, $v \in C$ and $g \in H(f)$;
3. the mapping φ is continuous.

Thus Eq. (3) generates a cocycle $\langle C, \varphi, (Y, \mathbb{R}, \sigma)\rangle$ and a non-autonomous dynamical system $\langle (X, \mathbb{R}_+, \pi), (Y, \mathbb{R}, \sigma), h\rangle$, where $Y := H(f)$, $X := C \times Y$, $\pi := (\varphi, \sigma)$ and $h := pr_2 : X \to Y$.

Remark 2 Let F be a mapping from $H(f)\times C \to \mathbb{R}^n$ defined by equality $F(g,x) = g(0,x)$ for any $(g,x) \in H(f) \times C$, then F possesses the following properties:

1. F is continuous;
2. $F(g^\tau, x) = g(\tau, x)$ for any $(g, x, \tau) \in H(f) \times C \times \mathbb{R}$;
3. Eq. (3) (and its H-class (4)) can be rewritten as follows

$$x'(t) = F(\sigma(t,g), x_t) \ \ (g \in H(f)). \tag{5}$$

Let $C_+ := \{\phi \in C : \ \phi \geq 0$, i.e., $\phi(t) \geq 0$ for any $t \in [-r, 0]\}$ be the cone of nonnegative functions in C. By C_+ on the space C there is defined a partial order: $u \leq v$ if and only if $v - u \in C_+$.

Condition (F2). Equation (3) is monotone, that is, the cocycle $\langle C, \varphi, (H(f), \mathbb{R}, \sigma)\rangle$ generated by (3) possesses the following property: if $u \leq v$, then $\varphi(t, u, g) \leq \varphi(t, v, g)$ for any $t \geq 0$ and $g \in H(f)$.

Recall (see, e.g. [18, ChV]) that a function $f \in C(\mathbb{R} \times C, \mathbb{R}^n)$ is said to be *quasi-monotone* if $(t, u), (t, v) \in \mathbb{R} \times C$, $u \leq v$, and $u_i(0) = v_i(0)$ for some i, then $f_i(t, u) \leq f_i(t, v)$.

Lemma 1 ([10]) *Let $f \in C(\mathbb{R} \times C, \mathbb{R}^n)$ be a quasi-monotone function, then the following statements hold:*

1. *if $u \leq v$, then $\varphi(t, u, f) \leq \varphi(t, v, f)$ for any $t \geq 0$;*
2. *any function $g \in H(f)$ is quasi-monotone;*
3. *$u \leq v$ implies $\varphi(t, u, g) \leq \varphi(t, v, g)$ for any $t \geq 0$ and $g \in H(f)$.*

Corollary 1 *Let $f \in C(\mathbb{R}\times C, \mathbb{R}^n)$ be a regular and quasi-monotone function, then the cocycle $\langle C, \varphi, (H(f), \mathbb{R}, \sigma)\rangle$, associated by Eq. (3), is monotone.*

Remark 3 If the function $f \in C(\mathbb{R}\times C, \mathbb{R}^n_+)$ is quasi-monotone, then F ($F(g,x) = g(0,x)$ for any $(g,x) \in H(f) \times C$) is also so, i.e., for any $(g, u), (g, v) \in H(f) \times C_+$, $u \leq v$, and $u_i(0) = v_i(0)$ for some i, then $F_i(g, u) \leq F_i(g, v)$.

Theorem 6 ([18, Ch.V]) *Let $f, g \in C(\mathbb{R} \times C_+, \mathbb{R}^n)$ be regular and assume that either f or g is quasi-monotone. Assume also that $f(t, \phi) \leq g(t, \phi)$ for all $(t, \phi) \in \mathbb{R} \times C_+$. If $\phi, \psi \in C_+$ satisfy $\phi \leq \psi$, then $\varphi(t, \phi, f) \leq \varphi(t, \psi, g)$ for all $t \geq 0$.*

Corollary 2 *Assume that $f \in C(\mathbb{R} \times C_+, \mathbb{R}^n)$ is regular, quasi-monotone and $\langle C_+, \varphi, (H(f), \mathbb{R}, \sigma)\rangle$ is the cocycle in C_+ generated by Eq. (3) (respectively, by its H-class (4)). Then the condition*

$$\mathcal{F}(g,x) \leq F(g,x)$$

for any $(g,x) \in H(f) \times C_+$ implies that

$$\phi(t, x, g) \leq \varphi(t, x, g)$$

for any $t \geq 0$, $g \in H(f)$ and $x \in C_+$, where $\mathcal{F} \in C(H(f) \times C_+, \mathbb{R}^n)$ is some regular function and $\langle C_+, \phi, (H(f), \mathbb{R}, \sigma)\rangle$ (shortly, ϕ) is the cocycle generated by equation

$$x' = \mathcal{F}(\sigma(t, g), x_t) \quad (g \in H(f)).$$

Proof This statement follows directly from Theorem 6. □

Condition (F3). The cone C_+ is positively invariant with respect to cocycle φ generated by Eq. (3), i.e., $\varphi(t, \phi, g) \in C_+$ for any $(t, \phi, g) \in \mathbb{R}_+ \times C_+ \times H(f)$.

Lemma 2 *Assume that the function $f \in C(\mathbb{R} \times C, \mathbb{R}^n)$ is regular, quasi-monotone and $f(t, 0) \geq 0$ for any $t \in \mathbb{R}$. Then C_+ is a positively invariant subset of the cocycle φ, generated by Eq. (3), i.e., $\varphi(t, x, g) \in C_+$ for any $(t, x, g) \in \mathbb{R}_+ \times C_+ \times H(f)$.*

Proof Let $g \in H(f)$, then it is easy to check that under the condition of Lemma 2 the function g is also regular, quasi-monotone and $g(t, 0) \geq 0$ for any $t \in \mathbb{R}$. Note that $F(g, x) = g(0, x) \geq 0$ for any $(x, g) \in C_+ \times H(f)$. By Theorem 6 we have $\phi(t, x, g) \leq \varphi(t, x, g)$ for any $t \geq 0$, $g \in H(f)$ and $x \in C_+$, where φ is the cocycle generated by Eq. (3) (respectively, Eq. (5)) and ϕ is the cocycle defined by equation $x' = 0$, i.e., $\phi(t, x, g) = x$ for any $x \in C_+$, $t \geq 0$ and $g \in H(f)$. By Corollary 2 we have $\varphi(t, x, g) \geq x$ for any $(t, x, g) \in \mathbb{R}_+ \times C_+ \times H(f)$. This means that $\varphi(t, x, g) \geq 0$, i.e., $\varphi(t, x, g) \in C_+$ for any $(t, x, g) \in \mathbb{R}_+ \times C_+ \times H(f)$. Lemma is proved. □

Condition (F4). For any bounded subset $A \subset C$ the set $f(\mathbb{R} \times A)$ is bounded in $\mathbb{R}^n$.

Lemma 3 ([10]) *Let $\varphi(t, u, f)$ be a bounded on $\mathbb{R}_+$ solution of equation (3), then under the condition (**F4**) the set $\varphi(\mathbb{R}_+, u, f) \subset C$ is pre-compact.*

Definition 12 A solution $\varphi(t, u_0, f)$ of Eq. (3) is said to be compact on $\mathbb{R}_+$ if the set $Q := \overline{\varphi(\mathbb{R}_+, u_0, f)}$ is a compact subset of C, where by bar we mean the closure in C and $\varphi(\mathbb{R}_+, u_0, f) := \{\varphi(t, u_0, f) : t \in \mathbb{R}_+\}$.

Let $f \in C(\mathbb{R} \times C, \mathbb{R}^n)$, $\sigma(t, f)$ be the motion (in the shift dynamical system $(C(\mathbb{R} \times C, \mathbb{R}^n), \mathbb{R}, \sigma)$) generated by $f, u_0 \in C, \varphi(t, u_0, f)$ be a solution of equation (3), $x_0 := (u_0, f) \in X := C \times H(f)$ and $\pi(t, x_0) := (\varphi(t, u_0, f), \sigma(t, f))$ be the motion of skew-product dynamical system $(X, \mathbb{R}_+, \pi)$.

A solution $\varphi(t, u_0, f)$ of Eq. (3) is called [6, Ch.I], [16, 17] *compatible* (respectively, *strongly compatible* or *uniformly compatible*) if the motion $\pi(t, x_0)$ is comparable (respectively, strongly comparable or uniformly comparable) with $\sigma(t, f)$ by character of recurrence.

4 A Class of Delay Differential Equations with a First Integral

Considered equation

$$x'(t) = f(t-1, x(t-1)) - f(t, x(t)), \tag{6}$$

where $f \in C(\mathbb{R} \times \mathbb{R}, \mathbb{R})$.

Denote by $H(f)$ the closure in $C(\mathbb{R} \times \mathbb{R}, \mathbb{R})$ of $\{f_\tau | \ \tau \in \mathbb{R}\}$ w.r.t. compact-open topology, where $f_\tau(t, u) := f(t + \tau, u)$ for any $(t, u) \in \mathbb{R} \times \mathbb{R}$.

Along with Eq. (6) we consider its H-class, i.e., the family of equations

$$x'(t) = g(t-1, x(t-1)) - g(t, x(t)) \ \ (g \in H(f)). \tag{7}$$

Assume that:

($\mathcal{F}1$) Eq. (6) is regular, i.e., for each Eq. (7) the conditions of existence, uniqueness and extendability on $\mathbb{R}_+$ are fulfilled;

($\mathcal{F}2$) the function f is non-decreasing in x;

($\mathcal{F}3$) $f \in C(\mathbb{R} \times \mathbb{R}, \mathbb{R})$ and for any bounded subset M from $\mathbb{R}$ there exists a positive constant $L = L(M, f)$ such that $|f(t, u_1) - f(t, u_2)| \leq L|u_1 - u_2|$ for any $(t, u_1), (t, u_2) \in \mathbb{R} \times M$;

($\mathcal{F}4$) $f(t, 0) = 0$ for any $t \in \mathbb{R}$.

Remark 4

1. Note that Eq. (6) is regular if f satisfies the following conditions: $f(t, 0) = 0$ for any $t \in \mathbb{R}$, f is monotone increasing and locally Lipschitz in $x \in C$ [2].
2. If the function $f \in C(\mathbb{R} \times \mathbb{R}, \mathbb{R})$ possesses the property ($\mathcal{F}3$) with the constant $L = L(M, f) > 0$, then every function $g \in H(f)$ possesses the property ($\mathcal{F}3$) with the same constant $L = L(M, f)$.

Along with Eq. (6) we consider equation

$$x'(t) = F(t, x_t), \tag{8}$$

where $F \in C(\mathbb{R} \times C, \mathbb{R})$ is defined by equality

$$F(t, \phi) := f(t-1, \phi(-1)) - f(t, \phi(0)) \tag{9}$$

for any $(t, \phi) \in \mathbb{R} \times C$.

Lemma 4 *For any compact subset K from C there exists a compact subset $\mathfrak{K}$ from $\mathbb{R}$ such that*

$$\max_{|t|\leq l,\ \phi\in K}|F(t+p,\phi)-F(t+q,\phi)|\leq$$

$$\max_{|t|\leq l,\ u\in\mathfrak{K}}|f(t+p-1,u)-f(t+q-1,u)|+\max_{|t|\leq l,\ u\in\mathfrak{K}}|f(t+p,u)-f(t+q,u)|$$

for any $p,q\in\mathbb{R}$ and $l>0$, where $f\in C(\mathbb{R}\times\mathbb{R},\mathbb{R})$ and $F\in C(\mathbb{R}\times C,\mathbb{R})$ is defined by (9).

Proof Let $p,q\in\mathbb{R}$, $l>0$ and K be a compact subset from C, then there exists a compact subset $\mathfrak{K}$ from $\mathbb{R}$ such that $\phi(s)\in\mathfrak{K}$ for any $s\in[-1,0]$ and $\phi\in K$. Note that

$$|F(t+p,\phi)-F(t+q,\phi)|\leq|f(t+p-1,\phi(-1))-f(t+q-1,\phi(-1))|+$$

$$|f(t+p,\phi(0))-f(t+q,\phi(0))|\leq\max_{|t|\leq l,\ \phi\in\mathfrak{K}}|f(t+p-1,u)-f(t+q-1,u)|+$$

$$\max_{|t|\leq l,\ \phi\in\mathfrak{K}}|f(t+p,u)-f(t+q,u)|$$

and, consequently,

$$\max_{|t|\leq l,\ \phi\in K}|F(t+p,\phi)-F(t+q,\phi)|\leq$$

$$\max_{|t|\leq l,\ u\in\mathfrak{K}}|f(t+p-1,u)-f(t+q-1,u)|+\max_{|t|\leq l,\ u\in\mathfrak{K}}|f(t+p,u)-f(t+q,u)|$$

for any $p,q\in\mathbb{R}$ and $l>0$. Lemma is proved. □

Note that on the space $C(\mathbb{R}\times\mathbb{R},\mathbb{R})$ (respectively, $C(\mathbb{R}\times C,\mathbb{R})$) is defined the compact-open topology and a shift dynamical system $(C(\mathbb{R}\times\mathbb{R},\mathbb{R}),\mathbb{R},\sigma)$ (respectively, $(C(\mathbb{R}\times C,\mathbb{R}),\mathbb{R},\sigma)$) with respect to time t.

Corollary 3 *The function $F\in C(\mathbb{R}\times C,\mathbb{R})$ defined by (9) is uniformly comparable by character of recurrence with $f\in C(\mathbb{R}\times\mathbb{R},\mathbb{R})$.*

Proof This statement directly follows from Lemma 4 and corresponding definition of uniformly comparability of functions by their character of recurrence. □

Corollary 4 *If the function $f\in C(\mathbb{R}\times\mathbb{R},\mathbb{R})$ is stationary (respectively, τ-periodic, quasi-periodic with the frequency basis $\nu_1,\nu_2,\ldots,\nu_m$, Bohr almost periodic, Levitan almost periodic, almost recurrent, recurrent) in $t\in\mathbb{R}$ uniformly w.r.t. u on every compact $\mathfrak{K}$ from $\mathbb{R}$, then the function F defined by (9) is also stationary (respectively, τ-periodic, quasi-periodic with the frequency basis $\nu_1,\nu_2,\ldots,\nu_m$, Bohr almost periodic, Levitan almost periodic, almost recurrent, recurrent) in $t\in\mathbb{R}$ uniformly w.r.t. ϕ on every compact K from C.*

Proof This statement follows from Theorems 1, 4 and Corollary 3. □

Lemma 5 *The following statements hold:*

1. *condition* $(\mathcal{F}1)$ *(respectively,* $(\mathcal{F}2)$*) implies* $(F1)$ *(respectively,* $(F2)$*) for Eq. (8);*
2. *if there exists a point* $x_0 \in \mathbb{R}$ *such that*

$$\|f\|_0 := \sup_{t\in\mathbb{R}} |f(t, x_0)| < +\infty, \tag{10}$$

then $(\mathcal{F}3)$ *implies* $(F3)$ *for Eq. (8).*

Proof The first statement is evident.

Let A be an arbitrary bounded subset of C and R be a positive number such that

$$\sup_{\phi\in A}(\|\phi\|_C + |x_0|) \le R.$$

To prove the second statement we note that if condition (10) holds, then by Lemma 2 condition $(F3)$ follows from $(\mathcal{F}3)$. Assume that condition $(\mathcal{F}3)$ is fulfilled. If A is a bounded subset of C, then we have

$$|F(t,\phi)| = |f(t-1,\phi(-1)) - f(t,\phi(0))| \le$$
$$|f(t-1,\phi(-1)) - f(t-1,x_0| + |f(t-1,x_0) - f(t,x_0)| +$$
$$|f(t,x_0) - f(t,\phi(0)| \le L(R)|\phi(-1) - x_0| + L(R)|x_0 - \phi(0)| + 2\|f\|_0\| \le$$
$$2L(R)(||\phi||_C + |x_0|) + 2\|f\|_0 \le 2L(R)R$$

and, consequently, the set $F(\mathbb{R}\times A)$ is a bounded subset from $\mathbb{R}^n$. Lemma is proved. □

Lemma 6 ([2]) *Let* ϕ *and* ϕ *be given in* C, $\varphi(t,\phi,f)$ *and* $\varphi(t,\psi,f)$ *the solutions of equation (6). Suppose that* $\psi \le \phi$*, then* $\varphi(t,\psi,f) \le \varphi(t,\phi,f)$ *for any* $t > 0$.

Let $C := C([-1,0],\mathbb{R})$ and $\varphi(t,\phi,g)$ be the solution of equation (7) passing through $\phi \in C$ at initial moment $t = 0$.

Define $V : C \times H(f) \to \mathbb{R}$ by

$$V(\phi,g) := \phi(0) + \int_{-1}^{0} g(\tau,\phi(\tau))d\tau.$$

It is easy to check that V possesses the following two properties:

(V1) For any $c \le d$ in $\mathbb{R}$ there is $M = M(c,d) > 0$ such that $|V(\phi,g) - V(\psi,g)| \le M\|\phi - \psi\|$ for any $\phi,\psi \in [c,d]_C := \{\phi \in C |\ c \le \phi(\tau) \le d$ for all $\tau \in [-1,0]\}$ and $g \in H(f)$;

(V2)

$$V(\phi, g) - V(\psi, g) \geq \phi(0) - \psi(0)$$

for any $\phi \geq \psi$ in C and $g \in H(f)$.

Lemma 7 ([13]) *The following statement holds:*

$$V(\pi(t, (\phi, g))) = V(\phi, g)$$

for any $(\phi, g) \in C \times H(f)$ *and* $t \geq 0$.

Lemma 8 *Under conditions* $(\mathcal{F}1)$*–*$(\mathcal{F}3)$ *for any* $c \leq d$ *in* $\mathbb{R}$ *and* $g \in H(f)$*, there exists* $K = K(c, d) > 0$ *such that*

$$\|\varphi(t, \phi, g) - \varphi(t, \psi, g)\| \leq K \|\phi - \psi\|$$

for any $\phi, \psi \in [c, d]_C$ *and* $t \geq 0$.

Proof This statement may be proved using the same ideas as in the proof of Lemma 6.3 from [13]. Below we will present the details of this proof.

For $c \leq d$ in $\mathbb{R}$ and $g \in H(f)$, let $M = M(b, c, f) > 0$ be defined as in the property (V1). We first show that

$$|\varphi(t, \phi, g) - \varphi(t, \psi, g)| \leq M \|\phi - \psi\| \tag{11}$$

for any $c \leq \psi \leq \phi \leq d$ and $t \in \mathbb{R}_+$. Indeed, by Lemma 6, we have $\varphi_t(\psi, g) \leq \varphi_t(\phi, g)$ for any $t \in \mathbb{R}_+$. It then follows from (V1), (V2) and Lemma 7 that

$$|\tilde{\varphi}(t, \phi, g) - \tilde{\varphi}(t, \psi, g)| = |\varphi(t, \phi, g)(0) - \varphi(t, \psi, g)(0)| =$$
$$\varphi(t, \phi, g)(0) - \varphi(t, \psi, g)(0) \leq V(\varphi(t, \phi, g), g_t) - V(\varphi(t, \psi, g), g_t) =$$
$$V(\pi(t, (\phi, g)) - V(\pi(t, (\psi, g)) = V(\phi, g) - V(\psi, g) \leq M \|\phi - \psi\|$$

for any $t \in \mathbb{R}_+$. For any $\phi, \psi \in [c, d]_C$, we define

$$\alpha(s) := \min_{s \in [-1,0]} \{\phi(s), \psi(s)\} \text{ and } \beta(s) := \max_{s \in [-1,0]} \{\phi(s), \psi(s)\}.$$

Clearly, $c \leq \alpha \leq \phi$, $\psi \leq \beta \leq d$. By the definition of α and β, it then follows that

$$\beta(s) - \alpha(s) = \big(\phi(s) + \psi(s) + |\phi(s) - \psi(s)|\big)/2 -$$
$$\big(\phi(s) + \psi(s) - |\phi(s) - \psi(s)|\big)/2 = |\phi(s) - \psi(s)|,$$

for any $s \in [-1, 0]$, which implies that

$$\|\alpha - \beta\| = \|\phi - \psi\|.$$

By (11), we have

$$|\varphi(t, \alpha, g) - \varphi(t, \beta, g)| \leq M\|\alpha - \beta\|,$$

for any $t \in \mathbb{R}_+$. Moreover, Lemma 6 implies that

$$\varphi(t, \alpha, g) \leq \varphi(t, \phi, g), \quad \varphi(t, \psi, g) \leq \varphi(t, \beta, g),$$

for any $t \in \mathbb{R}_+$. Thus, we obtain

$$|\varphi(t, \phi, g) - \varphi(t, \psi, g)| \leq |\varphi(t, \alpha, g) - \varphi(t, \beta, g)| \leq M\|\alpha - \beta\| = M\|\phi - \psi\|,$$

for any $t \in \mathbb{R}_+$. It then follows that

$$\|\varphi_t(\phi, g) - \varphi_t(\psi, g)\| \leq K\|\phi - \psi\|,$$

where $K := \max\{1, M\}$. □

Corollary 5 *Under conditions ($\mathcal{F}$ 1)–($\mathcal{F}$ 4) every solution $\varphi(t, \phi, g)$ of Eq. (7) is bounded on $\mathbb{R}_+$.*

Proof Let $\phi \in C$ and $c > 0$ such that $-c \leq \phi(s) \leq c$ for any $s \in [-1, 0]$ and $K = K(-c, c)$ be a positive constant from Lemma 8. By condition ($\mathcal{F}$ 4) we have $\varphi(t, 0, g) = 0$ for any $t \in \mathbb{R}$ and according to Lemma 8 we have

$$\|\varphi(t, \phi, g)\| \leq K\|\phi\|$$

for any $t \in \mathbb{R}_+$. □

Lemma 9 *Under conditions ($\mathcal{F}$1)–($\mathcal{F}$ 4) every solution $\varphi(t, u_0, f)$ of Eq. (6) is positively uniformly Lyapunov stable.*

Proof According to Lemmas 3 and 5 (item ii) $\varphi(\mathbb{R}_+, u_0, f)$ is a pre-compact subset of C and, consequently, there are $c_0 \leq d_0$ ($c_0, d_0 \in \mathbb{R}$) such that $c_0 \leq \varphi(t, u_0, f) \leq d_0$ for any $t \in \mathbb{R}_+$. Let δ_0 be a fixed positive number and

$$W_0 = [c_0 - \delta_0, d_0 + \delta_0]_C := \{\phi \in C | \; c_0 - \delta_0 \leq \varphi(s) \leq d_0 + \delta_0, \; \forall \, s \in [-1, 0]\}.$$

If we suppose that the statement of Lemma is false then there are $\varepsilon_0 > 0$, $\delta_n \to 0$ ($\delta_n > 0$) as $n \to \infty$, $t_n \geq 0$, $u_n \in W_0$ and $s_n \geq t_n$ such that

$$\|\varphi(t_n, u_n, f) - \varphi(t_n, u_0, f)\| < \delta_n \tag{12}$$

and

$$\|\varphi(s_n, u_n, f) - \varphi(s_n, u_0, f)\| \geq \varepsilon_0$$

for any $n \in \mathbb{N}$.

Since the sequence $\{\varphi(t_n, u_0, f)\} \subset \varphi(\mathbb{R}_+, u_0, f)$ is pre-compact and $\delta_n \to 0$ as $n \to \infty$, then from (12) it follows that the sequence $\{\varphi(t_n, u_n, f)\}$ is pre-compact too. Without loss of generality we can assume that $\delta_n \leq \delta_0$ for any $n \in \mathbb{N}$ and, consequently, $\{\varphi(t_n, u_n, f)\} \subseteq W_0$. Denote by $K_0 := K(c_0 - \delta_0, d_0 + \delta_0) > 0$ the positive constant figuring in Lemma 8. Let $\tau_n := s_n - t_n$. In virtue of Lemma 8 we have

$$\begin{gathered}\varepsilon_0 \leq \|\varphi(s_n, u_n, f) - \varphi(s_n, u_0, f)\| = \|\varphi(t_n + \tau_n, u_n, f) - \varphi(t_n + \tau_n, u_0, f)\| = \\ \|\varphi(\tau_n, \varphi(t_n, u_n, f), f^{t_n}) - \varphi(\tau_n, \varphi(t_n, u_0, f), f^{t_n})\| \leq \\ K_0\|\varphi(t_n, u_n, f) - \varphi(t_n, u_0, f)\| \leq K_0\delta_n \qquad (13)\end{gathered}$$

for any $n \in \mathbb{N}$. Passing to the limit in (13) as $n \to \infty$ we obtain $\varepsilon_0 \leq 0$. The last inequality contradicts to the choice of the number ε_0. The obtained contradiction proves our statement. The lemma is proved. □

Under conditions $(\mathcal{F}1)$–$(\mathcal{F}4)$ the following statement holds.

Lemma 10 ([13]) *Let $g \in H(f)$ and let $u(t)$ and $v(t)$ be two bounded solutions defined on $\mathbb{R}$ of (7)with $u(t) \leq v(t)$ for any $t \in \mathbb{R}$ there exists a number $\tau > 0$ such that whenever $u(s) < v(s)$ for some $s \in \mathbb{R}$, then $u(t) \ll v(t)$ for all $t \geq \tau + s$.*

Theorem 7 *Suppose that the function $f \in C(\mathbb{R} \times \mathbb{R}, \mathbb{R})$ is almost recurrent in $t \in \mathbb{R}$ uniformly with respect to u on every compact subset from $\mathbb{R}$.*

Then under Conditions $(\mathcal{F}1)$–$(\mathcal{F}4)$ the following statements hold:

1. *for any solution $\varphi(t, v, g)$ of Eq. (7) there exists a solution $\varphi(t, \gamma_v, g)$ of (7) defined and bounded on $\mathbb{R}$ such that:*
   a. *$\varphi(t, \gamma_u, g)$ is a strongly compatible solution of (7);*
   b. *$\lim\limits_{t \to \infty} |\varphi(t, v,)g - \varphi(t, \gamma_v, g)| = 0$;*
2. *if the function $f \in C(\mathbb{R} \times \mathbb{R}, \mathbb{R})$ is stationary (respectively, τ-periodic, quasi-periodic with the frequency basis $\nu_1, \nu_2, \ldots, \nu_m$, Bohr almost periodic, almost automorphic, almost recurrent, recurrent) in $t \in \mathbb{R}$ uniformly with respect to u on every compact subset from $\mathbb{R}$, then $\varphi(t, \gamma_u, f)$ is also stationary (respectively, τ-periodic, Bohr almost periodic, quasi-periodic with the frequency basis $\nu_1, \nu_2, \ldots, \nu_m$, almost automorphic, almost recurrent, recurrent).*

Proof Let $Y := H(f)$, $(Y, \mathbb{R}, \sigma)$ be the shift dynamical system on $H(f)$ and $\langle C, \varphi, (Y, \mathbb{R}, \sigma)\rangle$ be the cocycle generated by Eq. (6) (respectively, by family of equations (7)). Under the conditions of Theorem 7 the cocycle above possesses the following properties:

1. by Lemma 6 the cocycle φ is monotone;
2. according to Lemmas 3, 5 and Corollary 5 every solution $\varphi(t, \psi, g)$ of Eq. (7) is pre-compact;
3. in virtue of Lemma 9 every solution $\varphi(t, \psi, g)$ of Eq. (7) is positively uniformly Lyapunov stable;
4. by Lemma 10 the cocycle satisfies condition (**C4**).

To finish the proof of Theorem it is sufficient to apply Theorems 1, 4 and 5. □

Acknowledgments This research was supported by the State Program of the Republic of Moldova "Multivalued dynamical systems, singular perturbations, integral operators and non-associative algebraic structures (20.80009.5007.25)".

References

1. Arino, O.: Monotone semi-flows which have a monotone first integral. In: Lecture Notes in Mathematics, vol. 1475, pp. 64–75 (1991)
2. Arino, O., Seguier, P.: About the behaviour at infinity of solutions of $x'(t) = f(t-1, x(t-1)) - f(t, x(t))$. J. Math. Anal. Appl. **96**, 420–436 (1983)
3. Cheban, D.N.: Global pullback attractors of C-analytic nonautonomous dynamical systems. Stochastics Dyn. **1**(4), 511–535 (2001)
4. Cheban, D.N.: Asymptotically Almost Periodic Solutions of Differential Equations, ix+186 pp. Hindawi Publishing Corporation, New York (2009)
5. Cheban, D.N.: Global Attractors of Nonautonomous Dynamical and Control Systems, 2nd edn., vol. 18, xxv+589 pp. Interdisciplinary Mathematical Sciences. World Scientific, River Edge (2015)
6. Cheban, D.N.: Nonautonomous Dynamics: Nonlinear Oscillations and Global Attractors, xxii+434 pp. Springer Nature Switzerland AG (2020)
7. Cheban, D.: Almost periodic and almost automorphic solutions of monotone differential equations with a strict monotone first integral. Buletinul Academiei de Stiinte a Republicii Moldova. Matematica **3**, 39–74 (2020)
8. Cheban, D.: Poisson stable motions of monotone sub-linear non-autonomous dynamical systems (submitted)
9. Cheban, D.: On the structure of the Levinson center for monotone non-autonomous dynamical systems with a first integral. Carpatian J. Mat. **38**(1), 67–94 (2022)
10. Cheban, D., Liu, Z.: Poisson stable solutions of monotone differential equations. Sci. China Math. **62**(7), 1391–1418 (2019)
11. Cooke, K.L., Yorke, J.A.: Some equations modelling growth processes and gonorrhea epidemics. Math. Biosci. **16**, 75–101 (1973)
12. Hale, J.K.: Theory of Functional-Differential Equations, x+365 pp. Springer, New York (1977)
13. Jiang, J., Zhao, X.: Convergence in monotone and uniformly stable skew-product semiflows with applications. J. Reine Angew. Math. **589**, 21–55 (2005)

14. Levitan, B.M., Zhikov, V.V.: Almost Periodic Functions and Differential Equations, xi+211 pp. Cambridge University Press, Cambridge (1982)
15. Sell, G.R.: Lectures on Topological Dynamics and Differential Equations. Van Nostrand Reinhold Math. Studies, vol. 2. Van Nostrand–Reinbold, London (1971)
16. Shcherbakov, B.A.: Topologic Dynamics and Poisson Stability of Solutions of Differential Equations, 231 pp. Ştiinţa, Chişinău (1972; in Russian)
17. Shcherbakov, B.A.: The comparability of the motions of dynamical systems with regard to the nature of their recurrence. Differ. Equ. **11**(7), 1246–1255 (1975; in Russian) [English translation: Differ. Equ. **11**(7), 937–943]
18. Smith, H.L.: Monotone Dynamical Systems. An Introduction to the Theory of Competitive and Cooperative Systems. Mathematical Surveys and Monographs, vol. 41, x+174 pp. American Mathematical Society, Providence (1995)

Periodic Solutions in a Differential Delay Equation Modeling Megakaryopoiesis

Anatoli F. Ivanov and Bernhard Lani-Wayda

Abstract We consider a scalar nonlinear differential delay equation which was recently proposed as a mathematical model for platelet production (megakaryopoiesis). The equation has a unique positive equilibrium about which solutions tend to oscillate. We show that periodic oscillations in the model always exist when the equilibrium is linearly unstable. Several methods of proof are proposed. They include an adapted version of established ejective fixed point techniques, and application of a more recent theorem for nonlinear semiflows. We indicate how an analogous result can be obtained for a different class of equations frequently used in applications.

1 Introduction

It is well known that differential delay equations serve as mathematical models for a large variety of phenomena in applied sciences, in particular in biology and physiology [5, 7, 12, 15]. This contribution is motivated by the recent paper [3] where a relatively simple delay equation is proposed as a model of platelet production in human body (see Eq. (1) below). As it is observed from clinical data and measurements, the total number of platelets is stable with possible relatively small regular deviations in time. In the mathematical model such dynamics correspond to the existence of a unique positive equilibrium with typical oscillatory behavior of solutions about it. The physiological system exhibits the negative feedback property, when a deviation from the equilibrium in one direction forces the system to move in the opposite direction. This leads to the existence and typical nature of the so-called slowly oscillating solutions in the mathematical model, for which the time distance

A. F. Ivanov (✉)
Department of Mathematics, Pennsylvania State University, Dallas, PA, USA
e-mail: aivanov@psu.edu

B. Lani-Wayda
Mathematisches Institut der Justus-Liebig-Universität, Giessen, Germany
e-mail: Bernhard.Lani-Wayda@math.uni-giessen.de

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_4

between consecutive passages through the equilibrium level is greater than the delay time in the system.

Important mathematical questions arise then for the differential delay model under consideration, which would provide theoretical explanation of the observed phenomena. The first one is the global stability of the unique equilibrium, when any initial function results in a solution that is attracted by the equilibrium. This question is studied and answered affirmatively in paper [10]. The second one is the existence of slowly oscillating periodic solutions, which is answered in the present work. We show that the mere instability of the equilibrium implies the existence of such periodic solutions, which are likely to be stable and hence experimentally observable (although our methods do not prove the stability). Existence of periodic solutions is an information significantly beyond oscillatory behavior of most solutions. We would also like to mention that some initial studies on the qualitative analysis of the mathematical model are initiated in the original paper [3].

We consider the differential delay equation

$$\dot{x}(t) = -\mu x(t) + f(x(t))\, g(x(t-\tau)), \tag{1}$$

with strictly decreasing and positive C^1 functions $f, g : [0,\infty) \to \mathbb{R}$ having negative derivative, and $\mu, \tau > 0$. It has a unique equilibrium $x^* > 0$ as the positive solution of the equation $f(x)g(x) = \mu x$. Setting $\tilde{f}(y) := f(x^* + y) - f(x^*)$, $\tilde{g}(y) := g(x^* + y) - g(x^*)$, one has $\tilde{f}(0) = 0 = \tilde{g}(0)$, $\tilde{f}'(0) = f'(x^*) < 0$, $\tilde{g}'(0) = g'(x^*) < 0$, and the negative feedback properties $\tilde{f}(y)\cdot y < 0$, $\tilde{g}(y)\cdot y < 0$ $(y \neq 0)$. Setting $y(t) := x(t) - x^*$ transforms Eq. (1) to

$$\dot{y}(t) = -\mu y(t) + \tilde{f}(y(t)) \cdot g(x^* + y(t-\tau)) + f(x^*)\tilde{g}(y(t-\tau)). \tag{2}$$

The linearization of equation (2) at zero is given by

$$\begin{aligned}\dot{y}(t) &= -\mu y(t) + f(x^*)g'(x^*)y(t-\tau) + f'(x^*)g(x^*)y(t)\\ &= -Ay(t) - By(t-\tau),\end{aligned}$$

with the positive numbers $A := \mu - f'(x^*)g(x^*)$ and $B := -f(x^*)g'(x^*)$. The associated characteristic equation

$$\lambda + A + Be^{-\lambda\tau} = 0 \tag{3}$$

is well-studied, e.g., in [8]. We make an assumption of instability, which automatically implies oscillatory behavior, concerning the eigenvalue (solution of (3)) with maximal real part:

The characteristic equation has a leading unstable eigenvalue $\lambda = \rho + i\omega$ with $\rho > 0$. (4)

Remark 1 Define $s_{\min}$ as the smallest positive solution of the equation $\tau A + s \cdot \cos[\sqrt{s^2 - (\tau A)^2}] = 0$. Condition (4) is satisfied if and only if

$$B\tau > s_{\min}. \tag{5}$$

Then the characteristic equation (3) has no real roots, and $\omega \in (\pi/(2\tau), \pi/\tau)$.

Proof Equation (3) is equivalent to $\tau\lambda + \tau A + B\tau e^{-\lambda\tau} = 0$, and the latter has roots with positive real part if and only if condition (5) holds, as shown in [8]. Note that $s_{\min}$ has to satisfy $s_{\min}^2 > (\tau A)^2 + \pi^2/4$, so it follows from (5) that $B\tau > \pi/2$. Now if λ solves (3) then $\tilde{\lambda} := \tau(\lambda + A)$ solves $\tilde{\lambda} + B\tau e^{\tau A} e^{-\tilde{\lambda}} = 0$. It is well known that this equation has no real roots if $B\tau e^{\tau A} > e^{-1}$ (see e.g. Proposition 4 in [3]), and under condition (5) one has $B\tau e^{\tau A} > B\tau > \pi/2 > e^{-1}$, so that $\tilde{\lambda}$ (and hence λ) cannot be real. The inequalities for ω are proved in [8]. □

It is shown in [3] that, in absence of real eigenvalues, Eq. (1) exhibits oscillations about x^*. We mention the result on sustained oscillations by O. Arino in this context [1]. Recall that a slowly oscillating solution is such that the distance between any two zeros of $x(t) - x_*$ is greater than the delay τ. In the present note we show that the oscillatory behavior includes slowly oscillating *periodic* solutions, with one positive and one negative semi-cycle.

We assume that for every $\varphi \in C$ there exists a unique solution $x^\varphi(t)$ to Eq. (1) or (2) defined for all $t \geq 0$. Such existence is guaranteed e.g. by the assumption of Lipschitz continuity of nonlinearity f. Then Eqs. (1) and (2) induce semiflows on the space $C = C^0([-\tau, 0], \mathbb{R})$ with the max-norm $||\cdot||_\infty$. The solution segment $x_t \in C$ is defined as $x_t = x^\varphi(t + s)$, $s \in [-\tau, 0]$. The set of non-negative initial functions is invariant under the semiflow for (1), in accordance with the biological interpretation.

2 Cone and Return Map

We set

$$K := \{\varphi \in C \mid \varphi = 0 \text{ on } [-\tau, z^*] \text{ for some } z^* \in [-\tau, 0], \text{ and } \varphi > 0 \text{ on } (z^*, 0]\}.$$

It is proved in [10], Proposition 2.3, that there exist positive numbers m, M with $m < x^* < M$ such that the order interval $\{\varphi \in C \mid m \leq \varphi \leq M\}$ is attracting and invariant for the semiflow induced by Eq. (1). For the transformed equation (2) this implies that all its solutions y^φ satisfy $y^\varphi(t) \in [m^+, M^+]$ $\forall t \geq 0$ for arbitrary initial function φ with $\varphi(s) \in [m^+, M^+]$ $\forall s \in [-\tau, 0]$, where $m^+ = m - x^*$, $M^+ = M - x_*$.

Proposition 1 (Oscillating Solutions) *For a solution $y = y^\varphi$ of (2) with $\varphi \in K, \varphi \le M^+$ one has $y(t) \in [m^+, M^+]$ for all $t \in [-\tau, \infty)$. If y has two consecutive zeroes z_j, z_{j+1} then the maximum of $|y|$ between them occurs in $[z_j, z_j + \tau]$.*

Proof The corresponding solution x of (1) has initial function $\varphi + x^*$ with values in $[x^*, M^+ + x^*] \subset [m, M]$ and hence only values in the last interval. The corresponding bounds for y follow. Consider now two consecutive zeroes z_j, z_{j+1} with $y > 0$ on (z_j, z_{j+1}). If $y(t) > 0$ and $y(t-\tau) > 0$ for some t then Eq. (2) shows $\dot{y}(t) < 0$, hence the maximum of y on $[z_j, z_{j+1}]$ occurs in $[z_j, z_j + \tau]$. An analogous argument in case $y < 0$ on $[z_j, z_j + \tau]$ proves the assertion about maxima of $|y|$. □

We can modify f and g outside the interval $[m, M]$ to bounded C^1 functions $\hat{f}, \hat{g}$ with the same properties; the corresponding changes of $\tilde{f}, \tilde{g}$ do not affect solutions as described in Proposition 1. In particular, we then have a constant $L > 0$ such that $|\tilde{f}(y)| \le L|y|$ for all y. Choose now $\nu > \mu + g(x^*) \cdot L$ and define

$$K_\nu := \{\varphi \in K \mid t \mapsto \varphi(t)\, e^{\nu t} \text{ is increasing on } [-\tau, 0]\} \cup \{0\}.$$

Note that K_ν is a closed, convex cone in C, and that for $\varphi \in K_\nu$ one has

$$||\varphi||_\infty \ge \varphi(0) \ge e^{-\nu\tau} ||\varphi||_\infty. \tag{6}$$

Proposition 2 *For $\varphi \in K$, the corresponding solution $y = y^\varphi$ of Eq. (2) satisfies $y > 0$ on $(z^*, z^* + \tau]$ (with z^* as in the definition of K), is defined on $[-\tau, \infty)$, and has infinitely many zeroes $z_1 < z_2 < z_3 \ldots$ in $(0, \infty)$, all of which are simple, and $z_{j+1} - z_j > \tau$ $(j \in \mathbb{N})$. (I.e., y is slowly oscillating.) The solution segments $y_{z_j+\tau}$ satisfy $y_{z_j+\tau} \in (-1)^j K_\nu$ $(j \in \mathbb{N})$.*

Proof On $[0, z^* + \tau]$, Eq. (2) reduces to the ODE $\dot{y}(t) = -\mu y(t) + \tilde{f}(y(t))g(x^*)$ with zero as an equilibrium; the uniqueness of solutions and $y(0) > 0$ imply $y > 0$ on $[0, z^* + \tau]$. For $t > z^*$ with $y(t) > 0$ and $y(t-\tau) > 0$, Eq. (2) shows $\dot{y}(t) < 0$. Hence, if we had $y(t) > 0$ for all $t \ge 0$, then $y(t) \to 0$ monotonically. The condition on absence of real characteristic values excludes this (see, e.g., [3], Proposition 4 and Theorem 4.1), hence y has a first positive zero $z_1 > z^* + \tau$, and $\dot{y}(z_1) < 0$. With $\nu > 0$ chosen as above, consider now $w(t) := e^{\nu t} y(t)$ on $[z_1, z_1 + \tau]$. As long as $y(t) < 0$ (so certainly close to the right of z_1),

$$\begin{aligned}
\dot{w}(t) &= e^{\nu t}[\nu y(t) + \dot{y}(t)] \\
&= e^{\nu t}[(\nu - \mu)y(t) + \underbrace{\tilde{f}(y(t))}_{\le L|y(t)| = -Ly(t)} \underbrace{g(x^* + y(t-\tau))}_{\le g(x^*)} + \underbrace{f(x^*)\tilde{g}(y(t-\tau))}_{\le 0}] \\
&\le e^{\nu t} \underbrace{[\nu - \mu - g(x^*)L]}_{>0} y(t),
\end{aligned}$$

hence $\dot{w}(t) < 0$ as long as $y(t) < 0$. If y had a first zero z in $(z_1, z_1+\tau]$ then $\dot{w} < 0$ on $[z_1, z)$ would imply $w(z) < 0$ and hence $y(z) < 0$, a contradiction. Thus $y < 0$ and $\dot{w} < 0$ on $(z_1, z_1+\tau]$. It follows from

$$\frac{d}{dt}[e^{\nu t} y(z_1+\tau+t)] = \frac{d}{dt}[e^{\nu(z_1+\tau+t)} y(z_1+\tau+t)] \cdot e^{-\nu(z_1+\tau)}$$
$$= \dot{w}(z_1+\tau+t) e^{-\nu(z_1+\tau)} \le 0, \quad t \in [-\tau, 0],$$

that $y_{z_1+\tau} \in -K_\nu$. As above, $y(t) < 0$ for all $t > z_1$ is impossible, and with the same argument as for z_1 we obtain inductively the sequence (z_j) as asserted. □

Corollary 1 *The map $K \ni \varphi \mapsto y^{\varphi}_{z_2(\varphi)+\tau}$ maps K continuously into K_ν.*

Proof We see from Proposition 2 that this map takes values in K_ν. The continuity follows from the simplicity of zero z_2 and from continuity of the semiflow. □

Proposition 3 (Bound for Return Times) *There exists $T_1 > 0$ such that for all $\varphi \in K_\nu$ one has $z_2(\varphi) + \tau \le T_1$.*

Proof Using Lipschitz bounds for $\tilde{f}$ and $\tilde{g}$, a bound for g, and (6), one sees that there exist constants $c_1, c_2 > 0$ such that for a solution y with $y_0 = \varphi \in K_\nu$ one has

$$||\dot{y}|_{[0,\tau]}||_\infty \le c_1(||y_\tau||_\infty + ||\varphi||_\infty) \le c_1[||y_\tau||_\infty + e^{\nu\tau}\varphi(0)]$$
$$\le c_1[||y_\tau||_\infty + e^{\nu\tau}||y_\tau||_\infty] = c_2||y_\tau||_\infty.$$

Assume now that there exists a sequence (φ_n) in K_ν such that $z_1(\varphi_n) \to \infty$. The sequence $(y^{\varphi_n}_\tau)$ and, in view of the above estimate, also the rescaled sequence defined by $\eta_n := \dfrac{y^{\varphi_n}_\tau}{||y^{\varphi_n}_\tau||_\infty}$ both satisfy the conditions of the Arzelà–Ascoli theorem, so we can assume both are convergent. If $y^{\varphi_n}_\tau \to \psi^* \ne 0 \in C$, then $\psi^* \ge 0$, the solution of equation (2) with initial segment ψ^* has a first simple zero z_1^*, and hence y^{φ_n} has a simple zero close to z_1^*, a contradiction. In case $y^{\varphi_n}_\tau \to 0$, a familiar argument (as, e.g., in [11], proof of Lemma 5.7) shows that the solutions y^{φ_n} satisfy a non-autonomous linear equation with coefficients converging to the ones of the linearized equation, uniformly on compact intervals. The rescaled solutions $\dfrac{y^{\varphi_n}}{||y^{\varphi_n}_\tau||_\infty}$ satisfy the same non-autonomous equation, and their segments at time τ (namely, η_n) have a limit $\zeta^* \ne 0$. The solution of the linearized equation with initial segment ζ^* has a first simple zero $\hat{z}_1$, and it follows that the rescaled solutions (and hence also the y^{φ_n}) have a first zero close to $\hat{z}_1$ for large n, again a contradiction.

It follows that z_1 is bounded on K_ν. With an analogous argument one sees that z_2 is bounded as well, since the segments $-y^{\varphi}_{z_1(\varphi)}$, $\varphi \in K_\nu$, are again in K_ν. □

Remark 2 In the above proof, the estimate for $||\dot{y}|_{[0,\tau]}||_\infty$, which uses the particular form of the equation, could be replaced by using the general fact that

within the class of slowly oscillating solutions there is a constant $k > 0$ such that the inequality $||y_{t+1}||_\infty \geq k||y_t||_\infty$ is always satisfied (uniformly non-superexponential decay); see e.g. [6], where the uniformity is not explicitly stated in Theorem 2.2, but actually proved in Proposition 2.8.

It follows from Corollary 1 and Proposition 3 that the return map

$$P : K_\nu \to K_\nu, \varphi \mapsto y^\varphi_{z_2(\varphi)+\tau} \text{ if } \varphi \neq 0,\ P(0) = 0$$

(as in Theorem 3.4 of [11], Theorem 2 below in the present work) is well-defined, continuous (use that the semiflow is uniformly continuous on $[0, T_1] \times \{0_C\}$) and compact (since certainly $z_2(\varphi) + \tau \geq \tau$).

For $\varphi \in C$, let $\pi_\lambda \varphi$ denote the complex spectral projection of φ to the (complex) one-dimensional subspace of $C^0([-\tau, 0], \mathbb{C})$ corresponding to the leading eigenvalue $\lambda = \rho + i\omega$; then $(\pi_\lambda \varphi)(t) = \Pi(\varphi) \cdot e^{\lambda t}$ for $t \in [-\tau, 0]$, with a complex linear functional Π. Associated to λ is a (real) two-dimensional subspace U of C spanned by ψ_1, ψ_2, where $\psi_1(t) = e^{\rho t}\cos(\omega t)$ and $\psi_2(t) = e^{\rho t}\sin(\omega t)$. The real spectral projection of $\varphi \in C$ to U is $\pi_U \varphi = c_1\psi_1 - c_2\psi_2$, where $c = \Pi(\varphi) = c_1 + ic_2$ with Π as above. To provide a lower bound for π_U, it is sufficient to provide a lower bound for the functional Π. Up to a nonzero constant factor, Π is given by

$$\tilde{\Pi}\varphi := \varphi(0) - B \cdot J(\varphi), \text{ where } J(\varphi) := \int_{-\tau}^{0} e^{-\lambda(\tau+s)}\varphi(s)\, ds$$

(see Corollary 2.5 in [11]).

Proposition 4 (Lower Bound for Spectral Projection) *There exists $c > 0$ such that $|\tilde{\Pi}\varphi| \geq c||\varphi||_\infty$ for $\varphi \in K_\nu$, with an analogous estimate for Π.*

Proof Consider $\varphi \in K_\nu$. We write $||\varphi||_{L^1}$ for $\int_{-\tau}^{0} |\varphi(x)|\, dx$. From $\mathrm{Re}(\lambda) > 0$ and (6) we see that for $\varphi \in K_\nu$

$$|J(\varphi)| \leq ||\varphi||_{L^1} \leq \tau \cdot ||\varphi||_\infty \leq \tau e^{\nu\tau}\varphi(0). \tag{7}$$

We have $\omega \in (\frac{\pi}{2\tau}, \frac{\pi}{\tau})$, see Remark 1. It follows that

$$\mathrm{Im}(J(\varphi)) = \int_{-\tau}^{0} e^{-\rho(\tau+s)} \cdot \underbrace{(-\sin(\omega(\tau+s))}_{\leq 0} \cdot \varphi(s)\, ds.$$

Thus, with $\sigma := \min_{-\tau/2 \leq s \leq 0} |\sin(\omega(\tau+s)| > 0$, and $\tilde{\varphi}(s) := e^{\nu s}\varphi(s)$, we obtain

$$|\,\mathrm{Im}(J(\varphi))| \geq \sigma \cdot \int_{-\tau/2}^{0} e^{-\rho(\tau+s)}e^{-\nu s}e^{\nu s}\varphi(s)\, ds \geq \sigma_1 \int_{-\tau/2}^{0} \tilde{\varphi}(s)\, ds,$$

with a constant $\sigma_1 > 0$. Since $\tilde{\varphi}$ is increasing, we can conclude

$$|\operatorname{Im}(J(\varphi))| \geq \frac{\sigma_1}{2} \int_{-\tau}^{0} \tilde{\varphi}(s)\, ds \geq \underbrace{\frac{\sigma_1}{2} e^{-\nu\tau}}_{=:\sigma_2} \int_{-\tau}^{0} \varphi(s)\, ds = \sigma_2 ||\varphi||_{L^1}. \tag{8}$$

Now if $||\varphi||_{L^1} \leq \frac{\varphi(0)}{2B}$ then we see from (7) that

$$|\tilde{\Pi}\varphi| \geq \varphi(0) - B|J(\varphi)| \geq \varphi(0) - B \cdot \frac{\varphi(0)}{2B} = \frac{\varphi(0)}{2} \geq \frac{e^{-\nu\tau}}{2} ||\varphi||_{\infty}.$$

If however $||\varphi||_{L^1} \geq \frac{\varphi(0)}{2B}$ then (8) shows

$$|\tilde{\Pi}\varphi| \geq |\operatorname{Im}(\tilde{\Pi}\varphi)| = |\operatorname{Im}(J(\varphi))| \geq \sigma_2 ||\varphi||_{L^1} \geq \frac{\sigma_2}{2B}\varphi(0) \geq \frac{\sigma_2}{2B} e^{-\nu\tau} ||\varphi||_{\infty}.$$

The asserted lower estimate for $\tilde{\Pi}$ and hence for Π follows. □

Remark 3 In the passage following Lemma 2.9 in [11], it was mentioned that the conditions for a lower bound of the spectral projection are automatically satisfied for dimensions $N = 1, 2, 3$; however, this is true only for $N = 2, 3$, while for $N = 1$ one has to use the argument of Proposition 4 instead.

3 Periodic Solutions

Theorem 1 *Under the instability assumption (4), Eq. (2) (and hence Eq. (1)) has a slowly oscillating periodic solution, repeating after one positive and one negative semi-cycle.*

Based on the above preliminaries, we indicate three methodically different approaches to the proof of this result. The preparations from Sect. 2 provide a detailed proof in case of approaches I and III; we only sketch the ideas of approach II.

I. Recall that the fixed point $0 \in K$ is called ejective under the map P if there exists an open neighborhood $0 \ni U \subset C$ such that for every $\varphi \in K \cap U, \varphi \neq 0$, there is an integer $m = m(\varphi)$ such that $P^m(\varphi) \notin K \cap U$. Basics of the ejective fixed point theory, as applied to periodic solutions of differential delay equations, can be found in monographs [4, 9].

For the return map P from Section 2, a reasoning similar to the one in [8] shows that 0 is an ejective fixed point. Hence, as in the paper [8], Browder's ejective fixed point theorem (see Theorem 2, p. 88 in [8]) implies the existence of a nonzero fixed point of P in the cone K_ν, leading to a periodic solution of the described type. We carry out the proof of ejectivity, assuming that f and g are of class C^2.

Equation (2) can be rewritten as

$$\dot{y}(t) = F(y(t), y(t-\tau)), \tag{9}$$

where $F(y,z) := -\mu y + \tilde{f}(y)g(x^* + z) + f(x^*)\tilde{g}(z)$. We have $F(0,0) = 0$, and

$$\partial_1 F(0,0) = -\mu + \tilde{f}'(0)g(x^*) < -\mu < 0, \quad \partial_2 F(0,0) = f(x^*)\tilde{g}'(0) < 0.$$

With the positive numbers $A := -\partial_1 F(0,0)$ and $B := -\partial_2 F(0,0)$, and the nonlinear part

$$H(y,z) := F(y,z) - DF(0,0)(y,z),$$

Eq. (9) can be written as

$$\dot{y}(t) + Ay(t) + By(t-\tau) = H(y(t), y(t-\tau)). \tag{10}$$

The associated characteristic equation (3) is solved, in particular, by the leading eigenvalue $\lambda = \rho + i\omega$. We have $\rho > 0$ and, from Corollary to Lemma 3 of [8], p. 89, $\pi/(2\tau) < \omega < \pi/\tau$.

There exist $\delta_0 > 0$ and $c_2 > 0$ such that if $\delta \in (0, \delta_0]$ and $y, z \in [-\delta, \delta]$ then

$$|H(y,z)| \le c_2\delta^2. \tag{11}$$

Now assume that 0 is not ejective, and consider a solution $y : [-\tau, \infty) \to \mathbb{R}$ of (9) with initial segment $0 \neq \varphi \in K_v$ such that $\sup_{t\ge 0} |y(t)| = \delta$, for some $\delta \in (0, \delta_0]$. In view of Proposition 1, there exists a zero z_j of y such that $||y_{z_j+\tau}||_\infty \ge \delta/2$, and Proposition 2 shows $y_{z_j+\tau} \in \pm K_v$. Since $y(z_j + \tau + \cdot)$ is also a solution of equation (2), we can assume

$$y_0 = \varphi \in K_v, \, ||\varphi||_\infty \ge \delta/2, \text{ and } |y(t)| \le \delta \le \delta_0 \text{ for all } t \ge -\tau.$$

Recall the Laplace transform defined by $(\mathcal{L}y)(\lambda) := \int_0^\infty e^{-\lambda t} y(t)\, dt$. It has the well-known properties (see [11], formula (2.2))

$$(\mathcal{L}\dot{y})(\lambda) = -x(0) + \lambda(\mathcal{L}y)(\lambda)$$
$$[\mathcal{L}y(\cdot - \tau)](\lambda) = e^{-\lambda\tau}\left[\int_{-\tau}^0 e^{-\lambda t} y(t)\, dt + (\mathcal{L}y)(\lambda)\right].$$

Inspired by [8] (pages 92-93), we now apply the Laplace transform with the leading eigenvalue λ to Eq. (10), and we abbreviate the right hand side of that equation as $h(t)$. This gives

$$-y(0) + \lambda(\mathcal{L}y)(\lambda) + A(\mathcal{L}y)(\lambda) + Be^{-\lambda\tau}\int_{-\tau}^0 e^{-\lambda t} y(t)\, dt + Be^{-\lambda\tau}(\mathcal{L}y)(\lambda) = (\mathcal{L}h)(\lambda).$$

In view of the characteristic equation (3), the terms with $(\mathcal{L}y)(\lambda)$ cancel out, and one obtains

$$-y(0) + Be^{-\lambda\tau}\int_{-\tau}^{0} e^{-\lambda t}y(t)\,dt = (\mathcal{L}h)(\lambda),$$

or, with $\tilde{\Pi}$ as in Proposition 4,

$$\tilde{\Pi}\varphi = -(\mathcal{L}h)(\lambda).$$

(Compare Propositions 2.2. and 2.3 from [11], where the relation between the spectral projection and the Laplace transform of the linear part of the equation is detailed.) From (11) we obtain a constant $\tilde{c}_2 > 0$ such that $|-(\mathcal{L}h)(\lambda)| \le \tilde{c}_2\delta^2$, and from Proposition 4 we see that $|\tilde{\Pi}\varphi| \ge c||\varphi||_\infty \ge c\delta/2$, so

$$c\delta/2 \le |\tilde{\Pi}\varphi| = |(\mathcal{L}h)(\lambda)| \le \tilde{c}_2\delta^2,$$

which gives a contradiction for small enough δ. It follows that there exists a number $\delta_1 \in (0, \delta_1]$ such that every nonzero solution y with $y_0 \in K_\nu$ satisfies $\sup_{t\ge 0}|y(t)| \ge \delta_1$. This proves ejectivity of the return map P at zero, and Theorem 1 follows from the ejective fixed point theorem, as in [8].

II. Using the modification of f and g to functions bounded in $C^1(\mathbb{R}, \mathbb{R})$, as indicated after Proposition 1, one can consider a homotopy of the form

$$\dot{y}(t) = -\mu y(t) + (1-s)\cdot\tilde{f}(y(t))\cdot g(x^* + y(t-\tau)) + s\cdot f'(x^*)\cdot y(t)\cdot g(x^*) + \\ + f(x^*)\tilde{g}(y(t-\tau)),\ s \in [0,1],$$

which transforms Eq. (2) (for $s = 0$) to the equation

$$\dot{y}(t) = [-\mu + f'(x^*)\cdot g(x^*)]\cdot y(t) + f(x^*)\tilde{g}(y(t-\tau))$$

(for $s = 1$). The latter is of the type considered in Proposition 2.1 of [14], and the homotopy satisfies the admissibility conditions of Theorem D from [14], p. 282 for the case $N = 1$, which then gives a slowly oscillating periodic solution. (The results from [14] formally require f and g to be C^∞, but that is a minor issue.) In particular, the feedback conditions and the linearized equation are preserved throughout the homotopy for all $s \in [0, 1]$, as well as the a-priori bounds and bounds on return times, which apply in particular to periodic solutions.

One caution is in place here: In order to have that the fixed point index of all return maps P_s associated to the equation with homotopy parameter s and to the set of slowly oscillating solutions (which was computed to be $+1$ for a prototype equation in [14]) actually indicates existence of a nontrivial periodic solution, one has to separate the set $\{\varphi \mid P_s(\varphi) = \varphi \text{ for some } s \in [0,1]\}$ from zero—for this purpose one may require in addition that zero is hyperbolic for Eq. (2). Then a

sufficiently small neighborhood of zero cannot contain periodic solutions (due to the saddle point property), and, due to the bound on period lengths; also, no initial values of periodic solutions, since such solutions would have to have 'long' periods. Compare the remarks on p. 303 of [14], before formula (9.2).

III. With the splitting $C = U \oplus S$, where U is the (real) eigenspace associated to λ and S is the complementary spectral subspace associated to the remaining eigenvalues (with real parts less than that of λ), conditions (A1) and (A2) of Theorem 3.4 from [11] for semiflows on Banach spaces are satisfied, even with the slightly stronger form ($\widetilde{\text{A1}}$) instead of (A1).

($\widetilde{\text{A1}}$) $(E, |\cdot|_E)$ is a Banach space, and $\Phi : \mathbb{R}_0^+ \times E \to E$ is a continuous semiflow, $\Phi(t, 0) = 0$ for all t, $D_2\Phi$ exists on $\mathbb{R}_0^+ \times E$, and is continuous as a mapping into $L_c(E, E)$.

(A2) The operators $T(t) := D_2\Phi(t, 0) \in L_c(E, E)$ form a C^0-semigroup of linear operators. There exist real numbers $\alpha < \beta$ with $\beta > 0$ and a decomposition $E = U \oplus S$ into $T(t)$-invariant closed subspaces, where $U \neq \{0\}$, and a constant $K > 0$ such that

$$\forall t \geq 0 : \ |T(t)u|_E \geq K^{-1} e^{\beta t} |u|_E \ (u \in U), \quad |T(t)s|_E \leq K e^{\alpha t} |s|_E \ (s \in S).$$

We quote this theorem, only with different notation for the return time, and under the slightly stronger assumption ($\widetilde{\text{A1}}$):

Theorem 2 *Assume that the semiflow $\Phi : \mathbb{R}_0^+ \times E \to E$ on the Banach space $(E, |\cdot|_E)$ satisfies assumptions ($\widetilde{\text{A1}}$) and* (A2). *Let $\{0\} \neq \mathfrak{K} \subset E$ be closed and convex with $0 \in \mathfrak{K}$. Assume that $0 < t_1 < T_1$, that map $\theta : \mathfrak{K} \setminus \{0\} \to [t_1, T_1]$ is continuous, and $\Phi(\theta(\psi), \psi) \in \mathfrak{K}$ for $\psi \in \mathfrak{K} \setminus \{0\}$. Define $P : \mathfrak{K} \to \mathfrak{K}$ by*

$$P(0) := 0, \qquad P(\psi) := \Phi(\theta(\psi), \psi) \text{ for } \psi \neq 0.$$

We further assume: (1) P is compact; (2) $P(\psi) \neq 0$ if $\psi \neq 0$;
(3) $\exists c > 0 : \ \forall \varphi \in \mathfrak{K} : ||\pi_U \varphi|| \geq c||\varphi||$;
(4) There exist a continuous linear functional $\eta : E \to \mathbb{R}$ and $c_1 > 0$ such that

$$\forall \varphi \in \mathfrak{K} : c_1 |\varphi|_E \leq \eta(\varphi).$$

Then P has a fixed point φ^* in $\mathfrak{K} \setminus \{0\}$, corresponding to a periodic trajectory $\Phi(\cdot, \varphi^*)$ of the semiflow with period $\theta(\varphi^*)$.

With K_ν in place of $\mathfrak{K}$ and with the return map P constructed above in Sect. 2, conditions (1) and (2) are satisfied (the return time is clearly bounded from below by τ, and bounded above by T_1 from Proposition (3). Proposition 4 shows that condition (3) holds. As in [11], the linear functional η is simply $\eta(\varphi) = \varphi(0)$, for which the lower bound on K_ν is given in (6), so condition (4) holds. Thus Theorem 1 follows from Theorem 2.

Theorem 3.4 from [11] was stated for general semiflows, with the intention of further applications. We find such an application in the platelet production model from paper [3].

4 Further Extensions

Closely related to (1) is the following differential delay equation

$$\dot{x}(t) = F(x(t-\tau)) - G(x(t)). \tag{12}$$

It serves as mathematical model for a number of biological phenomena [12], as well as a model describing market fluctuations studied in economics [2, 13]. Equation (12) is considered under the following basic assumptions induced by applications:

(H_1) Functions F and G are defined and continuously differentiable on the positive semiaxis $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geq 0\}$ and are positive for all $x > 0$. In addition G is increasing with $G'(x) > 0\ \forall x \in \mathbb{R}_+$;

(H_2) There is a unique positive value $x_* > 0$ such that $F(x_*) = G(x_*)$, and in addition the following inequalities are satisfied

$$F(x) > G(x)\ \forall x \in (0, x_*) \quad \text{and} \quad F(x) < G(x)\ \forall x \in (x_*, \infty). \tag{13}$$

Assumption (H_2) implies the existence of the unique positive equilibrium x_* in model (12), in agreement with the applied interpretation. An important property in addition to (13), frequently required of model (12), is the (non-local) negative feedback assumption about the unique positive equilibrium. For this purpose the well-defined interval map $\Phi := G^{-1} \circ F$ is introduced with the following assumption in place:

(H_3) $F'(x_*) < 0$ and there exists a closed finite interval $I_0 = [a, b] \subset \mathbb{R}_+$, $x_* \in I_0$, such that $\Phi(I_0) \subseteq I_0$ and $(\Phi(x) - x_*)(x - x_*) < 0\ \forall x \in [a, b], x \neq x_*$.

In case of monotone decreasing F the hypothesis (H_2) and the negative feedback assumption (H_3) are satisfied automatically. The linearization of (12) about the positive equilibrium x_*, $\dot{y}(t) = F'(x_*)\, y(t-\tau) - G'(x_*)\, y(t)$, produces the same characteristic equation (3) as above, with $A = G'(x_*) > 0$ and $B = -F'(x_*) > 0$. The main periodicity result of Sect. 3, Theorem 1, extends to Eq. (12) as follows.

Theorem 3 *Assume that hypotheses (H_1), (H_2), and (H_3) are fulfilled, and that the corresponding characteristic equation (3) has a solution with positive real part. Then the differential delay equation (12) has a non-trivial periodic solution slowly oscillating about the equilibrium x_*.*

The proof is accomplished along the same lines as the proof of Theorem 1. Theorem 3 generalizes a similar result from [12].

Acknowledgments We thank the referees for useful remarks that helped us to improve the final presentation.

References

1. Arino, O.: A Note on "The Discrete Lyapunov Function". J. Differ. Equ. **104**, 169–181 (1993)
2. Bélair, J., Mackey, M.C.: Consumer memory and price fluctuations in commodity markets: an integrodifferential model. J. Dynam. Differ. Equ. **1**, 299–325 (1989)
3. Boullu, L., Adimy, M., Crauste, F., Pujo-Menjouet, L.: Oscillation and asymptotic convergence for a delay differential equation modeling platelet production. Discrete Contin. Dynam. Systems B **24**(6), 2417–2442 (2019)
4. Diekmann, O., van Gils, S., Verdyn Lunel, S.M., Walther, H.-O.: Delay Equations: Complex, Functional, and Nonlinear Analysis. Springer, New York (1995)
5. Erneux, T.: Applied Delay Differential Equations. Surveys and Tutorials in the Applied Mathematical Sciences, vol. 3. Springer, Berlin (2009)
6. Garab, Á.: Absence of small solutions and existence of Morse decomposition for a cyclic system of delay differential equations. J. Differ. Equ. **269**(6), 5463–5490 (2020)
7. Hadeler, K.P.: Delay equations in biology. In: Springer Lecture Notes in Mathematics, vol. 730, pp. 139–156 (1979)
8. Hadeler, K.P., Tomiuk, J.: Periodic solutions of difference differential equations. Arch. Rat. Mech. Anal. **65**, 87–95 (1977)
9. Hale, J.K., Verduyn Lunel, S.M.: Introduction to Functional Differential Equations. Applied Mathematical Sciences. Springer, Berlin (1993)
10. Ivanov, A.F.: Global asymptotic stability in a differential delay equation modeling megakaryopoiesis. Funct. Differ. Equ. **28**(3–4), 103–116 (2021)
11. Ivanov, A.F., Lani-Wayda, B: Periodic solutions for an N-dimensional cyclic feedback system with delay. J. Differ. Equ. **268**(9), 5366–5412 (2020)
12. Kuang, Y.: Delay Differential Equations with Applications in Population Dynamics. Mathematics in Science and Engineering, vol. 191. Academic Press, Cambridge (1993)
13. Mackey, M.C.: Commodity price fluctuations: price dependent delays and nonlinearities as explanatory factors. J. Econ. Theory **48**(2), 497–509 (1989)
14. Mallet-Paret, J.: Morse decomposition for delay differential equations. J. Differ. Equ. **72**(1), 270–315 (1988)
15. Smith, H.: An Introduction to Delay Differential Equations with Applications to the Life Sciences. Texts in Applied Mathematics, vol. 57. Springer, Berlin (2011)

Discrete and Continuous Models of the COVID-19 Pandemic Propagation with a Limited Time Spent in Compartments

Olzhas Turar, Simon Serovajsky, Anvar Azimov, and Maksat Mustafin

Abstract The paper considers discrete and continuous models of the epidemic propagation with a limited time spent in compartments. It contains a comparative analysis carried out for the influence of process parameters on both models. The problem of system identification is solved. Namely, we first estimated the accuracy of the solution of the inverse problem on the model data. Then the system is identified based on real data on the spread of COVID-19 in Kazakhstan, after which a forecast is made for the propagation of the epidemiological situation.

1 Introduction

COVID-19 pandemic has largely updated the development of mathematical models for epidemic propagation. Modern mathematical models of epidemiology go back to the work of R. Ross of malaria propagation research [1] and SIR model proposed by W. Kermack and A. McKendrick [2]. This model is based on the division of the entire population into three compartments of susceptible, infected and recovered. This model describes the transition of people from the compartment of susceptible to infected and then recovered. It is represented by a system of differential equations that describe the change in the size of each of these population compartments over time. A natural generalization of the SIR model is the SIRD model, in which it is assumed that some part of the infected people die, forming an additional group of deceased [3], and its simplification is the SIS model, in which the recovered

O. Turar
Astana IT University, Astana, Kazakhstan
e-mail: olzhas.turar@astanait.edu.kz

S. Serovajsky (✉) · M. Mustafin
Al-Farabi Kazakh National University, Almaty, Kazakhstan
e-mail: serovajskys@mail.ru

A. Azimov
Satbayev University, Almaty, Kazakhstan
e-mail: anvar.aa@mail.ru

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_5

patients do not develop immunity, i.e. recovered immediately join the group of susceptible [4]. The last two models generalize the SIRS model, where those who have recovered do not lose immunity immediately, but after some time, which means that the recovered group is present, but is not the end state of the population [5].

These models do not take into account latency period during which individuals have been infected but are not yet infectious themselves. This shortcoming is overcome in the SEIR model, in which a compartment of exposed is added [6]. Subsequently, other groups of the population were taken into account. Particularly, deceased patients were also considered in [7, 8], asymptomatic patients in [9], and hospitalized and critical patients in [10, 11]. In addition to ordinary differential equations, partial differential equations are also used to describe the propagation of an epidemic over a certain territory [12]. There are a significant number of stochastic models for the development of epidemics [3].

Along with continuous models, discrete models characterized by difference equations [13] are also used. In [14, 15], a discrete model with a limited time spent in exposed and patient compartments is considered. It assumes that after some time, each person from these groups will certainly move to another group: in particular, the contact will either become infected or most likely not get infected, the patient will either recover or die. In this paper, we consider a modification of this model, as well as its continuous analog. Three groups of patients are considered: undetected, isolated, and hospitalized. Undetected are not reflected in official statistics and are the main carrier of infection. Isolated are included in official statistics, but are not hospitalized and are carrier of infection. Being seriously ill hospitalized may die, but they are in strict isolation and are not carrier of infection.The flow diagram of the compartmental transitions is shown in Fig. 1.

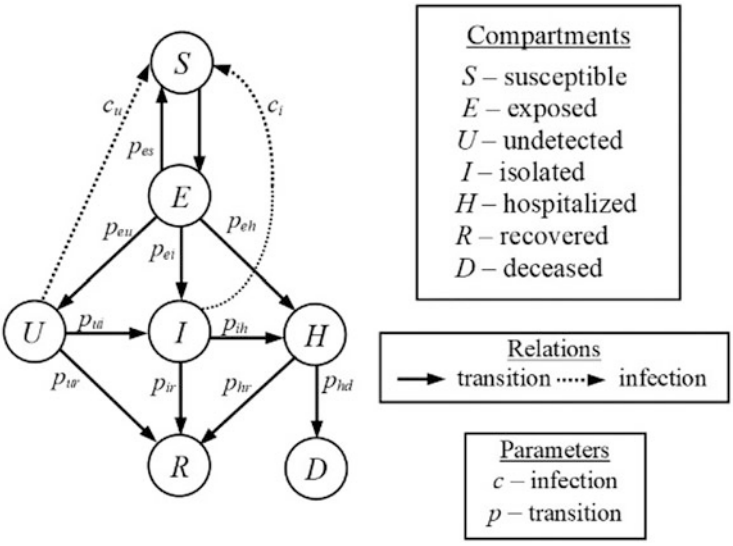


Fig. 1 Flow diagram of the compartmental transitions

A comparative analysis of the models is carried out with an assessment of the influence of the parameters included in them. We solve the problem of model identification with an evaluation of the solution accuracy. As an example, the propagation of COVID-19 in Kazakhstan is considered. The forecast results are compared with real data.

2 Discrete Model

The paper considers a discrete model of the epidemic propagation with a limited time spent in compartments. The state of the system in model described by the functions of the discrete argument S_k, E_k, U_k, I_k, H_k, R_k, D_k, which describe, respectively, the number of susceptible, exposed, undetected, isolated, hospitalized, recovered, and deceased at time k. Let's call the amounts of days spent in the exposed, undetected, isolated, and hospitalized compartments as n_e, n_u, n_i and n_h, respectively. The number of people in each compartment at a given time is the found as a sum of numbers of specific discrete function for the respective period i.e.

$$E_k = \sum_{j=1}^{n_e} e_k^j, \qquad U_k = \sum_{j=1}^{n_u} u_k^j, \qquad I_k = \sum_{j=1}^{n_i} i_k^j, \qquad H_k = \sum_{j=1}^{n_h} h_k^j, \tag{1}$$

where e_k^j, etc. denotes the number of exposed, etc. at time k and on the j-th day of being in the compartment. Each exposed, etc. of the given day of being in his compartment, a day later goes to the category of next day of being in this group, if it was not the last day of being in the compartment, what corresponds to equalities $e_{k+1}^{j+1} = e_k^j$, $j = 2, ..., n_e - 1$, etc.

The number of people in the compartments of susceptible, exposed and all groups of patients at the next time is equal to their number at the previous time, plus those who entered and minus those who left this compartment in this day:

$$S_{k+1} = S_k - (c_u U_k + c_i I_k)\frac{S_k}{N} + p_{es} e_k^{n_e}, \tag{2}$$

$$E_{k+1} = E_k + e_{k+1}^1 - e_k^{n_e}, \qquad U_{k+1} = U_k + u_{k+1}^1 - u_k^{n_u}, \tag{3}$$

$$I_{k+1} = I_k + i_{k+1}^1 - i_k^{n_i}, \qquad H_{k+1} = H_k + h_{k+1}^1 - h_k^{n_h}. \tag{4}$$

The number of people who recovered and died at a subsequent point in time is equal to their number at the previous point in time plus those who were included in this day to the specific compartment:

$$R_{k+1} = R_k + p_{ur} u_k^{n_u} + p_{ir} i_k^{n_i} + p_{hr} h_k^{n_h}, \qquad D_{k+1} = D_k + p_{hd} h_k^{n_h}. \tag{5}$$

In the given formulas, c_u and c_i denote the contagiousness of undetected and isolated patients, respectively, $p_{\alpha\beta}$ is the proportion of people in the compartment α passing into the compartment β, and N is the size of the entire population. In this case, the following equalities are natural:

$$p_{es} + p_{eu} + p_{ei} + p_{eh} = 1, \quad p_{ui} + p_{ur} = 1,$$
$$p_{ih} + p_{ir} = 1, \quad p_{hr} + p_{hd} = 1. \tag{6}$$

The number of exposed and all forms of patients on the first day of being in corresponding compartment defined by patients passed from other groups and determined by the equalities:

$$e^1_{k+1} = (c_u U_k + c_i I_k)\frac{S_k}{N}, \quad u^1_{k+1} = p_{eu} e^{n_e}_k, \quad i^1_{k+1} = p_{ei} e^{n_e}_k + p_{ui} u^{n_u}_k,$$
$$h^1_{k+1} = p_{eh} e^{n_e}_k + p_{ih} i^{n_i}_k. \tag{7}$$

The initial states of the system S_0, E_0, U_0, I_0, H_0, R_0, D_0 considered as known, namely distribution of exposed and all forms of patients at the initial moment are considered uniform in terms of days being in compartments.

The given equalities constitute a discrete model of the epidemic propagation. Passing to the limit as k tends to infinity, we establish the equilibrium position of the system. We also note that the increments in equalities (5) are positive. Thus, we establish the validity of the following statement.

Theorem 1 *The discrete system has a unique equilibrium position, specifically, the limiting values of the numbers of exposed and all forms of patients are equal to zero, and the numbers of recovered and deceased do increase with time.*

Preliminary calculations were carried out with the following values for length of periods: $n_e = 14$, $n_u = 3$, $n_i = 5$, $n_h = 7$. The contagiousness coefficients were assumed to be $c_u = 3.18$, $c_i = 0.171$, i.e. undetected patients are considered the main carrier of infection. Distribution of exposed parts moving to other compartments: $p_{eu} = 0.154$, $p_{ei} = 0.145$, $p_{eh} = 0.022$, $p_{es} = 0.679$; transition parameters for undetected patients: $p_{ui} = 0.03$, $p_{ur} = 0.97$; for isolated patients: $p_{ih} = 0.021$, $p_{hr} = 0.982$; for hospitalized: $p_{hd} = 0.018$, $p_{ir} = 0.979$. The total population was assumed to be 18,699,640 people, which corresponds to the population of the Republic of Kazakhstan in September 2020 according to the World Bank, according to datacatalog.worldbank.org. At the initial moment of time, it was believed that there were 140 contact patients, while there were no sick, recovered or deceased. Figure 2 shows the corresponding computation results.

Graphs present that at the initial stage of the process, which lasts about 200 days, there is a slight increase in morbidity and mortality, which is explained by the initially small number of infected and corresponds to the beginning of the development of the epidemic. In the next 200 or so days, there is an exponential increase in the number of sick and dead. Over the next two months, the number of

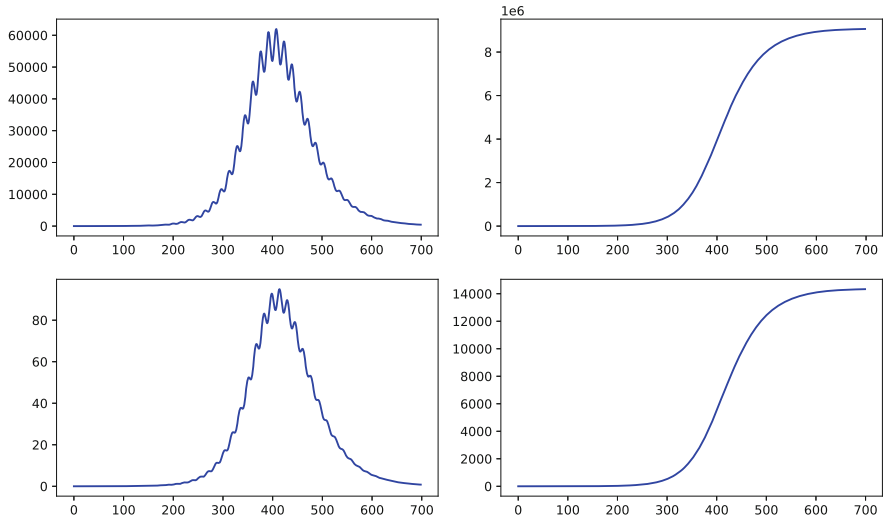


Fig. 2 The number of infected people (top) and deaths (bottom) according to the discrete model by days: the number of each day (left) and total up to the moment (right)

simultaneously ill and dying continues to grow, but the rate of growth is gradually slowing down. After reaching the maximum number of daily infected and dying, the epidemic gradually fades, which takes about the same time as the entire previous period. It is characterized by the fact that both the increase and decrease in the number of infected and deceased people every day is non-monotonic, see the left graphs in Fig. 2.

Let's note that with the considered choice of system parameters, the peak time of the epidemic falls on the 461st day from the start of the computations, when the maximum number of simultaneously sick people is observed—approximately 280 thousand people, which is 1.5% of the total population. The epidemic ends within 990 days, and the total number of recovered people is approximately 10.3 million, which is 54.85% of the total population, the number of deceased is 14,460 people, which is 0.14% of the total number of infected.

Table 1 describes the impact of the contagiousness coefficient of undetected patients c_u. With the growth of this parameter, both the maximum number of simultaneously ailing people and the total number of infected and deceased people increase, while the time of reaching the epidemic peak and the duration of the epidemic are reduced. This indicates a more intensive development of the epidemic. At the same time, the percentage of recovered and deceased from the total number of cases remains practically unchanged.

The coefficient of contagiousness of isolated patients has qualitatively the same effect on the system, but significantly lower degree of influence. The influence of all parameters characterizing the proportion of exposed patients passing into different categories is similar to the influence of contagiousness coefficients, since their

Table 1 Influence of the infectiousness coefficient of undetected patients

c_u	Epidemic peak time, day	Maximum number of infected in one day	End time of the epidemic	Total number of infected people, percentage from total amount of people	Total number of deceased people
2.829	570	184,098	1189	8,789,696, 47.05%	12,392
3.180	461	281,187	990	10,256,695, 54.85%	14,460
3.448	406	355,695	898	11,100,528, 59.36%	15,649

increase also leads to an increase in the total number of infected people. An increase in parameters characterizing the frequency of transition from milder forms of the disease to more severe ones, as well as the proportion of deaths among hospitalized, has practically no effect on the peak time of the epidemic, its duration, and also on the total number of infected cases, but leads to an increase in the number of deceased people.

Table 2 describes the effect of time in exposed compartment n_e. The growth of this parameter leads the peak time and the duration of the epidemic increase, while the maximum number of sick people at a time, as well as the total number of infected and dead, decreases. This indicates a less intensive development of the epidemic. At the same time, the percentage of recovered and deceased from the total number of infected practically does not change. The time spent in the undetected and isolated compartments has a similar effect, namely the influence of the first is the largest of three considered parameters, and influence of last is the smallest. Finally, the influence of the time spent in the hospitalized group has an even weaker effect on the system. With its increase, only the maximum number of patients at the same time increases.

3 Continuous Model

Let us describe a continuous analogue of the previously considered model. Here the same division of the entire population into compartments under the same assumptions. Thus, the state of the system is described by the functions of a continuous argument S, E, U, I, H, R, D, describing, respectively, the number of susceptible, exposed, undetected, isolated, hospitalized, recovered, and deceased at an arbitrary point in time. The process is described by the equations

$$\begin{aligned} S' &= -(c_u U + c_i I)S/N + p_{es}E/n_e, \\ E' &= (c_u U + c_i I)S/N - E/n_e, \\ U' &= p_{eu}E/n_e - U/n_u, \\ I' &= p_{ei}E/n_e + p_{ui}U/n_u - I/n_i, \\ H' &= p_{eh}E/n_e + p_{ih}I/n_i - H/n_h, \\ R' &= p_{ur}U/n_u + p_{ir}I/n_i + p_{hr}H/n_h, \\ D' &= p_{hd}H/n_h. \end{aligned} \tag{8}$$

with the corresponding initial conditions while retaining all the previously accepted notation.

Equating to zero the terms on the right side of equations (8) and solving the corresponding system of algebraic equations, we find the equilibrium position of

Table 2 Influence of time spent in a contact group

n_e, days	Epidemic peak time, day	Maximum number of infected in one day	End time of the epidemic	Total number of infected people, percentage from total amount of people	Total number of deceased people
7	294	429,488	623	10,433,949, 55.80%	14,710
14	461	281,187	990	10,256,695, 54.85%	14,460
18	540	235,494	1189	10,208,868, 54.59%	14,392

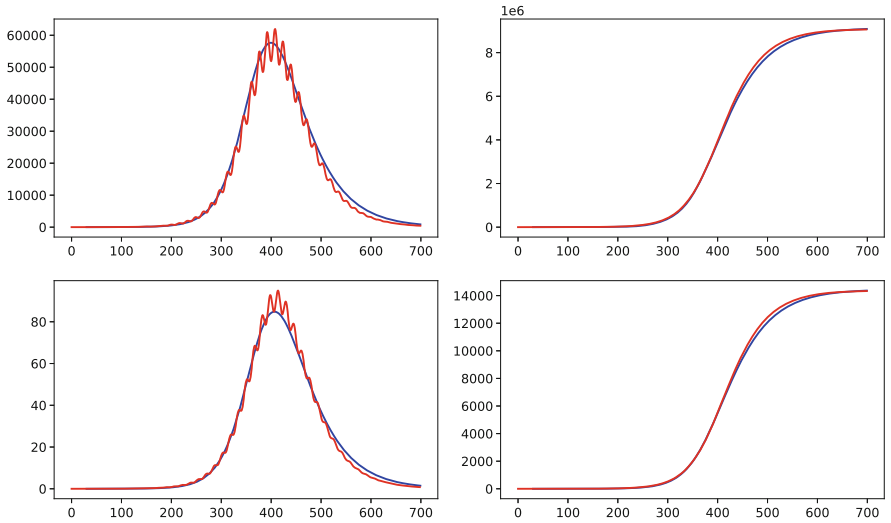


Fig. 3 The number of infected (top) and deceased (bottom) people according to continuous (blue) and discrete (red) models by days: for each day (left) and total so far (right)

the system. In addition, we note that the formulas on the right-hand sides of the last two equations (8) are positive. Thus, the following statement is true.

Theorem 2 *The continuous system has a unique equilibrium position, and the limit values for the number of exposed and all forms of patients are equal to zero, the number of recovered and deceased increases with time.*

Numerical analysis of the continuous model was carried out with the same set of parameters as for the discrete model. Figure 3 shows changes in the number of infected and deaths by day, for each day and in total up to some moment in accordance with continuous (blue) and discrete (red) models. As it shown on these graphs, the results of calculations for both models turn out to be quite close, although the changes of values per day according to the continuous model turns out to be smoother.

Table 3 shows the most important quantitative characteristics of both models for the considered set of parameters. Comparing the obtained results, one can note an extremely high degree of closeness of the results obtained based on considered models. At the same time, the duration of the epidemic according to the continuous model is longer than according to the discrete model by about a month or 3.5%, while the time to reach the peak of the epidemic in both models is almost the same. According to the continuous model, the total number of infected people is less than in the discrete model by about 4.5 thousand people, or only 0.1% of the total population, the proportion of recovered and deceased from the total number of infected people in both cases is almost the same. At the same time, the maximum number of simultaneously sick people in the continuous model is less by about

Table 3 Main quantitative characteristics of discrete and continuous models

Characteristic	Discrete model	Continuous model
Peak time of the epidemic	461 days	456 days
The maximum number of infected at the same time, its percentage from the total number of cases	281,187 people, 1.50%	260,916 people, 1.40%
End time of epidemic	990 days	1020 days
Total number of infected, its percentage from the total number of cases	10,256,695 people, 54.85%	10,252,178 people, 54.75%
Total number of recovered, its percentage from the total number of cases	10,242,235 people, 99.86%	10,237,724 people, 99.86%
Total number of deceased, its percentage from the total number of cases	14,460 people, 0.14%	14,453 people, 0.14%

20 thousand people, or 0.1% of the total amount of people. An increase in the duration of the epidemic with reduce of the total number of cases and the maximum number of simultaneously sick people with a constant proportion of deaths suggests that according to the continuous model, the intensity of the epidemic is lower than according to the discrete model.

It should be noted that each of the system parameters has a similar effect on both models, however, changing the parameters has a greater effect on the time of the peak of the epidemic and its duration for the continuous model and on the number of infected and deceased for the discrete model. For example, Table 4 lists the system characteristics for various values of period n_a in the undetected compartment: the first (upper) number in each cell corresponds to the discrete model and the second (lower) number corresponds to the discrete model. Here, when the parameter is increased by 2.5 times, the peak time of the epidemic decreases by more than 6 times for the discrete model and 7.5 times for the continuous model. At the same time, the duration of the epidemic is reduced by 3.2 times for the discrete model and by 4.5 times for the continuous model. The maximum number of sick people increases to 46. 4 times for the discrete model and 38.3 times for the continuous model. The total number of recovered and deceased people increases 4.44 times for the discrete model and 4.28 times for the continuous model.

4 Model Identification

To forecast the epidemic propagation based on mathematical models, it is necessary to properly select the parameters included in them by solving the corresponding inverse problems. As characteristics for which we practically do not have reliable

Table 4 The effect of time spent in the undetected compartment

n_e, days	Epidemic peak time, day	Maximum number of infected in one day	End time of the epidemic	Total number of infected people, percentage from total amount of people	Total number of deceased people
2	1564	17,847	2150	3,190,496, 17.06%	4492
	1801	19,081	3152	3,321,543, 17.76%	4682
3	461	281,187	990	10,256,695, 54.85%	14,460
	456	260,916	1020	10,252,178, 54.83%	14,453
5	255	827,191	674	14,147,239, 75.66%	19,945
	241	731,386	700	14,212,544, 76.00%	20,037

information, we note the initial values of E_0 and U_0 of exposed and undetected compartments, the contagiousness coefficients c_u and c_i, as well as the values p_{es}, p_{eu}, p_{ei} and p_{ar}, describing the proportion of the transition from one compartment to another. At the same time, we have relatively reliable information about the total number of cases, as well as hospitalized and deceased at certain points in time. As a result, we arrive at the following inverse problem.

Problem *It is required to choose such a vector* $q = (E_0, U_0, c_u, c_i, p_{es}, p_{eu}, p_{ei}, p_{ur})$ *so that the following conditions are fulfilled*

$$I_k[q] = \widetilde{I}_k, \qquad H_k[q] = \widetilde{H}_k, \qquad D_k[q] = \widetilde{D}_k, \qquad k = 1, \ldots, K,$$

there $I_k[q]$, $H_k[q]$, $D_k[q]$ *are the numbers of mild ill, hospitalized and deceased at time k, respectively, determined using a discrete model for a given value of the vector q, and* $\widetilde{I}_k$, $\widetilde{H}_k$, $\widetilde{D}_k$ *are the known values of the corresponding quantities, K is the number of time points at which information is measured.*

The problem is reduced to minimization of the following quantity

$$J(q) = \sum_{k=1}^{K} \left\{ \left(I_k[q] - \widetilde{I}_k \right)^2 + \left(H_k[q] - \widetilde{H}_k \right)^2 + \left(D_k[q] - \widetilde{D}_k \right)^2 \right\}.$$

Finding of the minimum is done using the trust-region method [16]. Considering that the size of the population N is quite large, the model is normalized, i.e. the number of each of the considered population compartments is preliminarily divided by N. To evaluate the effectiveness of the numerical algorithm, their values corresponding to the previous calculations are selected as the desired values of the parameters, and the corresponding solutions of the system under consideration are chosen as the "measurement results".

Table 5 shows found values of the sought parameters in comparison with their exact values, as well as the absolute and relative calculation errors. Obtained results

Table 5 The results of the calculation of the inverse problem

Parameter	Exact value	Found value	Absolute error	Relative error
S_0	0.81648	0.81640	0.00008	0.00001
E_0	0.10858	0.10864	0.00005	0.00046
U_0	0.00313	0.00316	0.00003	0.00958
c_u	3.18	3.16545	0.01455	0.00457
c_i	0.3	0.29677	0.00323	0.01075
P_{es}	0.679	0.67915	0.00015	0.00022
p_{eu}	0.154	0.15506	0.00106	0.00685
p_{ei}	0.145	0.14380	0.00120	0.00824
p_{eh}	0.022	0.02199	0.00001	0.00055
p_{ur}	0.8	0.79430	0.00570	0.00713

show that used algorithm is quite efficient. In particular, the largest error is observed in determining of the parameters U_0 and c_i, which is approximately 1%. The remaining parameters are restored with an error of fractions of a percent.

The calculations were also based on real information about the propagation of the COVID-19 epidemic in Kazakhstan, see index.minfin.com.ua. July 2020 data was used to tune the model. In this case, the following values of the identified parameters were obtained: $S_0/N = 0.068$, $E_0/N = 0.0876$, $U_0/N = 0.00248$, $c_u = 2.03$, $c_i = 0.517$, $p_{es} = 0.37$, $p_{eu} = 0.312$, $p_{ei} = 0.261$, $p_{eh} = 0.0574$, $p_{ui} = 0.99$, $p_{ur} = 6.24 \cdot 10^{-8}$ for discrete model and $S_0/N = 0.722$, $E_0/N = 0.0493$, $U_0/N = 0.00048$, $c_u = 4.95$, $c_i = 0.368$, $p_{es} = 0.19$, $p_{eu} = 0.139$, $p_{ei} = 0.594$, $p_{eh} = 0.0759$, $p_{ui} = 0.83$, $p_{ur} = 0.168$ for continuous model.

Using these values of the model coefficients, a forecast was made for the development of the epidemic for the next two months, i.e. August and September 2020. The results were compared with the actual course of the epidemic over the same period. Figure 4 shows graphs of changes in the number of infected (above) and deceased (below) by days based on discrete (left) and continuous (right) models. There, blue lines indicate real data, red lines indicate the results of calculations from the first month (July), the data for which were used to identify the model, and orange lines indicate the forecast for the next two months (August - September). As can be seen from the above graphs, the results of the forecast sufficiently reflect the course of the development of the epidemic. Forecasting for a longer period leads to a gradual decrease in the forecast accuracy.

As can be seen from the above graphs, the results of the forecast quite well reflect the course of the development of the epidemic. Although the parameter values of the models under consideration, reconstructed using real data, differ, the accuracy of the

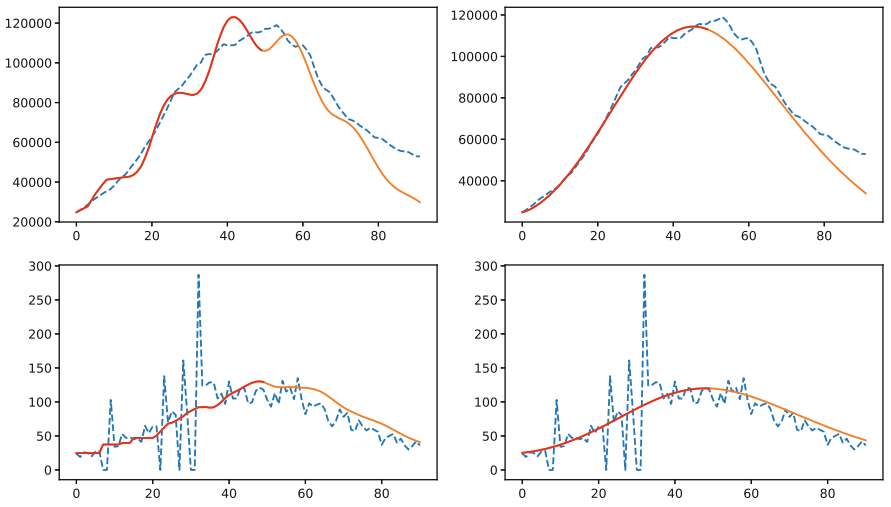


Fig. 4 Forecasted change in the number of infected (above) and deceased (below) by days compared to real data based on discrete (left) and continuous (right) models

forecast for both models is of the same order. We only note a smoother change in the functions in the continuous model compared to the discrete one. The results obtained indicate a rather high efficiency of the proposed models and the possibility of their use for forecasting epidemics.

We also note that forecasting for a longer period leads to a gradual decrease in the accuracy of the forecast. An increase in accuracy can be achieved by taking into account additional factors, in particular the effect of vaccination and the possibility of reinfection.

Acknowledgments This research was supported by the Grant No. AP09260317 *"Development of an intelligent system for assessing the development of COVID-19 epidemics and other infections in Kazakhstan"* of al-Farabi Kazakh National University.

References

1. Ross, R.: The Prevention of Malaria. John Murray, London (1911)
2. Kermack, W.O., McKendrick, A.G.: A contribution to the mathematical theory of epidemics. Proc. R. Soc. Lond. A **115**, 700–721 (1927)
3. Bailey, N.: The Mathematical Theory of Infectious Diseases and Its Applications (2nd edn.). Griffin, London (1975)
4. Bacaër, N.: Le Modèle Stochastique SIS pour une Épidémie dans un Environnement Aléatoire. J. Math. Biol. **73**, 847–866 (2016)
5. Wang, X.: An SIRS Epidemic Model with Vital Dynamics and a Ratio-Dependent Saturation Incidence Rate. Discrete Dynamics in Nature and Society (2015). https://doi.org/10.1155/2015/720682
6. Keeling, M.J., Rohani, P.: Modeling Infectious Diseases in Humans and Animals Illustrated Edition. Princeton University Press, Princeton (2007)
7. Sameni, R.: Mathematical Modeling of Epidemic Diseases; A Case Study of the COVID-19 Coronavirus. arXiv:2003.11371 (2020)
8. Krivorotko, O.I., Kabanikhin, S.I.: Matematicheskje modeli rasprostranenia COVID-19 (Mathematical Models of COVID-19 Propagation). Mathematical Center in Academcity, Novosibirsk (2021)
9. Mwalili, S., Kimathi, M., Ojiambo, V., Gathungu, D., Mbogo, R.: SEIR Model for COVID-19 Dynamics Incorporating the Environment and Social Distancing, vol. 13, p. 352 (BMC Res. Notes. 2020)
10. Unlu, E., Leger, H., Motornyi, O., et al.: Epidemic Analysis of COVID-19 Outbreak and Counter-Measures in France. medRxiv (2020). https://doi.org/10.1101/2020.04.27.20079962
11. Vynnycky, E., White, R.G. (eds.): An Introduction to Infectious Disease Modelling. Oxford University Press, Oxford (2010)
12. Huang, H., Wang, M.: The reaction-diffusion system for an SIR epidemic model with a free boundary. Discrete Contin. Dynam. Sys. B **20**(7), 2039–2050 (2015)
13. Brauer, F., Feng, Z., Castillo-Chavez, C.: Discrete epidemic models. Math. Biosci. Eng. **7**, 1–15 (2010)
14. Serovajsky, S., Turar, O.: Mathematical model of the epidemic propagation with limited time spent in exposed and infected compartments. J. Math. Mech. Comput. Sci. **4**(112), 162–169 (2021)
15. Serovajsky, S.: Mathematical Modelling. Chapman and Hall/CRC, London (2021)
16. Conn, A., Gould, N., Toint, P.: Trust-Region Methods. SIAM (2000)

Part III
Challenges in STEM Education

Some Aspects of Usage of Digital Technologies in Mathematics Education

Ján Gunčaga

Abstract Digital technologies have entered our daily lives and into schools. Computers, tablets, smartphones are a part of today's generation of children from birth; therefore they appear naturally also in education. Besides interactive whiteboards and notebooks in classrooms, children also often possess tablets and smartphones. For the teacher, it is a very actual question, how to implement advantages of these technologies in education. It appears that all of the formerly mentioned technologies have its place and they can use in effective way for achieving educational goals. Constructivist Theory of Learning has influence on mathematics education. It will be selected some topics from Slovak curricula for school mathematics. It will be discussed the aspect of visualization in mathematics education. The understanding of mathematics concepts can be deeper, motivation of pupils is greater in this case and, finally yet importantly, their creativity of students and pupils obtain strong support.

1 Introduction

Slovakia as a member country of the European Union has his educational documents formulated according the document "Recommendation of the European parliament and of the council of 18 December 2006 on key competences for lifelong learning" (see [20]). Mathematics education in lower secondary level is oriented to the development of the mathematical competence. The competences have definition in this document as a combination of knowledge, skills and attitudes appropriate to the context. Key competences are those, which all individuals need for personal fulfilment and development, active citizenship, social inclusion and employment. According [21] the mathematical competence is the ability to develop and apply mathematical thinking in order to solve a range of problems in everyday situations.

J. Gunčaga (✉)
Comenius University in Bratislava, Faculty of Education, Bratislava, Slovakia
e-mail: guncaga@fedu.uniba.sk

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_6

Building on a sound mastery of numeracy, the emphasis is on process and activity, as well as knowledge. Mathematical competence involves, to different degrees, the ability and willingness to use mathematical modes of thought (logical and spatial thinking) and presentation (formulas, models, constructs, graphs, charts). The document explains that essential knowledge, skills and attitudes related to this competence includes a sound knowledge of numbers, measures and structures, basic operations and basic mathematical presentations, an understanding of mathematical terms and concepts, and an awareness of the questions to which mathematics can offer answers. An individual should have the skills to apply basic mathematical principles and processes in everyday contexts at home and work, and to follow and assess chains of arguments. An individual should be able to reason mathematically, understand mathematical proof and communicate in mathematical language, and to use appropriate aids. A positive attitude in mathematics has its base on the respect of truth and willingness to look for reasons and to assess their validity. The State Educational Programme ISCED 2 Mathematics (see [27]) defines since 2010 mathematics education at the lower secondary level. This school subject is a part of the thematic area "Mathematics and working with information". It divides into following thematic areas:

- Numbers, variable and arithmetic operations with numbers,
- Relations, functions, tables, charts,
- Geometry and measurement,
- Combinatorics, probability, statistics,
- Logic, reasoning, proofs.

The main goal of the thematic area "Geometry and measurement" is that students obtain knowledge about base planar and space geometrical figures with inquiry-based methods and discover their properties. They learn to estimate, to measure and to calculate the size of the angle, length of some segment, surface and volume of some solid. They solve position and metrical tasks from reality. Space thinking plays one important role. In the year 2014 was innovated this program (compare with [11]) and big part of this new program is formulated in the form of standards. The beginning of this program formulates importance of the information and communication technologies (ICT) in the mathematics education. The role of the school subject mathematics is to develop the ability of students to use ICT tools for searching, elaborating, saving and presentation of information. The usage of the appropriate software should make easier heavy calculations or complicate algorithms. It brings concentration to the kern of the solved problem. Contents of the curriculum has the base on competences. Existing mathematical knowledge of the students and their experiences with the application of the existing knowledge is the base for discovery and presentation of the new mathematical notions. Education underlines the development of the students' abilities, mainly with active approach of students. Thematic area "Geometry and measurement" is oriented to the base plane and space geometrical figures such line, point, segment, triangle, quadrilateral, square, rectangle, circle, cube, rectangular parallelepiped, cylinder, cone, pyramid and sphere. Students learn their basic properties. Following

thematic area "Symmetries in the plane (axial and central)" is oriented to symmetry and congruence of the geometrical figures, central and axial symmetry, finding of axial and central symmetrical figures, construction of the picture according muster. The continuation in the sixth class is in the thematic area "The area and perimeter of the rectangle, square and rectangular in the decimal numbers, the units of area", "The angle and his measure, operations with angles" and "Triangle, congruence of the triangles". The mentioned geometrical figures are classified in these thematic areas such straight, right, acute and obtuse angle, angle bigger than straight angle; acute-angle triangle, rectangular and obtuse triangle. The work continues with the notions perimeter and area, units of perimeter and area. The continuation in the seventh class is in the thematic area "rectangular parallelepiped and cube, their surface and volume in the decimal numbers, transformation of the units of the surface and volume". It will be started in the eighth class with the enhancement of the knowledge about quadrilaterals and triangles in the frame of the thematic area "parallelogram, trapezium, perimeter and the area of the parallelogram, trapezium and triangle". The continuation id in the thematic area "Circle, circle line". It will be introduced here the notion of the Ludolph number , circle arc, central angle, sector of a circle, segment of a circle, the perimeter and the area of the circle, the length of the circle line. Geometrical activities in the eighth class are ending with the thematic area "Prism". The students obtain the knowledge about normal prisms, their networks, surfaces and volumes, the connections with cube and rectangular parallelepiped. The Pythagoras theorem in the rectangular triangle and his applications dominates in the ninth class. The geometric education after that is oriented to the pyramid, cylinder, cone, sphere and their surface and volume. The last thematic area is oriented to the triangle and similar triangle, similarity, similarity of triangles and planar geometrical figures.

2 PISA Testing and Mathematical Literacy

In the area of assessment of mathematical knowledge, a very important place belongs to the international testing of the OECD Programme for International Student Assessment (PISA). PISA gives attendance to the development of the mathematical literacy. This survey launched first time in 1997. Its goal is the evaluation of educational systems worldwide by testing the skills and knowledge of 15-year-old pupils. Since then, it has been conducted every third year. The survey focuses on several different aspects. According [16] the mathematical literacy is an individual's capacity to formulate, employ and interpret mathematics in a variety of contexts. It includes reasoning mathematically and using mathematical concepts, procedures, facts and tools to describe, explain and predict phenomena. It assists individuals to recognize the role that mathematics plays in the world and to make the well-founded judgements and decisions needed by constructive, engaged and reflective citizens. In its testing, PISA survey pays lot of attention to teaching styles. According to [18] the teacher of mathematics has a great opportunity. If he fills his allotted time

with drilling his students with routine operations, he kills their interest, hampers their intellectual development, and misuses his opportunity. But if he challenges the curiosity of his students by setting them problems proportionate to their knowledge, and helps them to solve their problems with stimulating questions, he may give them a taste for, and some means of, independent thinking. In 2012, there was a PISA measurement in Slovakia in 9-grade primary school pupils (lower secondary level). According to [8] 34 OECD countries and 31 OECD partner countries with approximately 510.000 pupils took part in the PISA 2012 measurement. In Slovakia, all 15-year old pupils born from January 1996 to December 1996 were included in the testing. It was made a stratified selection of schools and pupils forming a testing sample. Thus, 231 selected schools with 5.737 pupils were involved in the testing. The performance of Slovak pupils in mathematical literacy within the international PISA 2012 study was under the average of the involved OECD countries. The countries like Norway, Portugal, Italy, Spain, Russian Federation, United States of America, Lithuania, Sweden and Hungary had a performance comparable with the performance of Slovakia. When comparing the performance of Slovak pupils, statistically significant was the decrease of the achieved average score in the PISA 2012, as compared to all previous three-year cycles of the study. Between 2009 and 2012 it was a decrease from 497 to 482 points. There are three categories of mathematical procedures: express, use and interpret. According to [1] the worst results Slovakia achieved in the category interpret. Here the difference, as compared to the average of OECD countries, was as much as 24 points (473 points). According to [5] in the PISA study four content categories in mathematics distinguished in the year 2012:

- changes, relations and dependencies;
- quantity;
- space and shape;
- uncertainty and data.

It is important for geometry teaching the category Space and shape. This category obtain a wide range of phenomena that are encountered everywhere in our visual and physical world: patterns, properties of objects, positions and orientations, representations of objects, decoding and encoding of visual information, navigation and dynamic interaction with real shapes as well as with representations. Plane and space geometry serves as an essential foundation for space and shape, but the category extends beyond traditional geometry in content, meaning and method, drawing on elements of other mathematical areas such as spatial visualization, measurement and algebra (see also [16]). In the first category, Slovak pupils had 20 points less than the average of the OECD (474 points), in the second category 9 points less (486 points); in the third one, the result was the same as OECD average. In the fourth category, the result was the worst: only 472 points (21 points less than the average of the OECD). There was on 20th–30th April 2015 according to [17] next PISA measurement in Slovakia lower secondary schools including 15-year old pupils of the 9th grade. 6.350 pupils from 292 schools attended in this measurement. According to the initial results, the Slovak Republic achieved the performance of

475 points in the mathematical literacy. The performance of Slovak pupils was, like in 2012, statistically significantly lower than the average of OECD countries (490 points)—the difference was 15 points. Malta, Lithuania, Hungary, Israel and USA reached a performance comparable with Slovakia. A statistically significantly lower performance than Slovakia was reached by 4 OECD countries—Greece, Chile, Turkey and Mexico. The comparison of the performance of Slovak pupils in mathematics with 2012 testing showed a non-significant decrease of performance of 7 points. This means that in the PISA 2015 the Slovak pupils achieved a performance comparable to that of 2012. These results shows (see [4]), that the achievement of mathematical education has in the field geometrical education stagnate character without progress and on another fields low niveau under average of the OECD countries. It implies the need of the modernization and innovation in many parts of mathematics education.

3 The Aspect of Visualization in Geometry Teaching

The manipulation and interpretation of planar shapes and space figures in settings that call for tools ranging from dynamic geometry software and another ICT tools such using of different augmented reality applications. The aspect of visualization plays important role in the geometry teaching and this aspect is interdisciplinary. According [16] the aspect of visualization plays specific role in the scientific literacy. For example, interpreting data is such a core activity of all scientists that some rudimentary understanding of the process is essential for scientific literacy. Initially, data interpretation begins with looking for patterns, constructing simple tables and graphical visualizations, such as pie charts, bar graphs, scatterplots or Venn diagrams. Scientists make choices about how to represent the data in graphs, charts or, increasingly, in complex simulations or 3D visualizations. Nowadays exists many educational software, which make 3D visualizations of functions and space figures. The Van Hiele levels characterize the understanding of geometrical notions. Dutch mathematics teachers Pierre van Hiele and his wife Diana van Hiele-Geldof developed this theory. There are characterized according [9] and [10] following five levels:

1. Student can recognize geometric concepts, types and groups of geometric figures by their physical appearance, and in global way, without explicitly distinguishing their mathematical components or properties.
2. Student can recognize the mathematical components and properties of geometric concepts. He is able to verify conjectures through empirical reasoning and generalization. Student only formulates in this level basic logical relationships between mathematical properties of the geometrical figures.
3. Student is able to manage any logical relationship. He can to prove conjectures using informal deductive reasoning. He understand simple formal proofs, but he

is not able to construct themselves. He is able to classify geometric figures and groups of these figures and to compare them.
4. Student in this level is able to understand, why it is needed the rigorous reasoning. He can write formal deductive proofs. He understand, what does it mean axioms, hypotheses, definitions and other notions of logic.
5. Student can manage different axiomatic systems, he is able to analyze and compare properties in two axiomatic systems (for example triangle in the Euclidean geometry and spherical triangle in the spherical geometry).

The aspect of visualization and using ICT tools in mathematics education (such an augmented reality applications or educational software, for instance GeoGebra) can help students in the movement into higher Van Hiele level in the educational process (see also [2]). Vinner in [25] brings another point of view. If it will be shown the students some geometrical object, they obtain from teacher two types of information:

- Graphical: It includes pictures, drawings, physical objects, models and so on that students see in textbooks, blackboards, and with another ways. It works like a collection of photos. It is possible to give here using dynamical geometric systems and augmented reality applications.
- Verbal: It includes definitions, theorems, formulas, properties of plane and space geometric figures and so on that students read in textbooks, from screen of the computer other mobile devices, hear from teachers, other students by the collaborative lesson or other person. It works like a collection of newspaper cut-outs.

McLeod in [15] describe theory of cognitive thinking by Bruner (see [3]). His stages of understanding by the cognitive processes is useful also by mathematics education using mobile technologies in lower secondary level. Student by the understanding of geometric notions is able to be educated in following stages:

- *Enactive* stage appears first. It involves encoding action based information and storing it in our memory. Students work in this stage with different models of geometric figures in real or in the augmented reality mode. Nowadays, they can use possibilities of dynamic geometric systems.
- *Iconic* stage is typical by the fact, that information is stored visually in the form of images (a mental picture in the mind's eye). For some, this is conscious; others say they don't experience it. This may explain why, when students/pupils are learning a new subject, it is often helpful to have diagrams or illustrations to accompany the verbal information.
- *Symbolic* stage is the last. This is where information is stored in the form of a code or symbol, such as language. This is the most adaptable form of representation, for actions and images have a fixed relation to that which they represent. Cube is a symbolic representation of a single class of space figures. Symbols are flexible in that they can be manipulated, ordered, classified etc., so the user isn't constrained by actions or images. In the symbolic stage, knowledge is stored primarily as words, mathematical symbols, or in other symbol systems.

Bruner's constructivist theory suggests it is effective when faced with new material to follow a progression from enactive to iconic to symbolic representation; this holds true even for adult learners. A true instructional designer, Bruner's work also suggests that a learner even of the age in the lower secondary level is capable of learning any material so long as the instruction is organized appropriately.

4 Constructivist Theory of Learning

The constructivist theory of learning assumes that each person creates (constructs) his/her own knowledge of the world in which s/he lives. Constructivism tries to overcome the transmissiveness of traditional teaching—the transfer of "the teacher's knowledge" to the student. It deals with learning, alongside understanding (see [23] and [22]). According [14] the inductive (constructivist) approach in teaching characterizes by distinctly different characteristics from the deductive approach, while cognitive development and the learning process define as follows:

- Always based on the achieved level of learners' development,
- Provide meaningful learning,
- Enable learners to realize their own meaningful learning process,
- Influence learners so that they will modify their own knowledge schemes,
- Create and maintain a rich relationship between new knowledge and already existing knowledge schemes.

In connection with the intensive intersection of digital technologies into everyday life, teaching mathematics with ICT environment requires sufficient material and technical equipment. It needs also changes in educational approach, new communication methods in mathematics, a change in the status of the teacher of mathematics and the student, and an organizational change in mathematics lessons. A necessary condition for making changes in the teaching process is according [13] sufficient computer literacy among mathematics teachers, and their motivation and willingness to learn more in this field. The term 'constructionism' is a mnemonic for two aspects of the theory of science education. From constructivist theories of psychology, it is taken a view of learning as the reconstruction, rather than transmission, of knowledge. Then, the extension of the idea of manipulative materials to the idea that learning is most effective when part of an activity, which the learner experiences as constructing a meaningful product. The students during the lessons are active and they built new concepts, which help them to understand new notions according to their own personal needs. The opposite approach to constructionism is instructionism, which, in terms of teaching, typically involves the teacher giving instructions to children, such that they have limited opportunity for their own activity and personal way of thinking. Instructionism vs. constructivism looks like a split between two strategies for education: two ways of thinking about the transmission of knowledge. Behind all this is a split that goes beyond the acquisition of knowledge and touches on the nature of knowledge and the nature

of knowing. Constructivist instructional design, according [12], aims to provide generative mental constructions embedded in relevant learning environments, which facilitate knowledge construction by learners. The constructivist approach has many applications in different areas. It is possible to find a good example in the field of languages in [26] and in the field of science education for disabled children in [24]. Teaching mathematics provides scope for developing most of the competences defined by the International Society for Technology in Education, which are important for young people today. For example, using GeoGebra software, students can experiment, create and verify hypotheses. Within the project method, they can collaborate, communicate, collect and evaluate information from the Internet and process statistical data. Scientific thinking can be developed, for example, by a workshop method, where they not only create hypotheses using software, but learn to name problems and to argue.

5 Conclusions

Digital technology is being introduced into many school curricula, and "visualization has blossomed into a multidisciplinary research area, and a wide range of visualization tools have been developed at an accelerated pace" (compare with [19]). GeoGebra software is suitable for primary school mathematics instruction (specifically in teaching geometry to children in the fifth and sixth grade). It has great potential for use with interactive boards, and especially in the form of m-learning when students use smartphones and tablets. The option of using ready-made GeoGebra applets is very attractive for teachers. In addition, learning becomes more attractive, as teachers have the opportunity to replace transmissive teaching with the constructivist method to a great extent. Moreover, it also increases the digital literacy of students and teachers, which is a great benefit. In the future, more materials should be created and reviewed by experts on portals available for teachers, as well as classified according to topic units and students' age. Teacher training should focus on enhancing digital literacy, the ability to the work with GeoGebra and the methodology of teaching with digital technology. Presented examples offer innovative techniques in the teaching of spherical and plane geometry and promote the spatial imagination of pupils and students. Currently, similar features offer the 3D version of GeoGebra software. It will be organized in future many kinds of research, how can educational software to help by visualization and explanation of the geometrical concepts and to support by students space imagination in appropriate way. Another important goal is, how to support digital literacy by pupils and students in the educational process during mathematics and other natural sciences lessons (see [7] and [6]).

Acknowledgments Supported by grant KEGA 026UK-4/2022 entitled "The concept of Constructionism and Augmented Reality in STEM Education (CEPENSAR)".

References

1. Alföldyová, I., Polgáryová, E. et al.: Testovanie 9-2012 Priebeh, výsledky a analýzy. NÚCEM, Bratislava (2012)
2. Bayerl, E., Žilková, K.: Interactive Textbooks in mathematics education - what does it mean for students? In: Aplimat 2016 - 15th Conference on Applied Mathematics. Slovak University of Technology in Bratislava, Bratislava, pp. 56–65 (2016)
3. Bruner, J.S.: Toward a Theory of Instruction. Belkapp Press, Cambridge (1966)
4. Csachová, L., Gunčaga, J., Jurečkov á, M.: The educational research of mathematical competence. In: Focus on Mathematics Education Research, pp. 31–62. Nova Science Publishers, New York (2017)
5. Ferencová, J., Stovíčková, J., Galádová, A.: PISA 2012 Národná správa Slovensko. NÚCEM, Bratislava (2015)
6. Fuchs, K.J., Plangg, S.: Computer Algebra Systeme in der Lehrer(innen)bildung. WTM Verlag, Münster (2018)
7. Fuchsova, M., Korenova, L.: Visualisation in basic science and engineering education of future primary school teachers in human biology education using augmented reality. Eur. J. Contemp. Edu. **8**(1), 92–102 (2019)
8. Galádová, A., Lakatošová, D., Džuganová, M.: Tematická správa - PISA 2012 Matematická gramotnosť. NÚCEM, Bratislava (2015)
9. Gutierréz, A.: Geometry. In: MasterClass in Mathematics Education, pp. 151–164. Bloomsbury Academic, New York (2014)
10. Gutierréz, A., Jaime, A.: On the assesment of the van hiele levels of reasoning. Focus Learn. Problems Math. **20**(2–3), 27–46 (1998)
11. IŠVP: Inovovaný Štátny vzdelávací program Matematika nižšie stredné vzdelavanie. Štátny pedagogický ústav, Bratislava (2014). http://www.statpedu.sk/files/articles/dokumenty/inovovany-statny-vzdelavaci-program/matematika_nsv_2014.pdf. Accessed 7 March 2022
12. Karagiori, Y., Symeonu, L.: Translating constructivism into instructional design: potential and limitations. Edu. Technol. Soc. **8**(1), 17–27 (2005)
13. Korenova, L.: Digital technologies in teaching mathematics on the faculty of education of the comenius university in Bratislava. In: Aplimat 2016 - 15th Conference on Applied Mathematics, pp. 690–699. Slovak University of Technology in Bratislava, Bratislava (2016)
14. Kostrub, D.: Dieťa/žiak/študent - učivo - učite ľ, Didaktický alebo Bermudský trojuholník. Rokus, Prešov (2008)
15. McLeod, S.A.: Bruner. www.simplypsychology.org/bruner.html. Accessed 7 March 2022
16. OECD: PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving, Rev. edn. PISA, OECD Publishing, Paris (2017). http://dx.doi.org/10.1787/9789264281820-en. Accessed 7 March 2022
17. PISA 2015 Prvé výsledky výskumu 15-ročných žiakov z pohľadu Slovenska. http://www.nucem.sk/documents/27/medzinarodne_merania/pisa/publikacie_a_diseminacia/4_ine/Prve_vysledky_Slovenska_v_studii_OECD_PISA_2015.pdf. Accessed 7 March 2022
18. Polya, G.: How to Solve It: A New Aspect of Mathematical Method. Princeton University Press, Princeton, (1973)
19. Prodromou, T., Dunne, T.: Statistical literacy in data revolution era: building blocks and instructional dilemmas. Stat. Edu. Res. J. **16**(1), 38–43 (2017)
20. Recommendation: Recommendation of the European parliament and of the council of 18 December 2006 on key competences for lifelong learning. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32006H0962&from=EN. Accessed 7 March 2022 (2006)
21. Shulman, L.S.: Those who understand: knowledge growth in teaching. Edu. Res. **15**(2), 4–14 (1986)
22. Stoffová, V.: Počítač, univerzálny didaktický prostriedok. FPV UKF, Nitra (2004)
23. Turek, I.: Didaktika. Iura Edition spol s. r. o., Bratislava (2010)

24. Vančová, A., Šulovská, M.: Innovative trends in geography for pupils with mild intellectual disability. In: CBU International Conference on Innovations in Science and Education, pp. 337–344. CBU Research Institute, Prague (2016)
25. Vinner, S.: The role of definitions in the teaching and learning mathematics. Adv. Math. Think., pp. 65–81. Kluwer, Dordrecht (1991)
26. Šipošová, M.: Constructivist aspects in developing reading comprehension skills as a means of improving ESP learners' competencies. In: Aplikované jazyky v univerzitnom kontexte. TU, Zvolen, pp. 27–42 (2017)
27. ŠVP. Štátny vzdelávací program Matematika (Vzdelávacia oblasť: Matematika a práca s informáciami) Príloha ISCED 2. Bratislava: Štátny pedagogický ústav (2010). http://www.statpedu.sk/files/articles/dokumenty/statny-vzdelavaci-program/matematika_isced2.pdf. Accessed 7 March 2022

Teaching of STEM Lectures During the COVID-19 Time

Ján Gunčaga, Věra Ferdiánová, and Martin Billich

Abstract The COVID-19 pandemic situation has adversely affected mobility and international cooperation of students and workers throughout Europe. The transfer of knowledge from foreign experts who can point out the issue in another point of view is an integral part of university studies. However, the reaction of international agency CEEPUS have been greatly flexible a it allowed online and blending mobility to several countries. Thus, the aim of this article is to introduce the possibility how to implement online teaching with the use of foreign workers using available tools and means. The presentation shows practical experience within direct online teaching of STEM subjects in university courses. Primarily, the advantage of using GeoGebra software in online teaching of mathematical subjects is pointed out. Thanks to the aspect of GeoGebra visualization, there was no discomfort in transition of full—time teaching to distance teaching, because understanding of the given topic is not affected by changing the form of teaching. As a part of direct online teaching, a presentation of historical mathematical problems were created; with use of GeoGebra software it facilitated the conversion of historical tasks into a more modern form.

J. Gunčaga
Comenius University in Bratislava, Faculty of Education, Bratislava, Slovakia
e-mail: guncaga@fedu.uniba.sk

V. Ferdiánová (✉)
University of Ostrava, Faculty of Science, Department of Mathematics, Ostrava, Czech Republic
e-mail: vera.ferdianova@osu.cz

M. Billich
Catholic University in Ružomberok, Faculty of Education, Ružomberok, Slovakia
e-mail: martin.billich@ku.sk

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_7

1 Introduction: The CEEPUS Network

CEEPUS (Central European Exchange Programme for University Studies) is an exchange programme aimed at regional cooperation and mobility, especially within pre-arranged inter-university networks. The international agreement on the basis of which the cooperation is implemented is "CEEPUS III" (Agreement concerning the Central European Exchange Programme for University Studies), which entered into force on 1 May 2011, replacing the previous agreement CEEPUS II. The programme is intended for:

- undergraduate students
- postgraduate students
- academic staff

Participating countries: Albania, Bosnia and Herzegovina, Austria, Bulgaria, Croatia, Czech Republic, Hungary, Montenegro, Moldova, North Macedonia, Poland, Romania, Slovakia, Slovenia, Serbia. The universities of Pristina, Prizren and Peja in Kosovo also cooperate. International stays can take place at any eligible university abroad. Scholarship applications can be submitted either within an existing network (if the selected school is a partner) or as a CEEPUS freemover to any HEI.

At the beginning of the COVID-19 pandemic, there was a difficult situation in terms of mobility. For example, students were afraid to go abroad, as the image in the media was intimidating. Most universities did not even recommend trips abroad out of an abundance of caution. Consequently, the single European countries closed, air links between connections were cancelled and overall the times were not favourable for the implementation of any mobility with direct teaching activities.

Compared to the Erasmus project, the agency reacted very quickly and efficiently to the situation. In the Erasmus project, the Agency allowed, at most, the extension of certain types of projects such as credit mobility, etc. In contrast, student mobility was widely cancelled from the student's position, as the problem was the closed borders and in some countries the teaching at universities was only distance learning (e.g. in CR, SK, Italy, etc.):

Teacher mobility. It is subject to full-time employment and a number of teaching hours of 6 hours per week. The mobility is not intended for academic cooperation, but for cooperation within teaching experience and exchange of good practice.

- Student mobility: this is regular study mobility involving activities during the semester. The expected study period is from 3 months to 10 months. It is intended for all full-time students.
- Short term student: These are special types of practical mobilities designed for students to work on their Master's or Dissertation theses in collaboration with the host institution. The expected duration of the trip is at least 1 month.
- Blended mobility: In the case of virtual mobility, the applicant completes the professional program offered by the host institution without physical mobility, that is, without travelling to the host country. Hybrid (blended) mobility requires

a partial physical presence, so the mobility can be implemented partly in physical form and partly in virtual form [7].
- Online mobility: This is mobility that is professionally identical to teacher or student mobility, but it is implemented online.

2 Math Teaching by Using GeoGebra

Many educators say that the main goals of teaching mathematics are:

- the development of logical thinking
- the development of creative thinking
- the development of an autonomous person
- the development of the ability to solve problems

GeoGebra and educational software allows to implement these goals in mathematics education. GeoGebra is one of the most original mathematical tools that joins geometry, algebra and calculus. GeoGebra can be used for both teaching and learning mathematics from middle school through college to the university level. This open-source dynamic mathematics software help students to acquire more knowledge about geometric objects and to visualize adequate math process. GeoGebra enhances following key competencies for students:

- The skills of mathematical processing of the task.
- Ability to solve mathematical problems.
- Development of algoritmic thinking.
- Interpretation of the task results.
- Work with numerical experiments and graphical representations.

These tasks are important also during the online teaching in pandemic situation.

The basic idea of GeoGebra's environment is to provide two representations of each mathematical object in its algebra and geometry windows. If we change an object in one of these windows, its representation in the other one will be immediately updated. We can manipulate variables easily by dragging "free" objects around the plane of drawing, or by using sliders. However, GeoGebra has many ways to provide investigating geometrical proprieties, including the use of new commands with symbolic support for deriving, discovery and proving geometrical conjectures. The advantages we see in this geometry tool are:

- GeoGebra is user-friendly tool and offers easy-to-use interface, multilingual menus and commands, with minimal informatics experience required.
- GeoGebra was created to help students grasp some complicated or very abstract concepts in mathematics. It provides an opportunity to explore the world of mathematics in more details.
- GeoGebra stimulates teachers to use technology in investigations and visualization of mathematics.

- GeoGebra is good for producing and publishing complex and mathematically correct illustrations.
- GeoGebra provides an easy way to create interactive online materials. The worksheet files can be published as dynamic web pages.

2.1 Selected Examples in GeoGebra for Online Teaching for Future Math Teachers

Open questions above brings special teaching situation for teaching of future mathematics teachers. The task was, how to visualized and explain mathematics notions? There was very helpful the dynamic character of educational software such GeoGebra. The function in GeoGebra "Trace On" is important for visualization and explaining of the different kind of the sets in the plane with the given condition. It will be presented in the following examples.

Example 1 Draw the set of points, which have the same distance from the two given points A, B.

Solution It will be used here two circles k and m wit the radius r—changing parameter. It will be obtained by the using the function "Trace On" the line CD (see Fig. 1). This function is used for the points C, D, which are intersection points of the circles k and m. We obtain the line CD, which is the axis of the segment AB (Fig. 1).

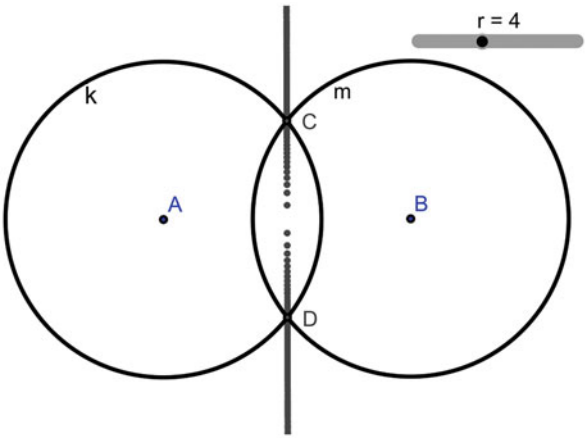


Fig. 1 The solution of the Example 1

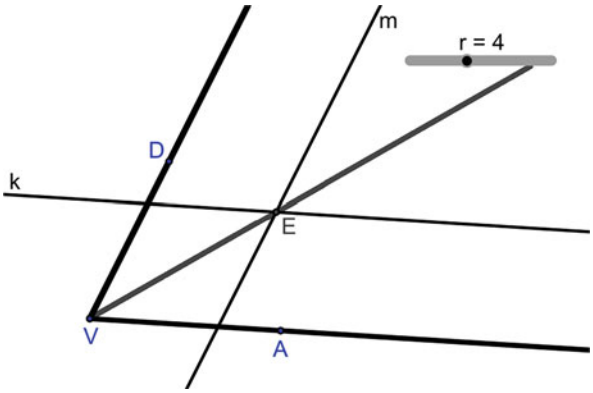


Fig. 2 The solution of the Example 2

Example 2 Draw the set of points, which have the same distance from the two given different rays with one common point V: $\overrightarrow{VA}$, $\overrightarrow{VD}$. These rays also create the angle $\angle AVD$.

Solution It will be used here two lines k and m and changing parameter r. k is parallel to $\overrightarrow{VA}$, belongs to half-plane $\overrightarrow{VAD}$ and his distance from $\overrightarrow{VA}$ is r. m is parallel to $\overrightarrow{VD}$, belongs to half-plane $\overrightarrow{VDA}$ and his distance from $\overrightarrow{VD}$ is also r. Now the intersection point of k and m is the point E. It will be used the function "Trace On" for the point E, so it will be obtained the ray $\overrightarrow{VE}$, which is the axis of the angle AVD (see Fig. 2).

Another type of school tasks suitable for future math teachers are examples from historical mathematical textbooks. These examples is possible to visualize with the help of educational software.

Example 3 There is given two different circles k and l with the common center S. On the circle with a smaller radius is given the point A. Draw another circle, which obtain the point A and touch the circles k and l (example from [5]).

Remark It is possible to draw some circle, which touch the circle with a smaller radius from inside or outside.

Solution Let's the radius of the circle k is s and the circle l is r. The line SA has two common points X and Y with the circle l with a bigger radius. Now we can draw circles m_1, m_2 with diameter XA and YA (They have radius $\frac{r+s}{2}$ and $\frac{r-s}{2}$, see Fig. 3).

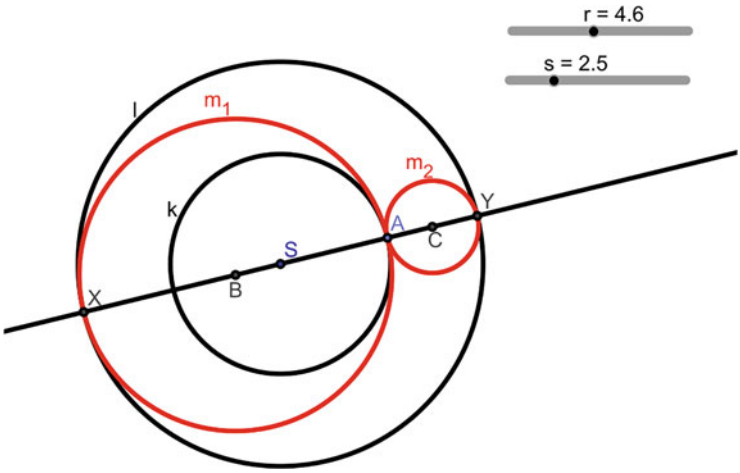


Fig. 3 The solution of the Example 3

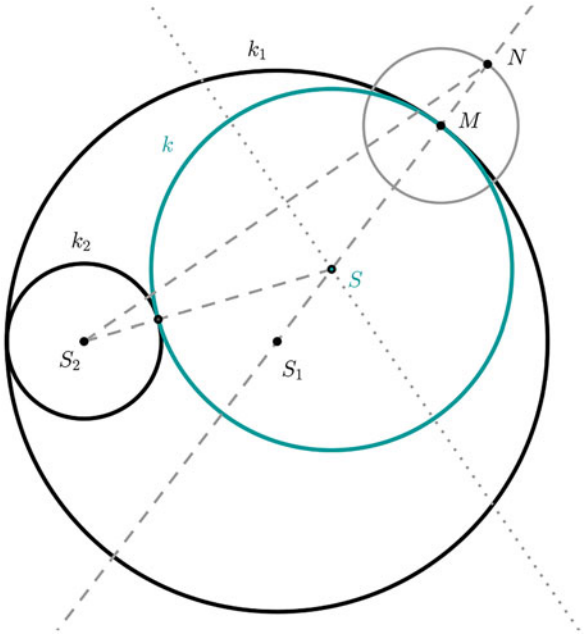
2.2 *Concept of Geometric Place with GeoGebra*

We begin with the construction of a circle tangent two non-congruent internally tangent circles, where the desired circle is tangent to the larger given circle at a given point M. Let the centres and radii of two given circles be represented by S_1, r_1 and S_2, r_2 respectively, where $r_2 < r_1$. There exists such a circle (see Fig. 4). The following procedure shows the construction of this circle, given by the centre S and radius r.

The center S must lie on line passing through center S_1 of the large circle and given point M. This center is equidistant from the circle with a smaller radius and given point M. Therefore, the centres S, S_2 and a point N forms an isosceles triangle, where point N (outside the larger circle) lies at a distance of r_2 from given point M on line passing through center S_1 and given point M, i.e. $|S_1N| = r_1 + r_2$. Now we can construct base S_2N of an isosceles triangle S_2SN, where the point S is unknown. However, if we construct the perpendicular bisector of the base S_2N, we get the third point S of our triangle, which is the center of desired circle.

In previous steps we constructed the initial model for more tasks based on concept of geometric place. First of all, we can demonstrate that the locus of all points S forms an ellipse. This statement is a conjecture, which can be supported or refuted with help of GeoGebra. One way to verify the truth of the conjecture could be to study what happens with the position of point S, when M is dragged along the larger circle are bound dragging with activated command *Trace On*. When M is dragged along the larger circle centered at point S_1, it can be observed that S seems to move along an ellipse as well. With usage of *Trace On* command, the described process is easy to see. Another way, In GeoGebra the locus of S could also be obtained by using the inbuilt *Locus* tool.

Fig. 4 Circle in the first task



In the end we found that: The locus of all points S forms an ellipse and the focal points of this ellipse are the centers S_1 and S_2 of the given circles. The major axis of the ellipse is $2a = r_1 + r_2$, where r_1 and r_2 are the radii of the given circles. An ellipse is usually defined as the set of all points in a plane for which the sum of distances from two given points (called foci) is fixed.

2.3 Extending Problem

Consider the similar problem as above. Begin with two non-congruent internally tangent circles centred on points S_1 and S_2. Construct a circle with given radius r to be tangent to both of them (tangent to the larger internally).

Solution Let the radii of two given circles be r_1 and r_2 ($r_2 < r_1$). Suppose S is the centre of the circle to be constructed. Then:

(i) S will be $|r_1 - r|$ from S_1, therefore it will be on a circle with radius $|r_1 - r|$ and centre S_1.
(ii) S will also be $r_2 + r$ from S_2, therefore it will be on a circle with radius $r_2 + r$ and centre S_2.

If we construct the circles described in (i) and (ii), their intersection S is the center of circle forming the answer to the problem. The availability of dynamic geometry software GeoGebra can be used to study main geometrical facts of existing points S.

Activity In this problem, the *slider* tool could be used to vary the value of given radius r as a parameter. We can construct some more centres and circles with required properties:

(1) For a few value of radius r construct the circles and plot their centres.
(2) Sketch the curve that contains the centres of all constructed circles.

In GeoGebra we can draw the searched locus of S again using the tool *Locus* (in a similar way as in the preceding problem). Figure 5 shows the completed task.

Equation of the Ellipse The analytical method for solving this problem requires to take certain coordinate system. Suppose the centres of the given circles are represented by S_1 and S_2 and their radii r_1 and r_2 respectively. The point of contact is assigned to be the Origin and the line joining the centres of the given circles is the x–axis. Let the centre of the drawn circle be $S(x, y)$ and its radius r. Also

$$|SS_1| + |SS_2| = (r_1 - r) + (r_2 + r) = r_1 + r_2 = \text{constant}$$

therefore, the locus of S being an ellipse. More about finding an equation of desired ellipse is elaborated in [1].

Note Depending on the relative positions of the given circles, their centres and radii, the locus of desired centres for tangent circles may be not only an ellipse but also a hyperbola.

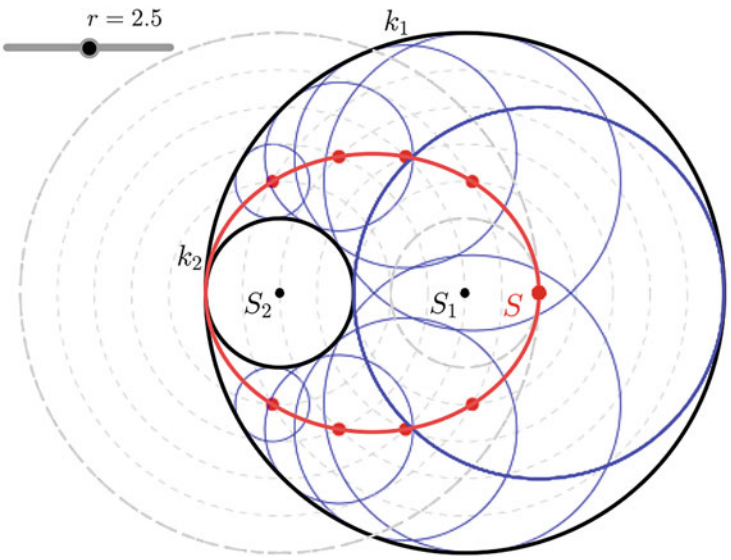


Fig. 5 Centres and circles with required properties

3 Conclusions

While looking for the previous presented examples with solutions questions, it is important to point out the reinforcement of the role of visualization by the computer.

- Visualization can often provide a simple and effective approach to discovering mathematical results, to problem solving, and to discover the concrete structure of the mathematical model by which students gain new knowledge or they learn the new mathematics notion.
- Visualization of relationships and connections in one model allows to derive new results in other mathematical areas and disciplines through new models which are isomorphic to this model (completely or only partially).
- Computer algebra systems (CAS) or dynamic geometry systems (DGS) bring the possibility to dynamically change the parameters of representation (graphs of functions, geometric shapes). It speeds up and makes for students easier to find connections between different mathematical notions and areas when they acquire new knowledge.

The usage of educational software such GeoGebra brings opportunity to solve more complex problems. Through solving this problems, the student can learn more from curricula and understand logical connections between different parts. It also supports cooperative learning. Visualization as an important supporting factor of the online teaching during the pandemic situation (see [4]). Many educational experts speak about hybrid or blended learning, the teaching and learning will be now different than before. It is expected that online teaching will have an important role in the teaching of external students in future (see [6]).

Acknowledgments Supported by grant VEGA 1/0033/22 “Discovery-oriented teaching in mathematics, science and technology education”. Many thanks also to our Network CEEPUS CIII-HU0028-15-2122 “Active Methods in Teaching and Learning Mathematics, Informatics and their Applications”.

References

1. Billich, M.: The use of geometric place in problem solving. In: Scientific Issues, Teaching Mathematics: Innovation, New Trends, Research. Ružomberok, pp. 7–14 (2008)
2. Csandova, E., Tothova, R., Korenova, L.: Uses of augmented reality in primary education. In: Prodromou, T. (Ed.) Augmented Reality in Educational Settings, pp. 80–100. Brill Sense, Leiden (2020)
3. Kenzig, M.J.: Lost in translation: adapting a face-to-face course into an online learning experience. Health Promot. Pract. **16**(5), 625–628 (2015). https://doi.org/10.1177/1524839915588295
4. Lopuchova, J.: Saturation of special educational needs for individuals with visual impairments. In: Hutyrova, M. (Ed.) Moznosti a limity vyzkumu ve specialni pedagogice, pp. 75–91. Palacky University in Olomouc, Olomouc (2013)

5. Močnik, F.: Mértani Nézlettan. (Visual Geometry). Lampel Róbert Sajátja, Pest (1856)
6. Vancova, A., Osvaldova, M.: Effects of Orff Schulwerk conception on music abilities of pupils with mental disorder. Ad Alta-J. Interdisciplinary Res. **9**(2), 261–264 (2019)
7. http://studyinhungary.hu/blog/virtual-mobility-is-now-available-in-the-ceepus-scholarship-programme

Extra-Curricular Activities to Promote STEM Learning

Natali Hritonenko, Victoria Hritonenko, and Olga Yatsenko

Abstract This chapter provides examples of extra-curricular activities proven to work well in face-to-face, hybrid, and virtual course settings. The included warm-up workouts, design-a-problem projects, decode-a-phrase sudoku, and cross-disciplinary word problems are designed to stimulate the learning of mathematical fundamentals and demonstrate both the versatility of mathematics and unity of science. They can be integrated in any STEM courses. The goal is to boost mathematical preparation, encourage math-anxious students, and entertain advanced students. Such problems transcend the boundaries of traditional mathematics and extend into other disciplines to stimulate out-of-box approaches to problem solving. All proposed activities are accompanied by detailed descriptions, examples, and students' feedback about challenges and benefits. They do not require mathematical proficiency above College Algebra and elements of Calculus, though can be easily extended to include more advanced topics.

1 Introduction

Two of the greatest scholars of all times, German mathematician Carl Friedrich Gauss (1777–1855) and Greek philosopher Aristotle (384 BC-322 BC), defined mathematics as *the queen of the sciences* and stated that *mathematical sciences particularly exhibit order and symmetry, and these are the greatest forms of the beautiful*. Indeed, mathematics is magnificent, fascinating, and exciting. However,

N. Hritonenko
Prairie View A&M University, Prairie View, TX, USA
e-mail: nahritonenko@pvamu.edu

V. Hritonenko (✉)
UC Berkeley Extension, Berkeley, CA, USA
e-mail: hritonenko@berkeley.edu

O. Yatsenko
University of Texas Health Science Center at Houston, Houston, TX, USA
e-mail: Olga.Yatsenko@uth.tmc.edu

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_8

mathematical disciplines are among the most challenging subjects for students in high school, college, and university. They are often considered both the greatest fear for students and a barrier for their academic success and career ambitions. Moreover, the Programme for International Student Assessment (PISA) ranked the US's achievements in mathematics almost at the bottom of the 35 industrialized nations and in the 38th place out of the 71 total countries surveyed [1, 2].

College Algebra and its subject-oriented satellites Contemporary Algebra and Finite Mathematics are college core course requirements. In addition, all sciences require a strong foundation of mathematics and are subjects to its rules. Multi-disciplinary instructions are becoming a major trend in modern educational curricula, and a strong mathematical background is a must. Students without a solid mathematical background may fall behind, lose interest in the topic, and fail not only their mathematics class, but also other classes. Instructors that teach Physics, Biology, and Engineering claim that about 65% of students with weak mathematical preparation either do not or barely pass their classes.

Students often give up before even trying to solve a problem that involves mathematical statements because of their fear of the subject, which just blocks their minds. In addition to a possibility of their inadequate mathematical preparation, it is probable that students do not study on a regular basis. Various surveys [3, 4] show that students should study on their own for 2–3 hours per week for every credit hour. It is unlikely that D-F-W students follow this suggestion that results in their poor preparation.

To fight the challenges in STEM education, numerous teaching techniques have been suggested to make mathematics more appealing to students. Publishers develop and constantly improve easy-to-read textbooks and software packages. Tutorial services are widely available and are even offered in some college for free. YouTube is full of video tutorials and helpful animations. Despite the plethora of these resources, more than a half of American students do not pass College Algebra on their first attempt [5, 6].

Challenges in STEM education appear in both face-to-face and online (i.e., internet, virtual) courses. With the rapid development of technology throughout the past several decades, along with other global circumstances, virtual learning is significantly changing the shape of modern education. Flexible schedules, the convenience of studying at a student's own pace, simultaneous career development, remote accessibility to learning, and financial savings on tuition, room and board are just a few advantages of online education. Thus, virtual instruction is growing at an extraordinary pace worldwide. The percentage of U.S. undergraduates taking at least one online class increased from 15.6% in 2004 to 43.1% in 2016, while the percentage of undergraduate students enrolled in fully online degree programs rose from 3.8% in 2008 to 10.8% in 2016 [7–10]. In comparison, the annual growth rate of online training is 8.5% in Germany, the leader of online education in the European Union [11, 12]. Students' satisfaction with online learning is reported to be high. To meet increasing demands of virtual education, colleges and universities have offered various online programs at lower or no cost [7–9, 12].

The COVID-19 pandemic has enormously accelerated virtual education [13]. Online learning emerged as a safe and sustainable option for schools and colleges during this challenging time. Some face-to-face classes or portions thereof need to be offered in virtual settings. New types of classes, such as internet/online synchronous and asynchronous, hybrid, and hyflex, have been immediately suggested and, in most cases, created. Students that generally prefer face-to-face learning have had to move online and handle switching to digital learning. Such students are quite different from the students who had initially chosen to pursue their education in an online setting, often facing unique challenges in both motivation and understanding. Traditional online students are at least mentally prepared for virtual education, though even they may not fully predict its challenges.

Most students who have been forced to move to an online setting by circumstances, e.g., COVID-19, are not happy with these changes. Surveys show that they fear being lonely, self-educated, and far away from their peers and professors. Such students believe they are not getting the same level of education, attention, and mentorship as during face-to-face instruction. Indeed, students must be more responsible for their own study and schedule when learning online as compared to face-to-face. Thus, online education creates new questions and calls for effective teaching strategies for making studying effective and engaging students into the learning process, potentially even pushing them toward success.

As it has been pointed out, challenges in STEM have various sources such as not sufficient students' mathematical background, prior STEM preparation that mostly concentrates on choosing a correct answer in a final test instead of understanding foundations. On the other hand, college instructors should accept all students signed up for their course and cover their course syllabus. Having inadequate preparation, students are lost in their classes. What can be done to bring students to the level needed and make a learning approach captivating and exciting to make students willing to practice even after a long workday? It is important to go beyond traditional academic training by helping students adapt to new challenges and providing a great scholastic environment for their education. A reasonable mix of graded assignments with non-traditional activities for practice and mastery encourages students to go beyond their regular "boring" homework and is beneficial to any study, especially in an online setting. Numerous recommendations, novel teaching practices, animations, tutorials, YouTube videos, and other mixed media have been developed and are ready to be integrated into the classroom to meet the demands and requirements of a course curriculum [see, e.g., [14–19].

This chapter presents several activities proven to work well in face-to-face, hybrid, and virtual settings. Their goals are to make mathematics classes interesting, entertaining, and, at the same time, educational. They aim to help students review and master mathematical fundamentals, stimulate students' interest in mathematics and motivate them to invest more time in their studies. Emphasis is made on both learning mathematical fundamentals and connecting mathematical formulas to other disciplines to demonstrate their versatility. Some of them can be integrated to mathematics classes, while others are interdisciplinary to be introduced to a variety of disciplines. The presented ideas can be modified to fit existing course

learning objectives or inspire design of new activities. All discussed activities allow reasonable flexibility. Deadlines of some activities can be set before a coming topic, while deadlines of others can be open to the last weeks of a term. Extra grade points can be offered as an incentive for their completion. Integrating activities into a course and analyzing the outcomes, an instructor can decide what activities work for a specific group of students better. Another highlight of the presented activities is memorizing basic mathematical formulas (widely used in science) while using them in multiple ways in different activities. As a result, students will be better prepared for both their future courses and SAT, ACT, GRE, MCAT, GMAT, and other standardized exams. Examples of interesting puzzles and projects that can be implemented in high school and college courses are provided along with supplemental students' responses as a reflection of their challenges and benefits. Although the mathematical level is acceptable for middle, high school, and college students with College Algebra or equivalent, these activities can be modified and extended to include any desired STEM topics.

2 Learning Through Games and Puzzles

Everybody enjoys playing games over work. Why not to add some science and incorporate them to a course curriculum to make a subject more intriguing and simultaneously foster student creativity, enhance student understanding, challenge stronger students, and help weaker students to catch up? Customized games and puzzles are powerful tools in the instructor's arsenal to combat classroom fatigue and, at the same time, repeat and review the key concepts in a relaxed and fun atmosphere [14–19]. The COVID-19 pandemic accelerated the trend towards the web-based live classes and caught many instructors struggling to adapt. The games such as Sudoku, Guess Who, and Math Bingo came to the rescue and nurtured the strong educational bond between the instructor and students and contribute to the positive educational outcomes. They have the power to stimulate non-traditional learners, encourage students in danger of falling behind, while the advanced students are entertained and have an opportunity dive deeper into the subject. The games should not require any knowledge out of a subject content, but, rather, deeper thinking and understanding. At the same time, these games can serve as a good review tool and help support understanding of the studied material.

Flexibility in the number and selection of extra-curricular activities allows easy adjustment to any students' preparation, class setting, and program requirements. Carefully crafted games and puzzles make mathematics classes entertaining and educational for students learn and review while playing. Basic mathematical rules, properties, and statements will naturally come to their mind and stay there. Small surveys given to students can guide instructors what way to go. As an engaging but nontraditional educational tool, they greatly contribute to successful STEM education when properly applied.

This section provides examples of two types of such activities, classroom Warm-ups games and puzzles for homework. Brief notes from students' surveys are shown in italic.

2.1 Warm-Ups Games

Warm-ups games consist of small questions easily solved without using a calculator. They can be played at the beginning of each class and at the middle of a long class to gain students' attention. Warm-ups games target to not only review fundamentals, master basic concepts, and prepare students for a new topic but also bring students' mind to the class. Indeed, coming from different classes students need some time to adjust their thoughts to another class.

The warm-ups format can vary and, depending on a group participation, can be played differently. The plan is to have 5–15 short questions or statements related to topics needed to be reviewed or studied in class. It is beneficial to ask additional questions or discuss students' responses, especially if the majority responses are incorrect. Examples of three types of Warm-ups games: *word problems, True/False*, and *what is greater*, are presented below.

Short Word Problems

Description Offer a few short answer problems that can be mentally solved. Some examples of such problems are below.

1. A woman bought a dress and paid $58 and a half of what it cost. How much did the dress originally cost?
2. The heaviest jackfruit grown was 76 lb. The heaviest green cabbage was just ln(sin90°) lb heavier than the heaviest jackfruit. What was the weight of heaviest green cabbage?
   By the way, both records were set up in the US, the heaviest jackfruit was grown in Hawaii in 2003, while the heaviest green cabbage was grown in Alaska in 1998.
3. It takes 2 days to paint a fence. How many days would it take to paint a twice wider and twice taller fence (working at the same pace)?
4. Divide 10 by a half and add ten. What do you get?
5. What is the lowest square number presented as the sum of squares of two other positive numbers?

It will bring more fun if at least some questions are complemented by interesting facts, like in Problem 2. It is noteworthy to mention that the last question, Problem 5, is a modification of a $15,000 question on *Who want to be a millionaire* [20]. Ideas of problems for this activity can be found in different sources [21, 22, and others].

True/False Games

Description Ask students to clap if a statement is TRUE, raise both hands up if it is FALSE, or stand up if it can be either TRUE or FALSE. Alternatively, students can be asked in advanced to prepare colored cards with FALSE or NEVER (black card), TRUE or ALWAYS (white card), POSSIBLE (red card) and to show the corresponding card. A few examples of statements related to reviewing odd/even functions are given below.

1. If $y = f(x)$ is even, then $y = f(-x)$ is odd.
2. $|f(x)|$ is even.
3. The product of two even functions is even.
4. The product of two odd functions is odd.
5. The inverse of an even function is even.
6. A graph of a function can be symmetric about the y-axis.
7. A graph of a function can be symmetric about the x-axis.

Seeing and analyzing students' responses help an instructor understand what needed to be discussed. More detailed questions after each statement, like *Why? When is it true? Why is it nonsense? Why is it impossible?*, lead to a deeper understanding of the concept.

What Is Greater?

Description Ask students to prepare three colored cards with '<', '=', '?' or four cards if '>' is added. The card '?' is to be shown when all signs are possible depending on additional information. Another option is to ask students to raise a left (right) hand if the left (right) expression is greater than the right (left) one, clap if they are equal. A few examples of questions that target trigonometric functions are below.

1. $\sin(x)$ or $\tan(x)$
2. $(\sin x+\cos x)^2$ or $\sin 2x$
3. $\sin(x)$ or $\sin^2(x)$
4. $\sec x^2$ or 0
5. $\sec^2 x^2$ or $1+\tan^2 x^2$
6. $\sec^2 x$ or 0
7. $\sin 2x$ or $2\sin x$, where x is in the first Quadrant
8. $\sin^2 3x+\cos^2 3x$ or 3

The importance of warm-ups and the following brief discussions cannot be overestimated. It may appear that students are passive during the first games, for they are not used to play in mathematics classes. However, with time, seeing the progress in their study that leads to better understanding and remembering the learning material, the students understand the value of these games and actively participate. Evaluations of warm-ups revealed that only a little more than a half of students

enjoyed them; however, when asked whether warm-ups should be continued, 100% of students say "Yes".

In their surveys, students note that the *"warm-ups games are more important than a class itself, they help remember basic formulas and their relations. They are fun and make a class alive. Warm-ups make"* the students *"to be in class on time"*.

2.2 Puzzles for Homework

Puzzles are a great asset for fighting challenges of STEM education and bringing some fresh air to a subject. They can come in different forms, like finding incorrect steps or solving a puzzle. They can be designed to be related to a certain topic or several topics.

The first example below aims to review basic algebraic formulas. The second sudoku tests knowledge on solving algebraic equations and systems.

Is 1 + 1 + 1 = 0 Correct? If not, find what step is incorrect in the proof below.

Proof

Step 1: Let $a = b$. Step 2: $a^3 = b^3$. Step 3: $a^3 - b^3 = 0$.
Step 4: $(a - b)(a^2 + ab + b^2) = 0$.
Step 5: $(a - b)(a^2 + ab + b^2)/(a - b) = 0/(a - b)$.
Step 6: $(a^2 + ab + b^2) = 0$.
Step 7: $a^2 + ab + b^2 = 0$.
Step 8: Let $a = b = 1$, then $1^2 + 1 \cdot 1 + 1^2 = 1 + 1 + 1 = 0$

Surprisingly, the most common answer shows a mistake is going from Step 3 to Step 4. The idea for such riddles can be found at different web-pages and publications, see, e.g., [23].

Sudoku Fill in the square (Fig. 1 below) with the letters A, B, E, F, G, L, N, R, U, such that each letter appears only once in each row, each column, and each small square. Find the hidden statement decoded as

1	2	3		4	5	6	7	8	9	10

and tell whether you agree with it. Each number stands for the letter from Sudoku that satisfies the following statements:

1. The row number of ***the first*** letter is a, and its column number is b such that $x = 9$ and $x = -1$ satisfy the quadratic equation $x^2 - bx - a = 0$.
2. The row number of ***the second*** letter is a, and its column number is b, such that the graphs of two lines described by the equations $y = 2x + a$ and $3y - bx = 9$ coincide.

Fig. 1 Sudoku

		F	E		N			
U	N			F			L	
	L		G					
L						N		E
		A	B		L	F		
B		U						G
					E		U	
	A			N			E	L
			L		A	G		

3. The row number of ***the third*** letter is a, such that the quadratic equation $x^2 - 2ax + 4 = 0$ has two equal positive solutions. The column number of the third letter is that solution.
4. If the row number of ***the fourth*** letter multiplied by 2 is added to its column number multiplied by 3, the result will be 22. If the column number multiplied by 2 is subtracted from the row number, the result will be 4.
5. The difference between the row and the column numbers of ***the fifth letter*** is 5, the difference between their squares is 65.
6. The column number of ***the sixth letter*** is 3 units more than its row number. If the row number is decreased by 2 and the column number is increased by 5, their new sum will be 18.
7. The row and column numbers of ***the seventh letter*** are prime numbers with the sum of 7 and positive difference.
8. The row number of ***the eight letter*** represents the side of the base, and its column number stands for the height of the box with a square base. Its surface area is 56, and the box is 4 units taller than wider.
9. The row number of the ***ninth*** letter is the width, and the column number is the length of a rectangle with the area of 40 and the perimeter of 26.
10. The row and column numbers of ***the tenth*** letter are the same. Their product is the value of the largest area of a rectangle with the perimeter of 16.

The decoded phrase shows FUN ALGEBRA. Yes, algebra is fun, indeed. Sudoku can contain different statements to be decoded, e.g., Viva Statistics [19], Great Integral, Area and Perimeter. Suguru, Kakuru, Inkies, and other number puzzles [24, 25] can be also modified to fit desirable goals. Students like such puzzles.

3 Projects in Mathematics Classes

Project-based learning is a very popular education strategy. Its value has increased significantly in the times of virtual training. Numerous research and education papers praise the benefits of integrating projects into course curriculum and describe their categories, designs, and rubrics [14, 25, 26, and others].

This section aims to discuss a special type of individual projects where students are requested to design a mathematical statement, an expression to simplify, a word problem, or any other type of mathematical problem. Depending on the project, a problem should lead to a certain answer or incorporated into a story. A student's class roll number, birthday, or a favorite number can be the answer to a problem. The story can be a fiction, tale, "My Spring Break", "My Great Nation", or any other relevant or fun theme. Mathematical concepts for the problems vary, but should remain related to the topics studied in class (or otherwise be reviewed). If applicable, students can be asked to solve their designed problem using two different methods and provide a comparative analysis of both ways. After the completion of their project, students can be asked to write their honest opinions about the project, along with its benefits and challenges. Students must be aware that they will earn maximum credit for this task even if they dislike the project as long as they justify their point of view. Similarly, zero points will be granted if they simply state, *I like the project*, without any further note. Finally, *a double-blind peer review evaluation* of projects is performed after submission of all projects. This part of project activity is not discussed in this chapter, though most students are in favor of this activity after a brief initial period of bewilderment and confusion.

Examples of two different project categories highly appreciated by students are provided below. Projects are accompanied with their description, and students' responses shown in italic. Objectives and targeted mathematical topics involved are omitted as they are quite visible.

3.1 Project "Absolute Value"

Description of the Project
Five relations $x + y = 1$, $|x + y| = 1$, $|x| + y = 1$, $x + |y| = 1$, $|x| + |y| = 1$, are given.

1. Sketch the graph of each relation. Provide mathematical reasonings for graphs.
2. What relations are functions, one-to-one functions, relations?
3. Design a real-world problem with the solution modeled by each function.

Below are just a few examples of the submissions of College Algebra students without any changes of their language.

Students' Submissions

- A gymnast is going to compete in the Olympic Games. He goes up the mountain top and then down the mountain at the same rate for each ski blade. Represent the equation of him going up and down the hill one mile per minute. What is the equation that represents his ski blades? Draw a sketch of the mountain. What is the equation related to this sketch?
- Pacman is headed one unit to the left side of the screen. While he is headed in that direction, he is going one unit down. He can only eat one piece of food at a time and there are no ghosts in his way. Trace his path. Draw a sketch of the graph of what the path like.
- You are drawing a map of the Yankee Stadium field. Second base is located at (0,1), and home plate is located at (0,−1). What is the equation of the graph knowing that when players on second and home base make a throw to either first or third base it makes a reflection across the x and y-axis. What is the equation to this shape?
- I combined all my questions into one problem just to clarify. Every man has a special way of proposing to that special someone who completes him. However, picking that special ring is what defines you.

  Before you can pick that special ring you must know which store to shop at. Each jewelry story presents its best stone by a math equation. The choices of stores you have are Jewelry$-x+y=1$, Jared$-|x+y|=1$, Szu$-1-|x|+y=1$, Super Jeweler– $x+|y|=1$, and Blue Nile$-|x|+|y|=1$. In order to find that special store you must graph the equations to figure which store is best to buy from. Which store is considered the best store a man should go to?
- A flashlight has been lit at 45° and reflected from the mirror at the same angle. Give a mathematics model of the projection.
- A machine fills Quaker Oatmeal containers with y ounces of oatmeal. After the containers are filled, another machine weighs them. If the container's weight differs from the desired y ounce weight by more than 1 ounces, the container is rejected. Write an equation that can be used to find the heaviest and lightest acceptable weights for the Quaker Oatmeal container.

Students' Responses

As a technologically advanced generation, students submit correct graphs of all relations, though it is challenging for them to provide the mathematical rationale for sketching graphs and determine why the graphs take this or that form. Thinking about the shape of the graphs *helps* them *visualize and understand the absolute value better*. Creating a word problem is *a tricky task* for most students. The most important benefit is that students are engaged in learning and *have to think outside the box*.

3.2 Project "Write a Story"

Description

1. Write a story (a story theme is assigned or related to a specific course topic).
2. Design five mathematical problems related to a certain mathematical concept and incorporate them to the story.

Examples of students' submission are provided below.

Story "Our Great Nation: Statue of Liberty" ***(A Fragment with Two Problems from the Story)***

Problem A Lady Liberty is one of the most iconic statutes in America. She greets immigrants from overseas and is visited by approximately four million people each year. This familiar attraction is a sign of freedom to many Americans. The statue, designed by Frédéric Auguste Bartholdi and dedicated on October 28, 1886, was a gift to the United States from the people of France. Suppose Lady Liberty was a student at ASU (Awesome Statues University) and the tablet she is holding is her research paper that she will be presenting in class. Given:

1. The tablet's length is 23 ft 7 in., and width is 13 ft 7 in.
2. The average paper's length is 11 in. and width is 8 in.
3. The average number of words that fit one page (single spaced, 12 pt. font, Arial) is 450.

How many words (single spaced, 12pt. Arial font) were most likely at the Lady Liberty's assignment? How many average size papers is that equivalent to?

Problem B A father and a son decided to visit the Statue of Liberty for Spring Break. The young boy was super excited about visiting and wanted to take lots of pictures to show his class when he returned to school on Monday. The two decided that for every ten steps, the father would take a picture of the son. The son would take a picture of the view every three more steps. And they would ask a stranger to take a picture of them both. The ended up with 80 pictures. How many of each type of picture did the boy have to show his class given that there are 354 steps? (Approximate your answers to the nearest whole number.)

Story "Spring Break"

For Spring Break my family and I went on vacations to Mexico. We traveled by car when we got to Refugio, Texas, I stopped at a Shell gas station to fill up. It was a surprise to me to find an old friend at that gas station. We filled our tank and left the gas station at the same time. I traveled 120 km/h heading south while my friend traveled 90 km/h heading west. At what rate was the distance between my friend and I increase in 3 hours?

In Mexico we went to Tampico. Tampico is a pleasant beach. My son wanted to take back 1152 cubic inches of sand back home with him. In order to bring this amount of sand, he needs to make a square base and open top box. He wants

to minimize the amount of material used to make the box, what should be the dimensions of his box?

My husband and I both enjoy playing volleyball. Therefore, we brought our beach ball with us. My husband started inflating the spherical ball. The volume of the ball is increasing by 3 cm?/s as the ball reached 6 cm. How fast is the radius changing at this point in time?

While playing volleyball we decided we needed to mark our playing area so that we can determine if the ball is in or out. My husband got out an 80ft rope. What would be the dimensions of the court to minimize the area?

Students' Responses

The first story was submitted by a student from a College Algebra class, while the second was a Calculus project involving topics on related rates and optimization. Even if their problems may have been a little na?ve and incorrect, the students were motivated to think and design. Students are praised for their work and gently directed if needed.

Topics can vary in tasks. For instance, a project on Laplace Transforms requires to construct the ODE and the initial value problem that have a given function is a solution. Then solve the initial value problem by two different ways and compare methods and results, and, finally, write a conclusion.

The project that involved designing a word problem received very high evaluations from students. Students say that *this project was actually fun (once they realized a story they could tell). One very challenging portion was trying to implement a problem into the story.* The students *have never been tasked to create a problem with a project usually it is the opposite, where they were tasked to solve a problem.*

From the examples in class as well as the homework, solving a problem was not an issue and they didn't think that making up a problem would be so difficult and that it is much easier *to solve a problem than to create it. In all students felt that it was very unique to say the least. It is one of the most demanding projects that they have completed intellectually. Although it was interesting, students also feel that there are a lot of vague or obscure ways to have completed the project, so it leaves a lot of room for error.* The students complain that *it is tedious at times to think of a problem and the project takes a lot of time.* Moreover, they *don't know whether the problem is good, and have to know the methods before making up a problem, though all of them recommend the project to continue.*

Design-a-problem projects spark curiosity in students, make mathematics appealing, and increase both their writing skills and mathematical culture. Indeed, the best way to engage students is to ask them to work on something with limited information given. The creativity of students' projects is incredible, even if the mathematical problems are sometimes na?ve or incorrect. Such projects are beneficial not only to students, but also to instructors, as instructors can find 'trouble' points in students' studies, get to know their students better, and gain some new knowledge they have never thought about before reading their students' stories. Reading problems created by students is enjoyable and fun.

4 Interdisciplinary Projects

As it has already been mentioned at the beginning of this chapter, mathematics is "the queen of the sciences". If mathematics is the queen, which subject is the mother of all sciences? Wikipedia, that knows everything, names mathematics as the mother of all sciences because it is a tool which solves problems of every other science. Needless to say, that it is impossible to find a discipline that does not require at least simple mathematics background. Indeed, new discoveries are made at the edges between different disciplines. Thus, it is important to emphasize the unity of all sciences. Cross-disciplinary education is an essential part of modern education. Elements from different disciplines can be incorporated to any subject.

Examples of two interdisciplinary projects are presented in this section. The first project can be offered in Calculus I or Business Calculus, the second one is suitable for College Algebra and any course on Biomathematics, for they do not require any background beyond these courses.

4.1 Project for Business, Management Sciences, Operations Research

Mankind wants to know the future. Predictions have been made since ancient times. Can you guess what year it was projected that "there is a world market for maybe five computers" and "there is no reason anyone would want a computer in their home"? Are you smiling as you read these when you have a computer, or even several, in addition to your phones, tablets, and other gadgets? Oh, yes. Probably people thought so a hundred years ago. No, both quotes were made less than a century ago, in 1943 by the Chairman of IBM, Thomas Watson, and in 1977 by the Founder of the Digital Equipment Corporation, Ken Olson. Technological development was not counted in those predictions.

Technological development and scientific innovations have changed our world. They lead to appearance of new equipment, which is more effective, less expensive, and requires less resources than the older models. Therefore, any company is continuously working toward the development of optimal modernization/ renovation strategies under improving technology. Mathematical techniques are an asset in finding a reasonable sustainable solution.

Let us consider a company that produces some goods, buys new more productive equipment, and scraps obsolete (but still functional) equipment of age T with the goal to maximize its total net profit over time [27–29]. The efficiency $b(v, t)$ at time t of the equipment installed at time v can be expressed as $b(v, t) = \exp(cv - d(t - v))$ and the cost of the new capital as $p(t) = \exp(ct)$, where $c > 0$ is the rate of technological progress. The rate $d > 0$ represents the impact of equipment age on its efficiency, and q is the initial equipment price. Using mathematical methods, it is proven in [27, 28] that under conditions $c + d > 0, c < r, q(r + c) < 1$, the

optimal service lifetime T of equipment that maximizes the discounted profit over the infinite horizon is constant and determined from the non-linear equation

$$(r+d)e^{(-(c+d)T)} - (c+d)e^{(-(r+d)T)} = (r-c)(1-(r+d)q), \tag{1}$$

where $r > 0$ is a discount rate over time.

Tasks

1. Analyze the behavior of the efficiency function $b(v,t)$ and the cost of the new capital $p(t)$ and their dependence on the parameters $c > 0$ and d. Consider positive and negative d. Provide applied interpretation and examples.
2. Find the optimal service lifetime T defined by the equation (1) under a small discount rate $r << 1$ and small technical progress and deterioration rates $c << 1, d << 1$. Interpret this result.
   Hint: Apply the first three terms of the Taylor series for $\exp(x)$.
3. Show that at a small discount rate $r << 1$, small technical progress rate $c << 1$, no deterioration ($d = 0$), and equipment price ($q = 1$), the equation (1) produces the celebrated result of Terborgh (1949) presented in [29] that the optimal equipment lifetime is $T = \sqrt{2/c}$.
   Hint: apply the Taylor series for $\exp(x)$ up to the second order.
4. Provide interpretation of your results. Describe dependence of the optimal service time T on parameters $c, d,$ and r.

4.2 Project for Bio-Medical and Pharmaceutical Sciences

The *Hill equation* or Hill-Langmuir equation

$$\log(T/(1-T)) = n\log[L] - \log K \tag{2}$$

is widely used in biochemistry and pharmacology to describe an effect of binding of one ligand to a macromolecule (such as a protein or an enzyme) on its capacity to bind additional ligand molecules [30, 31]. In the Hill equation, n is the Hill coefficient, T is the fraction of protein bound by ligand L, $[L]$ is the concentration of unbound ligand L, K is the dissociation constant between the ligand and protein. The Hill coefficient n plays an important role in describing an intricate relationship between an enzyme and sequential binding of its multiple ligand molecules. It is used by pharmaceutical companies to evaluate the ability of their drugs to bind, slow down, or inhibit an activity of a given enzyme, e.g., an enzyme that is important in cell division in order to design a drug to combat cancer or other diseases.

If $n < 1$, there is a negative cooperativity between binding the first and subsequent ligand molecules to a given enzyme. Ligand-bound protein has a decreased affinity to bind other ligands, which is useful for designing pharmaceutical drugs that are to inhibit or shut down activity of a target protein.

If $n = 1$, then the binding of the first ligand molecule to an enzyme has no effect on subsequent binding of additional ligand molecules.

If $n > 1$, then binding of the first ligand molecule to the enzyme enhances the enzyme's ability to bind subsequent ligand molecules. An example of a positive Hill coefficient is hemoglobin, which has four separate binding sites for individual oxygen molecules. Binding the first molecule of oxygen to the hemoglobin complex has a positive effect, makes it easier for that hemoglobin to bind the second molecule oxygen and even easier to bind the third and the fourth ones. The explanation of oxygen binding to hemoglobin was the original motivation behind Archibald Hill to derive his coefficient.

Tasks

1. Find the domain of each variable in the Hill equation (2).
2. Graph the Hill equation (2) for different ranges of parameter values. Interpret your results.
3. What shape does the graph of Hill equation (2) look like?
4. Is the relation described by Hill equation (2) a function? one-to-one function? Justify your response.
5. Describe the dependence of the Hill coefficient n on other parameters. Interpret your findings.
6. Present the Hill equation as a function $n = f([L], K)$ using just one logarithmic function.
7. Estimate the ranges of parameters when there is positive, negative, and no cooperative binding.
8. Find other applications of the Hill equation (2).
9. Find other functions that have a shape similar to the function described by the equation (2).

The idea of interdisciplinary projects can be found in numerous research papers, books, and just around us. A project can be adjusted to any course topic. Let us say, students-athletes take College Algebra only because it is a course requirement, and are not interested in these formulas, then think of a project that involves a soccer field, a pool table, etc. In general, cross-disciplinary projects stress applicability of mathematical statements and form interdisciplinary vision important to raising a new knowledgeable generation of scientists, practitioners, and educators.

5 Summary

This chapter provides just a few examples of different activities in hope of encouraging instructors to design their own puzzles and interesting mathematical problems that will assist them in helping students to grasp a topic and, simultaneously, enhancing their knowledge and appreciation of mathematics and other disciplines. Cartoons, inspiration stories, relaxing games, and other activities can be easily added to this list.

Most students listen to and work on stories about great mathematicians. Such stories make mathematics more appealing to them, especially if scientists are chosen properly. Why not talk about Napier and Bürgi while introducing logarithms, or about Descartes and Fermat while discussing solution of equations or rectangular coordinates? None of them were mathematicians but left an essential trace in development of mathematics we have today. Let students learn and discover. What's about the ever-so-popular cartoons and comic books? Many of them involve mathematics and careless errors. For instance, Calculus students can be asked to provide interpretation of the famous spiderman cartoon [32] and find an error there or explain the math behind the action [33]. Such tasks that can be integrated into mathematical courses making them not only educational but also entertaining. Educational activities should be prepared appropriately for different types of learners and associated with each mathematical topic. Depending on students' interests, age, and participation, the presented activities can be modified, extended, or switched entirely to other types. For instance, tricky and challenging problems are very popular among students.

Expected learning outcomes, design, advantages, disadvantages, and adaptation of each activity are to be carefully assessed. An analysis of students' participation (turning in), exam grades (academic performance), and students' surveys is a great asset to assess effectiveness of an introduced activity.

At the beginning, it takes quite some time and effort from an instructor to prepare such activities, set up a learning environment in class, and persuade students to start working on extra-curricular assignments, but it is rewarding. Keeping students involved and intensively engaged in their studies is crucial. As a benefit, students will feel that they are receiving extra attention from and solidarity with an instructor and appreciate this. Students become more enthusiastic about working on topics presented as puzzles and projects. They emphasize that *it makes them think and understand the methods and grasp a concept better, strengthens mathematical skills, allows them to design a problem they are comfortable with. Although the projects are challenging and time-consuming* and students *have to be familiar with all methods being learned, the projects make* them *think critically of the concepts such as their advantages and disadvantages* and *allow them to be a scientist.*

Acknowledgments The authors would like to thank the Chairs Dr. Jan Guncaga and Dr. Vladimir Mityushev for organizing and hosting the very productive Session *Challenges in STEM Education* at the 13th ISAAC Congress (Ghent, Belgium, August 2–6, 2021), session participants, and two anonymous reviewers for their valuable comments. Support of PVAMU FEP is acknowledged.

References

1. Programme for international student assessment PISA. OECD (2021). https://www.oecd.org/pisa/. Accessed 5 Jan 2022
2. Global e-learning market analysis 2019. Syngene Research LLP (2019). https://www.researchandmarkets.com/reports/4769385/global-e-learning-market-analysis-2019. Accessed 5 Jan 2022

3. Paff, L.: Questioning the Two-Hour Rule for Studying. Faculty Focus. Magna Publications, Madison (2017). https://www.facultyfocus.com/articles/teaching-and-learning/questioning-two-hour-rule-studying/. Accessed 28 Aug 2017
4. Rice University CTE: how much should we assign? Estimating out of class workload (2016). https://cte.rice.edu/blogarchive/2016/07/11/workload. Accessed 1 Dec 2021
5. Shakerdge, K.: High failure rates spur universities to overhaul math class. The Hechinger Report (2016), https://hechingerreport.org/high-failure-rates-spur-universities-overhaul-math-class/. Accessed 6 May 2016
6. College Algebra: Mathematical Association of America (2021) https://www.maa.org/college-algebra. Accessed 1 Dec 2021
7. 2021 online education trends report. Best Colleges. https://www.bestcolleges.com/research/. Annual-trends-in-online-education. Accessed 1 Dec 2021
8. Duffin, E.: Opinions of online college students on quality of online education US 2020. Statista (2020). https://www.statista.com/statistics/956123/opinions-online-college-students-quality-online-education/. Accessed 4 Nov 2021
9. 50 online education statistics: 2020/2021 data on higher learning and corporate training, Research.com (2020). https://research.com/education/online-education-statistics. Accessed 1 Dec 2021
10. Suzanne Smalley, S.: Half of all college students take online courses (2021). https://www.insidehighered.com/news/2021/10/13/new-us-data-show-jump-college-students-learning-online. Accessed Oct 13 2021
11. Palvia, S., Aeron, P., Gupta, P., Mahapatra, D., Parida, R., Rosner, R., Sindhi, S.: Online education: worldwide status, challenges, trends, and implications. J. Global Inf. Technol. Manag. **21**(4), 233–241 (2018)
12. Gaebel, M., Zhang, T.: Trends 2018: learning and teaching in the European higher education area. European University Association (2018). https://eua.eu/resources/publications/757:trends-2018-learning-and-teaching-in-the-european-higher-education-area.html. Accessed 11 Oct 2018
13. UNESCO: COVID-19 education: from disruption to response (2021). https://en.unesco.org/covid19/educationresponse. Accessed 1 Dec 2021
14. Hritonenko, N.: Student projects in the educational process. DELTA-K. J. Math. Council Alberta Teachers Assoc. **41**(1), 51–52 (2004)
15. Zhonggen, Y.: A meta-analysis of use of serious games in education over a decade. Int. J. Comput. Games Technol. **2019**, 4797032 (2019). https://doi.org/10.1155/2019/4797032
16. Hritonenko, N.: Logical problems in teaching statistics. DELTA-K. J. Math. Council Alberta Teachers Assoc. **40**, 73–76 (2003)
17. Hritonenko, N.: Mathematics with joy. DELTA-K. J. Math. Council Alberta Teachers Assoc. **39**(2), 9–13 (2002)
18. Vlachopoulos, D., Makri, A.: The effect of games and simulations on higher education: a systematic literature review. Int. J. Edu. Technol. High Edu. **14**, 22 (2017). https://doi.org/10.1186/s41239-017-0062-1
19. Hritonenko, N., Hritonenko, V., Yatsenko, O.: Engaging activities for enhancing mathematical learning. In: Advances in Social Science, Education and Humanities Research, pp. 98–102. Atlantis Press, Amsterdam (2021)
20. When not knowing math can cost you $15,000. YouTube (2007). https://www.youtube.com/watch?v=BbX44YSsQ2I. Accessed Dec 1, 2021
21. 100+ maths puzzles with answers! Mentalup. https://www.mentalup.co/blog/brain-teasers-2. Accessed 1 Dec 2021
22. Hritonenko, N., Yatsenko, Y.: USA Through the Lens of Mathematics. Chapman and Hall/CRC Press, New York (2021)
23. Campbell, E., Hritonenko, N.: Integrated approach in teaching trigonometry concepts. Teach. J. OOI Acad. **5**(1), 121–129 (2005)
24. Printable math and number puzzles suited for all kids, parents, teachers and math learners (2021). https://www.mathinenglish.com/puzzlemain.php. Accessed 1 Dec 2021

25. Printable puzzles, mazes and more! Krazydad (2021). https://krazydad.com/. Accessed 1 Dec 2021
26. Hritonenko, N., Yatsenko, O.: Projects to facilitate mathematical learning. In: Teaching Mathematics in Higher Education and Working with Gifted Students In Contemporary Context, Belarus, pp. 8–11 (2019)
27. Hritonenko, N., Yatsenko, Y.: Mathematical Modeling in Economics, Ecology and the Environment, 2nd edn. Springer, New York (2013)
28. Yatsenko, Y., Hritonenko, N.: Economic life replacement under improving technology. Int. J. Product. Econ. **133**, 596–602 (2011)
29. Terborgh, G.: Dynamic Equipment Policy. McGraw-Hill, New York (1949)
30. Abeliovich, H.: On Hill coefficients and subunit interaction energies. J. Math. Biol. **73**, 1399–1411 (2016)
31. Bellelli, A., Caglioti, E.: On the measurement of cooperativity and the physic-chemical meaning of the Hill coefficient. Curr. Protein Pept. Sci. **20**, 861–872 (2019)
32. Fun with integrals. Life Through A Mathematician's Eyes, Wordpress.com (2015). https://lifethroughamathematicianseyes.wordpress.com/2015/10/21/fun-with-integrals/. Accessed 1 Dec 2021
33. Clark, N.: Cabin Calculus. Blog (2013). http://anengineersaspect.blogspot.com/2013/04/cabin-calculus-cartoon-thursday.html. Accessed 1 Dec 2021

Usage of Online Platforms in Education of Mathematics in Transcarpathia at the Beginning of Quarantine

Gabriella Papp

Abstract Distance learning and e-learning as concepts have been in our minds for a long time. In March 2020, they suddenly gained great importance due to the introduction of quarantine and were immediately put into practice. It had to be applied in the everyday lives of teachers and students with surprising speed.

The goal of this research is to assess and demonstrate how teachers overcome the difficulties of mathematics education in distance learning. For this purpose, a month later after the beginning of distance education, I conducted a questionnaire survey among 20 teachers of mathematics in Transcarpathia who teach in several educational institutions with different work experiences. They were asked how education went on during quarantine, how they chose the platforms and methods needed to hold their lessons, what the checking and testing process was, what advantages and disadvantages they faced in distance learning.

1 Introduction

Due to the quarantine introduced during the pandemic, teachers had to face a new problem. The concept and practice of e-learning and distance learning had to be incorporated into everyday life, which were far removed from the methodology learned or their lessons. In this regard, teachers had to find solutions to questions such as, *"Which platform should be used?"*, *"How can they best to solve that changes in the teaching-learning process do not reduce students' knowledge?"*. What the teachers did in the educational process with a board, booklet, or interactive aids, sometimes playfully, yet accurately, can now only be done remotely using a video connection or written instructions. Under the renewed conditions, students will have a greater role in independently processing the curriculum, possibly searching the Internet.

G. Papp (✉)
University of Debrecen, Debrecen, Hungary

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_9

According to Frederick et al. (see [4]) from more complete definition of learning can be crafted a new one: *Learning is improved capabilities in knowledge and/or behavior as a result of mediated experiences that are constrained by interactions with the situation*. With this definition of LEARNING we are half-way to our goal of defining distance learning. Now consider that there is more than one purpose for learning. Recognizing that learning is a constant process that takes place wherever and whenever the individual is receptive, there must be accommodation made for the different purposes for learning (different learning intentions). After all, learning situations may be formal (contrived) or be self-directed in everyday settings (naturalistic). Learning may occur by design, or it might occur by chance. Therefore, with these possibilities in mind, the authors propose three major subcategories of learning: (1) instruction: objectives-driven learning; (2) exploration: without objectives; and (3) serendipity: unintended learning [4].

Digital technologies have made their way not only into our everyday lives, but nowadays they are also commonly used in schools. Computers, tablets and smartphones are now part of the lives of this new generation of students [6]. All subjects are important, and it is difficult to teach all of them that you suddenly have to apply this method, yet perhaps one of the most difficult situations is for mathematics teachers. Most of the time we spend our days writing on a board, taking description the proof, solving practical examples, which now has to be solved in a completely different environment, with the help of other tools. To overcome difficulties, many platforms can be used to create groups, solve tests and tasks.

Teachers must understand how technology, pedagogy, and content interrelate, and create a form of knowledge that goes beyond the three separate knowledge bases. Teaching with technology requires a flexible framework that explains how rapidly-changing, protean technologies may be effectively integrated with a range of pedagogical approaches and content areas [6].

1.1 Distance Learning

Distance education emerged as an alternative to traditional education in the 18th century as a differently conceivable and feasible form of education, teaching, and learning. In the beginning, the main tool was the letter in which the written materials were delivered to the students. Later, also using traditional mail, image, sound and video recordings were also transmitted [3].

We can read this about distance education in the 1987 Adult Education Small Lexicon, formulated by Gyula Csoma: Distance learning is a special way of remote control; a remote control-based management and learning system, which is organized for the acquisition of defined, prescribed and precisely structured knowledge, thinking and, to a limited extent, action operations in the context of work-based learning, in order to meet specific requirements. In the didactic system of distance education, the two stages of the teaching-learning process are as far apart as possible in space and time [7].

Bušelić in [1] puts distance learning is a field of education that focuses on teaching methods and technology with the aim of delivering teaching, often on an individual basis, to students who are not physically present in a traditional educational setting such as a classroom. It has been described as a process to create and provide access to learning when the source of information and the learners are separated by time and distance, or both [1]. The United States Distance Learning Association defined distance learning in 1998 as "the acquisition of knowledge and skills through mediated information and instruction, encompassing all technologies and other forms of learning at a distance." This is a definition that does not distinguish formal and informal learning, or different types of distance (temporal and physical) [4].

Distance learning offers a myriad of advantages which can be evaluated by technical, social and economic criteria. Also, distance learning methods have their own pedagogical merit, leading to different ways of conceiving knowledge generation and acquisition [1]. By Frederick et al. definition of distance learning is this: *distance learning is improved capabilities in knowledge and/or behaviors as a result of mediated experiences that are constrained by time and/or distance such that the learner does not share the same situation with what is being learned* [4].

1.2 E-Learning

Learning has a procedural and active character, which must lead to construction of knowledge by the learner on the background of the learners individual experience and knowledge [9]. New technologies are driving necessary and inevitable change throughout the educational landscape. Effective technology use, however, is difficult, because technology introduces a new set of variables to the already complicated task of lesson planning and teaching [6].

The concept of e-learning is used in several senses. In the broadest sense, technology-supported learning, computer-assisted learning, digital learning [7]. The e-Learning system must enable the learner to create the personal information landscape while working with the provided learning materials. The means are individual compilation and topical rearrangement of learning material, creating "pools" of especially important documents as well as the possibility to annotate and cross-reference material [9]. Most of the terms have in common the ability to use a computer connected to a network, that offers the possibility to learn from anywhere, anytime, in any rhythm, with any means [2]. Students have the opportunity to proceed on their own schedule independently of the teachers. This is called asynchronous learning. This method does not preclude communication between students and teachers, as choosing an asynchronous form of communication can answer all the questions [3].

Tavangarian in [9] summarizes this as follows: We will call e-Learning all forms of electronic supported learning and teaching, which are procedural in character and aim to effect the construction of knowledge with reference to individual experience,

practice and knowledge of the learner. Information and communication systems, whether networked or not, serve as specific media (specific in the sense elaborated previously) to implement the learning process [9].

According to [2] communication is the key when it gets difficult to try reaching out to students via texts, various messaging apps, video calls, and so on-content should be such that enable students for practice and also hone their skills. The quality of the courses should be improved continuously and teachers must try to give their best [2].

1.3 Digital Technologies

During mathematics classes, pupils can make use of digital technologies in various way:

- during numerical calculations so they can concentrate on the solution of the problem itself;
- for visualisation, modelling and simulation of problems and thus to obtain such a graphical representation of the problem, which pushes them towards a solution;
- as a source of educational materials e.g. e-books or videos, interactive educational materials;
- drilling exercises, a pupil can make use of electronic working sheets or e-tests to evaluate himself [6].

Digital technologies offer teachers a possibility to make use of new educational methods, e.g. the constructivist approach, controlled search, workshop method or peer instruction method. Digital technologies are very suitable for project teaching, too. Teachers can make use of blended learning, flipped classroom method, etc. Last but not least, the computers are used for electronic testing when knowledge of the pupils is measured [6].

Dhawan says that online programs should be designed in such a way that they are creative, interactive, relevant, student-centered, and group-based. Instructors indulged them in remote teaching few flatforms such as Google Hangouts, Skype, Adobe Connect, Microsoft teams, and few more, though ZOOM emerged as a clear winner. Also, to conduct smooth teaching-learning programs, a list of online etiquettes was shared with students and proper instructions for attending classes were given to them [2]. In my opinion, platforms for editing e-tests also play a significant role in distance learning.

The classical test consists of a set of test assignments and questions from concrete subject domain, related to an assessment system and offered for solving (accomplishment of certain activities) [8].

Sokolova and Totkov explain the e-tests theory: The classical taxonomy of test questions and assignments is based on the way by which examinees give their answers. Test questions and assignments are divided into two groups: free-form responds (open type)—the examinees construct their answers themselves;

questions and assignments with constructed answer (closed type)—examinees select the correct answer from a set of alternative answers [8].

According Korenova Therefore we can define the term "e-test" dually: 1. In a narrower meaning, the e-test is an electronically controlled didactic test with an option to enrich it with multimedia elements. 2. In a wider meaning, the e-test is an electronic interactive material based on a system of questions and searching for answers created not only for measuring, but also for reaching educational goals (hence can serve as tools for innovative teaching methods). Using e-test we are able not just to determine the students' knowledge, but with these new digital tools we can increase the students' motivation, use them during repetition, exercise, controlled discovery methods. The e-test is very attractive from the students' point of view, because the digital world is very close to them [5].

Test questions and assignments, which are included in a concrete e-test can be chosen on the basis of different principles and rules. Opinions of different authors expressed in the literature, are very contradictory [8].

The question arises, what kind of digital technology do the mathematics teachers of the surrounding Hungarian-language schools use? Is it one of the above-mentioned platforms or e-tests to assess knowledge even at the beginning of distance learning? The results of this research I presented below.

2 Methods

The target group of the research were mathematics teachers teaching in Transcarpathia, in Hungarian-language primary and secondary schools, as well as in higher education. The 8-item electronic questionnaire I edited using a Google Form and then made available on the social network. I got answers to my questions that what methods are used after the introduction of distance learning, what platform they do it on, and what advantages and disadvantages they see after overcoming the initial difficulties. In addition to selecting one and multiple choice items (close type), participants had to enter their own answers to the advantages and disadvantages questions (open type). Fewer than expected, only 20 responses were received, the results of which I will presented below as pilot research.

3 Results

While editing the survey questions, I considered important the question *"How many years of mathematical pedagogical experience?"*. I was curious about the differences between the different work experiences in choosing and applying the technology and platforms required by the new situation. Figure 1 shows the distribution of respondents' work experience.

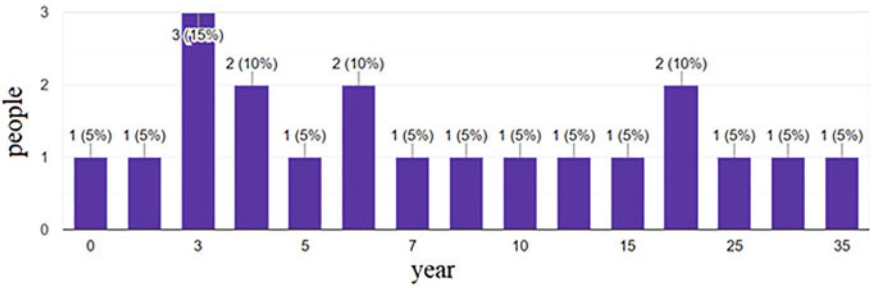


Fig. 1 Distribution of respondents' work experience

We can see that among the respondents, 8 persons have 1–5 years and another 5 persons have 6–10 years work experience. This suggests that more than half of the respondents are closer to applying the technologies due to their young age. This conclusion does not rule out the possibility that the use of technologies or learning to use the necessary new programs and platforms would be far from more experienced colleagues.

When choosing educational interfaces, in the spring of 2020, 60% of responding teachers marked Facebook Messenger, and a further 10% mentioned this application choosing the "other" answer option, thus listing more platforms. Regardless of work experience, they responded that the applied application was chosen because of its prevalence, as their students, or in the case of younger age, the parents of the students, had already used it in their daily lives. 10% used Google applications, and another 5–5% used other platforms like Geogebra groups, Microsoft Teams, EduBase and Smart Learning Suite.

55% of the responding instructors had already used video call to conduct the lesson at the time of the survey. Only 10% of respondents use the Zoom platform alone, and another 10% use other video applications in addition to Zoom. 15% also used it for measuring the level of knowledge when solving both oral and written tasks. An additional 25% created and applied e-tests using different test editing platforms, and 35% used traditional tests during the learning assessment. Other 10% was traditional and e-tests, 10% that students submit photos from their solved tasks and 5% does not use knowledge level measurement.

In Table 1, I summarized the advantages and disadvantages that respondents wrote while completing the questionnaire. During the review, I categorized and generalized the responses where the same things appeared. There were respondents who expressed several advantages and / or several disadvantages.

The table shows that the instructors in the survey see the advantage in educating students for independent, in which it helps a lot that they are separated from teachers in space and time, and that the instructional videos they can be viewed multiple times. In contrast, the disadvantages are that it is more difficult for students to learn this way, often either due to a lack of necessary tools or difficulties in mathematics. And in the case of knowledge assessment tests, it is difficult to decide whether the

Table 1 Advantages and disadvantages of distance learning according to the respondents

Advantages	Disadvantages
Students independent (4)	Harder to learn (3)
Raising awareness, interest (2)	Lack of device (electricity, internet, communication equipment) (4)
Repeatable/look back video lesson and curriculum (3)	It is difficult to accountability the students knowledges (a parent or child had answer?) (3)
Speed (3)	Lack of time (more time to prepare teacher) (3)
Different space and time (4)	No personal contact (4)
No advantage (2)	No disadvantages (1)

student solved the set task alone or with help. The lack of personal contact was also mentioned as a disadvantage.

Respondents included teachers who said there were no advantages or disadvantages to distance learning. In these cases, it can also be assumed that they did not want to answer the question.

4 Conclusion

In the spring of 2020, asynchronous learning introduced the application of new technologies in our countryside as well. The 20 mathematics teachers I interviewed jumped through this hurdle. They use multiple platforms, online materials and video calling to do their job accurately and conscientiously. To measure their students' level of knowledge, they perform classroom character lessons in a video call, or take online measurements using e-tests or correcting images submitted by students.

Examining the presented results, the possibility arises that it would be expedient to repeat the survey by looking for the same 20 respondents to compare the extent to which the technique they used has changed in the last school year of distance education. Due to anonymous responses, this would not be easy to do, but instead it would be appropriate to extend it to more respondents to I get more accurate values and a picture of the distance learning process.

References

1. Bušelić, M.: Distance learning – concepts and contributions. Oecon. Jadertina **2**(1), 23–34 (2012). https://doi.org/10.15291/oec.209
2. Dhawan, S.: Online learning: a panacea in the time of COVID-19 crisis. J. Edu. Technol. Syst. **49**(1), 5–22 (2020). https://doi.org/10.1177/0047239520934018
3. Fazekas, G., Kocsis, G., Balla, T.: Elektronikus oktatási környezetek (2014). https://regi.tankonyvtar.hu/hu/tartalom/tamop412A/2011-0103_10_elektronikus_oktatasi_kornyezetek/ch01.html. Accessed 18 July 2021

4. King, F.B., Young, M.F., Drivere-Richmond, K., Schrader, P.G.: Defining distance learning and distance education (2001). https://www.researchgate.net/publication/228716418_Defining_distance_learning_and_distance_education. Accessed 19 July 2021
5. Korenova, L.: Usage possibilities of e-tests in a digital mathematical environment. Usta ad Albim BOHEMICA č. **3**, 78–83 (2013). ISSN 1802-825X
6. Korenova, L.: What to use for mathematics in high school: PC, tablet or graphing calculator? Int. J. Technol. Math. Edu. **22**(2), 59–64 (2015). https://doi.org/10.1564/tme_v22.2.03
7. Nemes, G., Csilléry, M.: Kutatás az atipikus tanulási formák (távoktatás/e-learning) modelljeinek kifejlesztésére célcsoportonként, a modellek bevezetésére és alkalmazására. Nemzeti Felnőttképzési Intézet, Budapest (2006)
8. Sokolova, M., Totkov, G.: About test classification in e-learning environment. In: International Conference on Computer Systems and Technologies (2005). <https://www.researchgate.net/publication/251757911.21. Accessed June 2021
9. Tavangarian, D., Leypold, M.E., Nölting, K., Röser, M., Voigt, D.: Is e-learning the solution for individual learning? Electron. J. of e-Learn. **2**(2), 273–280 (2004). ISSN 1479-4403

The Use of Technologies to Promote Critical Thinking in Pre-service Teachers

Vanda Santos

Abstract In educational environments, technology is present as a resource that facilitates teaching and learning. In higher education, modern education in science, technology, engineering and mathematics (STEM) faces fundamental challenges. The objective of the present study is to analyse, the learning strategy with the use of technologies in mathematical activities, analyse mathematical activities and ways of thinking in Higher Education. The research methodology adopted consists of a case study, relating to a group of pre-service teachers in a public Higher Education Institution. A qualitative approach was adopted with the interpretation of data collected through the activities on GeoGebra Classroom, brief questionnaire and conducting individual interviews. It is concluded that technologies have a significant participation in the educational environment and support teaching and learning. The teacher, when perfecting his pedagogical practice, will be able to insert the technological tools in teaching and learning, to improve the interaction with students and further the improvement of learning with the modelled use of technology in the classroom.

1 Introduction

In educational environments, technology is present as a resource that facilitates teaching and learning. In the teaching and learning of mathematics, at all levels of education, it needs to integrate not only technology, but also the establishment of links with other areas of knowledge, namely with the sciences in general.

The purpose of education is not just to teach basic knowledge, but to use thinking skills such as creative thinking skills, problem solving skills, science and technology skills, as these are necessary skills for sustainability and lifelong education. According to Organisation for Economic Co-operation and Development, the concept

V. Santos (✉)
University of Aveiro, Campus Universitário de Santiago, Aveiro, Portugal
e-mail: vandasantos@ua.pt

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_10

of competency implies more than just the acquisition of knowledge and skills [11]. There are important skills, such as: learning to do, which includes problem-solving skills, critical thinking and collaboration; learning to be, which includes social and cross-cultural skills, personal responsibility and self-regulation; and learning to live together, which includes teamwork, civic and digital citizenship, and global competence [1, 11, 13, 16, 17]. Interdisciplinary knowledge is increasingly important for understanding and solving complex problems [11].

In the teaching and learning of mathematics, at all levels of education, it needs to integrate not only technology, but also the establishment of links with other areas of knowledge [7], namely with the sciences in general.

The advantage of using technology in the teaching of mathematics and its effects on professional development, namely in basic and secondary education, it is well studied [22]. According to the authors Jones [4] and Tomaschko et al. [20] are unanimous in stating that the teaching of mathematics, activities supported by technology, facilitate the development of positive attitudes that will lead to better learning and a greater taste for this science. Technologies enable students to work at higher levels of generalization or abstraction and the GeoGebra software (for all levels of education that combines together geometry, algebra, spreadsheets, graphing, statistics and calculus in a single application[1]) given the multiple geometric and symbolic representations that it offers of mathematical concepts, associating visualization and interactivity [14], can be a potentiator of solid content learning mathematicians and supporting Science, Technology, Engineering, (Arts) and Mathematics (STE(A)M) education innovating in teaching and learning. The use of GeoGebra Classroom, a virtual platform, can made possible preparing STEM practices for the teacher. They can assign tasks for students, observe in real time the development of the activity carried out by the students, it is possible to view which tasks students have (or have not) started, providing an immediate feedback and a better interaction between teacher and students [24].

In higher education, modern education in STEM face fundamental challenges [12]. Interdisciplinarity, is an approach with recognized potential to provide relevant experiences to students, bringing them closer to reality situations, and allowing them to establish connections between curricular topics to develop deeper learning in these areas, but also skills such as communication, problem-solving and critical thinking [9, 18]. The last skill, critical thinking, according to National Research Council is a central element of problem-solving at all levels of STEM education [8].

In mathematic teaching, according to Su et al. [19], students have the ability to improve and develop their critical thinking when learning mathematics by solving mathematical problems, identifying possible solutions and evaluating and justifying their reasons for doing mathematics, the results and thus gaining confidence in the way they think. [19] also mention that critical thinking, combined with mathematical reasoning, allows students to reflect on their own reasoning, as they must be taught

[1] https://www.geogebra.org/about.

to: identify scenarios; evaluate them; select problem-solving strategies; identify possible conclusions that have to be logical; describe and summarize a solution; and sometimes indicate how these solutions will apply to more advanced mathematical problems.

The objective of the present study is to analyse, the learning strategy with the use of technologies in mathematical activity and ways of thinking in Higher Education.

Overview of the Paper The paper is organised as follows: after the introduction, the methodology is presented in Sect. 2. In Sect. 3 the context of the study, participants and activities with the use of GeoGebra Classroom are described as well the interviews. In Sect. 4 discussion is made. In Sect. 5 conclusions are drawn.

2 Methodology

The research methodology adopted consists of a case study, relating to a group of pre-service teachers in a public Higher Education Institution. The case study is the most common qualitative method and is implemented when the researcher is interested in researching a singular, particular situation, defined in the development of the study. In fact, [23] states that a "case study is an empirical investigation that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident (p. 13)".

To cross-reference data from multiple sources of evidence, for example an interview allows triangulation of data, adding more rigor to the investigations [2, 10, 23]. Triangulation is a procedure where convergences between multiple and different sources of information are sought to form themes or categories in a study [3] .

3 Study Description

3.1 Context of the Study

The activities with the pre-service teachers took place during the academic year 2020/2021 but, due to the global pandemic by Covid-19, the face to face classess were interrupted in the middle of March of 2020 and a distance learning regime was adopted.

3.2 Participants

The participants (N=10) were pre-service students (Master Students for Training of Teachers for the 1st to 6th grades with emphasis on Mathematics and Natural Sciencese) at 2nd semester of 2021, between April and May of 2021 .

During the 1st semester these students worked, in the context of a project Formative "Train future teachers to teach children through Challenge Based Learning (CBL)", in articulation with other courses. In this experience, students were challenged to develop CBL projects [5] in accordance with a common motto for all courses. This motto was centered on the goal 11 of the United Nations Sustainable Development Goals (SDG–11), about Sustainable Cities and Communities [21] using technological resources and technologies to support active learning. This project allowed students to work with different areas in each topic, in a STEAM strand.

3.3 Activities Description and Results

The activities took place at the 2nd semester of 2021. A brief questionnaire were given befor the GeoGebra Classroom was introduced. A total of four activities in GeoGebra Classroom were given—these activities are in line with the level of education that will teach (students between 6 and 12 years old), were aligned with the national curricula. At the end an interviews were done by email about critical thinking.

Students have worked previously with GeoGebra software, offline. This was the first time in GeoGebra Classroom, a short tutorial about functionalities as a teacher and student was given. After this short tutorial, three questions, before the activities, were given trough the Zoom platform, such as: Do you have questions from the last GeoGebra Classroom class? 90% said no and 10% said yes (some doubts); Did the class arouse any curiosity? 100% said yes; Did you have any experience with the GeoGebra Classroom after class? 100% said no. The conclusion is they haven't any doubts with the use of GeoGebra Classroom.

The activities start with an exploratory activity, with two tasks about Eratosthenes arc measurement technique to measure the Earth's circumference. It was about Eratosthenes' the most famous accomplishment, the measurement of the circumference of Earth (Fig. 1). It is possible to see as teacher what the students are doing during the realization of the tasks, we see the eight students and task 2 ("Tarefa 2") at Fig. 2. The tasks placed to the groups were:

1.1 We have to measure the angle formed by the shadow of a stake (indicated further orange on the right) and observe that it is equal to the angle to the center of the Earth, formed by the two cities;

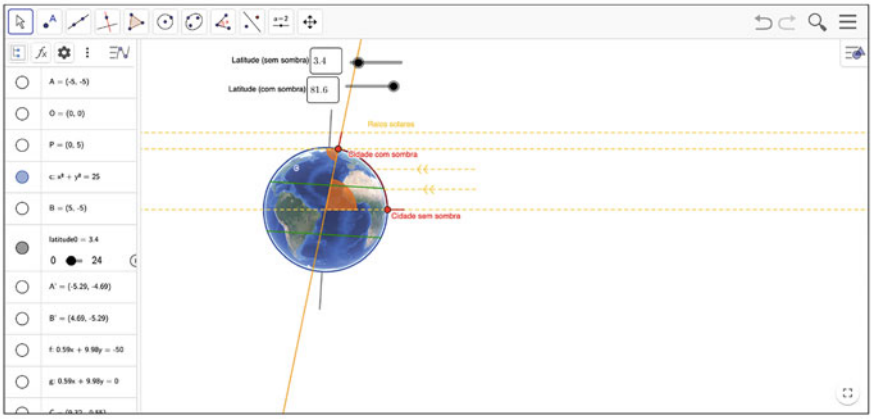


Fig. 1 Perimeter of the world

Fig. 2 Teachers view

1.2 Try sliding the sliders in the figure. Can you explain why are the three angles marked in orange geometrically equal? Use the selectors to position the two cities mentioned in the text: Alexandria (latitude 31.2°) and Aswan (latitude 24°).

We can see what the students have performed, it is possible visualize the students had completed the two tasks realted to this activity.

The second activity had eight tasks about the concepts of angles and parallelism—parallel rays; rays directly parallel; corresponding angles; alternate interior angles; alternate exterior angles; vertical angles and supplementary angles.

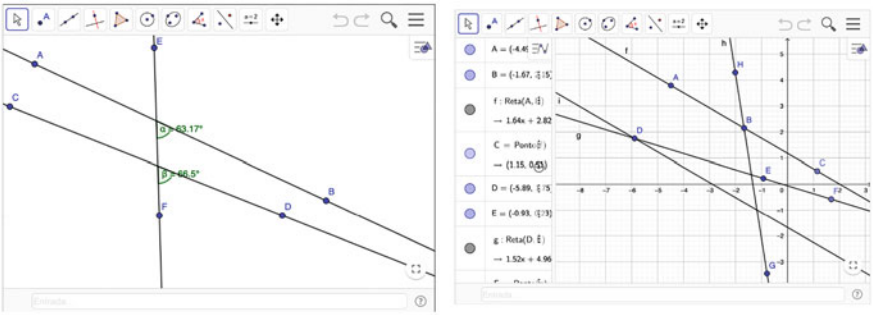


Fig. 3 Student's tasks

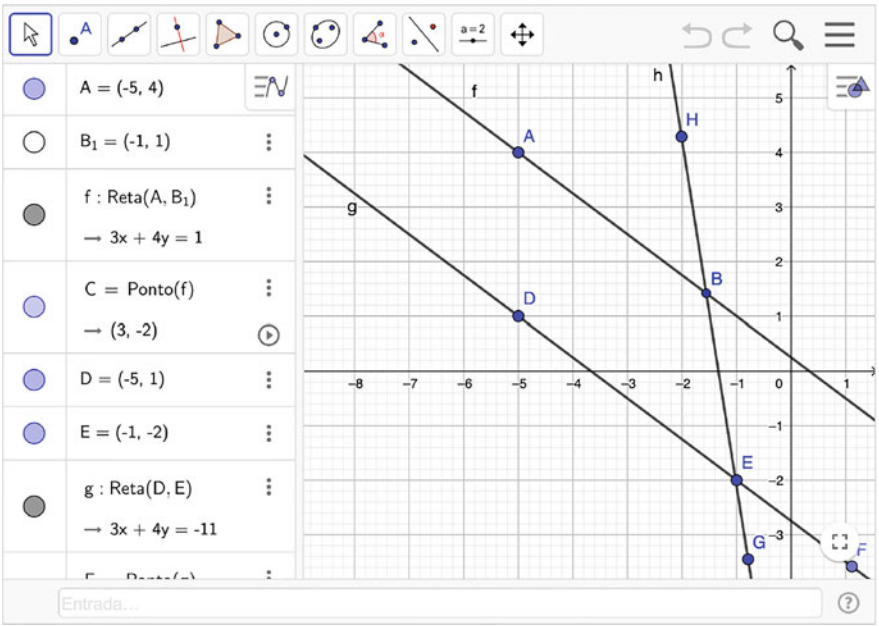


Fig. 4 Student's task

These concept are inline with what they will teach in the future. The tasks placed to the groups were [15] (Figs. 3 and 4):

2. Verify that the AC and DF lines are parallel;
   Using letters from the figure, show:

   2.1 Two parallel rays;
   2.2 Two rays directly parallel;
   2.3 A pair of corresponding angles;
   2.4 A pair of alternating internal angles;

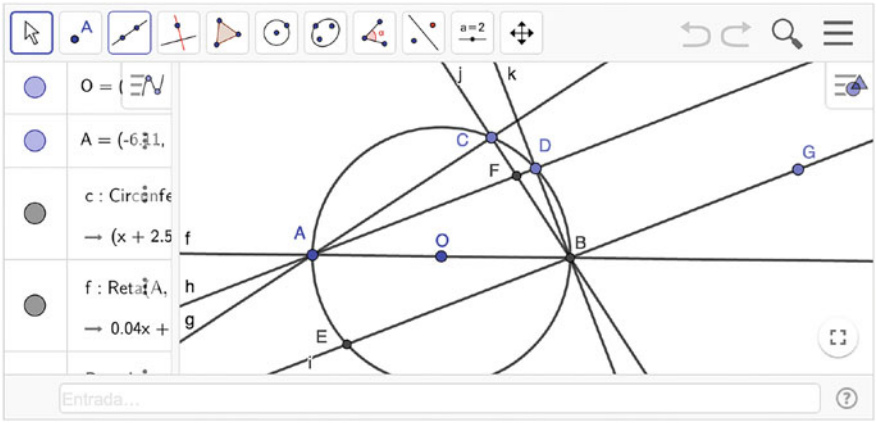


Fig. 5 Student's task

2.5 A pair of alternating external angles;
2.6 Two vertically opposite angles;
2.7 Two supplementary adjacent angles.

3. Do you keep the same answers if the AC and DF lines are not parallel?

The next activity were placed these tasks to the groups (Fig. 5):

4. See the figure below where a circle with center O and some lines that intersect the circle are represented.

4.1 Check that the line AD is parallel to the line EB;
4.2 Identify two inversely parallel lines;
4.3 Identify two corresponding angles, geometrically equal;
4.4 Consider the line AB as a secant line to the line EB and the other line chosen by you (it may vary from the first to the second question). Identify two geometrically equal alternate interior angles;
4.5 Identify two geometrically different alternate interior angles.

These five tasks are about same concepts—opposite parallel lines; corresponding, equal angles; equal alternate interior angles and different alternate internal angles.

The last three tasks was about Thales's theorem (congruent angles from perpendicular sides; congruent triangles) (Fig. 6).

5. One of Thales' theorems tells us that any triangle inscribed in a circle that has one of its sides coincident with a diameter is rectangular (the angle opposite the diameter is right).

5.1 Applying the theorem above identify two congruent acute angles of perpendicular sides;

Fig. 6 Student and teacher view

5.2 dentify two geometrically equal (but not coincident) triangles that have as vertices points identified in the figure above;
5.3 Justify the equality of the triangles you marked earlier.

We see in the image at right the students' work and verify some of them they haven't completed the task.

3.4 Interviews

Structured interviews were conducted using the email interview technique [6]. The purpose is to look into the development as a student and a thinker, the interview was composed of eigth questions. More particularly, the purpose was to determine the extent to which the tools and language of critical thinking have come to play an important part in the way the student go about learning, in school and in everyday life.

The data collection procedure consisted of: elaboration of the interview guide by the investigator; sending an e-mail to each of the participants with the questions with the respective response time; receiving responses sent by participants via e-mail. After receiving the responses by the participants, the content of the interviews was analysed, which consisted of sending the questions in the body of the email, inviting the participants to answer, by the researcher. The interviews proceeded to a content analysis in which it was organized by categories, such as the identification by code was assigned to each interview: Sn (to indicate the student who answered that interview with n=1…6 and the date). When asked about what does critical thinking mean to you? S1 says "Critical thinking involves the ability to assess what is perceived, whether through what is heard or observed, through a careful analysis of the fundamentals behind what is presented." (S1, 21st April, via email). About the role of critical thinking in class, S2 answer "The role of critical thinking in classes is to encourage students to think critically, that is, make them question why and awaken the desire to always know more." (S2, 21st April, via email). Asked how you approach learning new ideas, S3 reply "Nowadays, I believe that, to get closer to learning of new ideas, it is essential to promote a more interactive, more dynamic environment, create new experiences and, in a certain way, bring them

closer to everyday life." (S3, 21st April, via email). Regarding learning materials promote critical thinking, S1 says "These materials must always contextualize the topic addressed in class, and canbe educational games, new texts, books, movies. You can also use digital tools such as MindMeister, Neo K12: Flow Chart Games and ProcessOn." (S1, 21st April, via email). To judge the quality of intellectual work, the criteria you use are "Clarity, precision, accuracy, relevance and depth." (S6, 22nd April, via email). Concerning how does critical thinking apply to the study of mathematics, S6 says "Critical thinking applies to the study of mathematics, since it is only when students use critical thinking that they realize its applicability in everyday situations and problems. In other words, critical thinking in the study of mathematics allows students to analyze their everyday situations and think of innovative solutions for them, through the mobilization of mathematical content." (S6, 22nd April, via email). It was asked to give some examples of the use of critical thinking in your daily life, the responds was "I feel that, in my daily life, critical thinking is part of a set of mental processes that help me deal with major or minor events, from small decisions to major ones (why make one decision over another) . This is manifested a lot in what is presented to me also in the context of the classroom. I stop myself from accepting something as a truth, until I understand why it is the way it is." (S1, 21st April, via email). The last question was about to what extent did your teachers encourage you to use critical thinking and to explain the answer, S6 says "I think the situation in which I felt that my teachers encouraged me to use critical thinking the most was last semester, when we were proposed to develop an educational project that included a proposal for a solution to a locally and globally relevant problem. In this case, during the development of the project, I had to constantly use critical thinking, in order to understand whether the thought solutions were feasible or not." (S6, 22nd April, via email).

4 Discussion

The exploration of these three activities, in a total of 18 tasks, follows the work developed by these students in the 1st semester, the Formative project, which allowed them to have more tools to be able to develop STEM activities in the future using technologies, in this case GeoGebra. Since from the point of view of didactics in the discipline of Mathematics, GeoGebra is used by a wide international community of teachers, its applications in multiple educational contexts have been the subject of numerous studies focusing on student learning in the early years, basic education, higher education, and the teaching of other sciences. Thus, the acquisition of other means of teaching and forms of teaching interconnecting with other subjects allows these students to have a greater mastery of other skills, namely critical thinking, interdisciplinary knowledge and technology.

To promote critical thinking skills it is important questioning. Questions can, in addition to encouraging the promotion of critical thinking skills, facilitate individual student thinking and encourage them to start questioning other students and even

themselves. The analysis made from the responses to the interview allowed us to understand that future teachers have conceptions about critical thinking, revealing that critical thinking is important for solving everyday problems and that teachers should stimulate critical thinking in their students. Students are aware of the concept of critical thinking, because according to S1 "critical thinking is part of a set of mental processes that help me deal with major or minor events, from small decisions to larger ones (why to make one decision rather than another).".

4.1 Limitations of the Study

This study has some limitations, namely, the sample size of participants, which does not allow generalizations; the duration and the observation time, that is, a longitudinal study that allows to verify the effects of teacher training. Another aspect is the interview being by email can prove to be a more rigid/thought out structure of the responses, not being a natural conversation, because of circunstances we lived at that time, in which the issues on the topics addressed could have been further explored. It is recommended to use semi-structured interviews so that it is possible to conduct an interview where the issues under analysis are deepened.

5 Conclusion

It is concluded that the technologies have a significant participation in the educational environment and facilitate teaching and learning. Students realize that the use of technologies is important to promote more interactive lessons and that critical thinking is important to "always want to know more" as referred to by S2. The teacher, when perfecting his pedagogical practice, will be able to insert the technological tools in teaching and learning, to improve the interaction with students and facilitate the improvement of learning with the modelled use of technology in the classroom. The use of the GeoGebra classroom familiarizes students with the scientific process and improves the teaching and learning process and with the participation in the project Formative, in a STEM approach, allows them to improve their future integration of STEM in schools. These student have a prior experience with the project Formative and classes with GeoGebra (offline) was a positive point to these activities at GeoGebra Classroom, because they can use their experience to teach in a STEM approach with their future students.

Acknowledgments This work is funded by national funds through the FCT—Foundation for Science and Technology, I.P., within the scope of the project UIDB/00194/2020 and in the scope of the framework contract foreseen in the numbers 4, 5 and 6 of the article 23, of the Decree-Law 57/2016, of August 29, changed by Law 57/2017, of July 19.

References

1. Bastos, N.R.O., Santos, V.: Using digital tools to enhance active methodologies in higher education. In: EDULEARN21 Proceedings, pp. 12046–12053
2. Creswell, J.W.: Research Design: Qualitative & Quantitative Approaches. Sage, Thousand Oaks (1994)
3. Creswell, J.W., Miller, D.L.: Determining validity in qualitative inquiry. Theory Into Pract. **39**(3), 124–130 (2000)
4. Jones, K.: Providing a foundation for deductive reasoning: students' interpretations when using Dynamic Geometry software and their evolving mathematical explanations. Edu. Stud. Math. **44**(1–2), 55–85 (2000)
5. Kohn, R.K., Lundqvist, U., Malmqvist, J., Hagvall Svensson, O.: From CDIO to challenge-based learning experiences: expanding student learning as well as societal impact? Eur. J. Eng. Edu. **45**(1), 22–37 (2020)
6. Meho, L.I.: E-mail interviewing in qualitative research: a methodological discussion. J. Am. Soc. Inf. Sci. Technol. **57**(10), 1284–1295 (2006)
7. National Council of Teachers of Mathematics (NCTM): Principles and Standards for School Mathematics. National Council of Teachers of Mathematics, Reston (2000)
8. National Research Council (NRC): Learning to Think Spatially. National Academies Press, Washington (2006)
9. NCTM: Principles to Actions: Ensuring Mathematical Success for All. NCTM, Reston (2014)
10. O'Donoghue, T., Punch K.: Qualitative Educational Research in Action: Doing and Reflecting. Routledge, London (2003)
11. OECD: OECD Future of Education and Skills 2030: OECD Learning Compass 2030. A Series of Concept Notes. Concept Note Series, Paris (2018)
12. Pinheiro, M.M., Santos, V.: STEM learning in digital higher education: have it your way! In: Huang YM., Lai CF., Rocha T. (Eds.), Innovative Technologies and Learning. ICITL 2021. Lecture Notes in Computer Science, vol. 13117, pp. 567–578. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-91540-7_58
13. Santos V., Bastos N.R.O.: Critical thinking on mathematics in higher education: two experiences. In: Reis, A., Barroso, J., Lopes, J.B., Mikropoulos, T., Fan, C.W. (Eds.), Technology and Innovation in Learning, Teaching and Education. TECH-EDU 2020. Communications in Computer and Information Science, vol. 1384, pp. 156–167. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-73988-1_11
14. Santos, V., Quaresma, P.: Collaborative aspects in an elearning classroom. In: ICERI2019 Proceedings, vol. 11108 (2019)
15. Santos, V., Quaresma, P.: Exploring geometric conjectures with the help of a learning environment - a case study with pre-service teachers. Electron. J. Math. Technol. **354**, 27–42 (2022)
16. Santos, V., Pedrosa, D., Castelhano, M.: Multiplicity of perspectives in a collaborative environment: geometry workshop using the WGL platform. In: ICERI2020 Proceedings, pp. 5026–5035 (2020)
17. Santos, V., Pais, S., Hall, A.: Mathematics classes for tourism undergraduate students and pre-service teachers with active methodologies using technologies. Int. J. Technol. Math. Edu. **28**(3), 203–212 (2021).
18. Stohlmann, M., Moore, T., Roehring, G.H.: Considerations for teaching integrated STEM education. J. Pre-College Eng. Edu. Res. **2**(1), 28–34 (2012)
19. Su, H.F.H., Ricci, F.A., Mnatsakanian, M.: Mathematical teaching strategies: pathways to critical thinking and metacognition. Int. J. Res. Edu. Sci. **2**(1), 190–200 (2016)
20. Tomaschko, M., Kocadere, S.A., Hohenwarter, M.: Opportunities for participation, productivity, and personalization through GeoGebra mathematics apps. In: Handbook of Research on Mobile Devices and Smart Gadgets in K-12 Education, pp. 45–56. IGI Global, Hershey (2018)

21. United Nations: The sustainable development goals report 2016, United Nations, New York (2016)
22. Vlasenko, K., Chumak, O., Sitak, I., Lovianova, I., Kondratyeva, O.: Training of mathematical disciplines teachers for higher educational institutions as a contemporary problem. Univ. J. Edu. Res. **7**(9), 1892–1900 (2019)
23. Yin, R.K.: Case Study Research: Design and Methods, 3rd edn. Sage, Thousand Oaks (2003)
24. Zochbauer, J., Hohenwarter, M., Lavicza, Z.: Evaluating GeoGebra classroom with usability and user experience methods for further development. Int. J. Technol. Math. Edu. **28**(3), 183–191 (2021)

Alarming Changes in Polish Education vs Longlife and Remote Learning

Ryszard Ślęczka

Abstract The study presents the status and condition of Polish educational system. The article presents its strenghts and weaknesses and also the area of possible threats. The main part of the text presents the basic assumptions of Polish educational reforms, including the main benefits of joining the European Union structures. The last part of the article presents the issue of distance learning and online education.

1 Introduction

In 1990s Polish education system became fully modern and similar to the systems established in other member states of the European Union. The solutions introduced included 6-year primary school and two-stage secondary school. Higher education was divided into bachelor's (licencjat), master's (magister) and doctoral studies. In this context, mandatory education included primary school and stage I of the secondary school. It was considered priority and fully matching the challenges of the contemporary world.

2 Previous Solutions

Like other European community members, Poland has gradually reformed its education system. These reforms were aimed at improving the existing educational solutions and adapting them to the rapidly changing reality. First of them were introduced in our country in 1932, by adopting the Act on the School System [1]. During the People's Republic of Poland (PRL), Polish education system was

R. Ślęczka (✉)
Pedagogical University of Krakow, Krakow, Poland
e-mail: ryszard.sleczka@up.krakow.pl

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_11

subject to numerous changes. The act of 1961 deserves particular attention, as it introduced 8-year primary schools and 4-year secondary schools [2]. It is mentioned purposefully because the solutions it offered are surprisingly similar to the ones existing today. Some significant transformations took place after 1989. Polish education system shifted from the previous (communist) model of state-governed schooling to a more democratic, public and private structure. It was evident that this model was compliant with the provisions of the Universal Declaration of Human Rights, the International Covenant on Civil and Political Rights or the UN Convention on the Rights of the Child. Another important step was gradual joining the Bologna process, which kept bringing us closer to the modern educational policy in terms of school system and science.[1] In the context, the School Education Act (1991) [3] and the Act on the Implementation of the School System Reform (1999), are worth mentioning. Schools were given under the supervisions of the local governments and the whole education model became similar to other European and world-wide solutions. The curricula were remade (new teaching framework and module-based teaching). Active learning methods were promoted to become part of the modern style and model of teaching. The so called Bologna system was to ensure the development of the idea of lifelong learning and promote the European Higher Education Area [4].

The general assumptions were that 80% of the students should complete general education path followed with at least bachelor's degree. The remaining 20% could complete vocational schools. In order to improve the quality of education, a system of external exams conducted by the central and regional examination bodies was introduced [5]. Teachers were categorized into professional promotion groups. The four degrees of professional career are: trainee, contract, appointed and chartered [6]. Within this structure, teachers can be also awarded the honorary title of education professor. Educators were encouraged to take part in different forms of professional development such as: workshops, courses and post-graduate studies [7]. The educational policy of the recent years reveals attempts to introduce some neo-liberal solutions which would limit the role of the state and support new solutions regarding management and organization of the education system and financing thereof.

During this time, an attempt to introduce the European Qualification Framework was made to ensure greater clarity of qualifications obtained, increase learners' and workers' mobility and promote lifelong learning. The main objective was to create a universal model which would enable comparing qualifications gained in different countries and within different education systems. In our country, the National Qualification Framework was established, the goal of which was to support the reforms of the higher education system and education as a whole. The restructured

[1] In June 1999, in Bologna, ministers of higher education from 29 European countries signed the Bologna Declaration which established the European Higher Education Area. During the meetings in Prague (2001) and Berlin (2003), commitments to coordinate the educational policy were made, in order to create a comparable, competitive and globally attractive European higher education system.

curricula included learning outcomes which referred to the level of knowledge, skills, and social competence. They are confirmed using the ECTS (European Credit Transfer and Accumulation System) and ECYT (European Credit System of Vocational Education and Training) points, referring to the accumulated learning outcomes. Within few years that followed the reform, we had reached the European educational standards and could compare with other EU member states, in particular with the best economically developed ones.

3 Today's and Future Perspective

The present functioning of Polish schools is regulated by the 2016 act the Law on School Education (with further amendments) [8], which can be called archaic. It surely steers us away from the modern European and global educational standards. It has re-established 8-year primary schools, 4-year general secondary schools, 5-year technical upper secondary schools as well as 3-year stage I and 2-year stage II sectoral vocational schools. Sectoral vocational schools are an interesting solution, however, they lack adequate funding and support from business and industry sectors. Activity of the Ministry of Education in the recent years cannot be evaluated as positive. In many instances, especially regarding new solutions, this government unit follows the rule of preserving the status quo. Communications presented recently by the Ministry are also very alarming. Representatives of this resort and its organizational units more and more often present xenophobic and racist messages, attacking weaker individuals or people with a different sexual orientation. There are too many examples to recall them all here and they are not encouraging. Thus, perhaps it is worth mentioning some activities carried out during the last years and ask at least two crucial questions which could help us understand the present world.

In Europe, in the context of the wellbeing crisis and the COVID-19 pandemic, some initiatives to support young people were undertaken. One of them was "Youth on the Move" aimed at helping young people acquire relevant knowledge, skills and experience. The initiative was included in the Europe 2020 strategy and proposed 28 key actions to adapt education systems to the needs and interests of the young European citizens. It promoted international studies and trainings, which facilitate the entry of young people into the open labor market (European employment area). One of the programme's objectives was to reduce the share of early school leavers (increase the share of young people with tertiary education or its equivalent). This strategy was to facilitate joint solutions which should contribute to the increase of the employment rate. In the area of experience exchange, a dialogue between the member states and the European Commission was undertaken. One if its most important documents was "Employment Guidelines", approved annually by the Council of Europe. The document included joint employment report and important recommendations addressed to specific countries to improve the whole employment process. In our country, the "2020 Human Capital Development Strategy" was successfully implemented. It has been prepared in a way to facilitate full use of this

capital in social, political and economic life. The strategy was used to: increase the employment rate, prolong the professional activity and ensure better functioning of the senior citizens, improve the situation of persons and groups at risk of social exclusion, take care of the citizens' health and increase the effectiveness of the healthcare system, improve the level of civic competence and qualifications. Thanks to these and other activities Poland has the one the lowest unemployment rate in Europe and in the world.

Was it important for the young generation of Poles to participate in programmes which supported education system financially? We had access to the following European programmes: SOCRATES, LEONARDO DA VINCI, MINERVA and other. They facilitated European cooperation in terms of open and remote education, information and communication technologies and other similar initiatives. For example, GRUNDTVIG programme supported development by providing resources and services to be used in adult education. The second edition of LEONARDO DA VINCI facilitated the development of vocational education through international projects which focused on the improvement of the lifelong and vocational learning systems and the increase of employment opportunities through promoting innovation and entrepreneurship. SOCRATES-Erasmus was addressed to university students. This programme is very popular among the students as it allows them to gain new experiences and explore cultures in other countries. Along with increasing their competitiveness on the labor market, it also gave them a chance to shape their personality and individual value systems [9].

Was it a mistake to join the Eurydice system (the Education Information Network in Europe) which exists from 1980 and supports cooperation in the field of education? It works as an ongoing information exchange network between the national units (set up by education ministries) and the central unit in Brussels. The national units are responsible for providing basic information about their local education systems, which is then processed and published as general data about the school systems. One of the numerous publications of central unit was the report ("Key Data on Education in Europe") which presented the most important changes in education systems during the last decade. It provides statistical and qualitative descriptions of the phenomena that occurred in education. Its authors made an attempt to answer many questions critical for the future of the young generation (for example, what actions do European countries undertake to reduce the share of early school leavers).

4 In Developing Remote Education and Online Learning

Polish education system, like others in Europe and around the world, has introduced solutions rooted in the concepts of lifelong learning. These solutions using media, that is, television (EDUSAT channel) and Internet access with wide access to educational programs. However, experience has shown that the above mentioned forms of learning are still underused and not very popular compared to traditional,

stationary teaching forms. In this context, the experiences of the last dozen or so months (almost two years) of COVID-19 pandemic may be significant. We had an opportunity to test the whole education system nation-wide in terms of modern teaching forms (for example, remote and online teaching). It seems that conclusions drawn from the research conducted are not optimistic. The identified barriers include: under-funding of education system, old technology and poor awareness of modern challenges among students and teachers. More detailed limitations involve education drop-out or information and scientific chaos (remote school skipping, lack of psychological assistance addressed to students, teachers and parents, unwillingness to show one's own home). The positive aspects include the culture of openness, connected with exchange of experience as well as potential of remote education focused on developing students' individual talents. Students who do not have to always be top achievers.

5 Conclusion

As a summary, it must be stated that the existing legal regulations of the European Union enable maintaining full social, cultural and educational distinctness. According to the Art. 165 and 166 of the Treaty on the Functioning of the European Union, education remains within the sphere of competence of the member states. EU institutions focus only on coordinating and supporting the actions taken (the so called "open method of coordination"). It does not mean that education system chosen by certain national subjects is less important or irrelevant. On contrary, it seems essential when it comes to building the new European society. Thus, Polish approach to education should be expanded to include wider horizons instead of limiting it to tradition or someone's childhood memories. It should be treated globally, as a sphere where specific educational activities are implemented to prepare young people to enter the labor market, participate in cultural life or identify with the civic society.

References

1. Pęcherski, M., Świątek, M.: Organization of Education in Paland During 1917–1977. Podstawowe akty prawne, Warszawa (1978)
2. Act of 15 July 1961 r.: The Law on the Development of Education and Upbringing System. J. Laws No. 32, item 160 (1961)
3. School Education: Act of 7 September 1991. J. Laws No. 95, item 425 (1991)
4. Kraśniewski, A.: Bologna Process: Where is European Higher Education Heading? MEiN, Warszawa (2004)
5. Regulation of the Minister of National Education of 21 March 2001 on conditions and methods of assessment, classification and promotion of students and organizing examinations and tests in public schools. J. Laws No. 29, item 323 (2001)

6. Regulation of the Minister of National Education of 3 August 2000 on the professional promotion grades for teachers. J. Laws, No. 70, item 825 (2000)
7. Key Data on Teachers and School Leaders in Europe, Raport Eurydice. MEiN, Warszawa (2013)
8. Act of 14 December 2016 the Law on School Education. J. Laws, item 1082 (2021)
9. ERASMUS in Poland in the 2003/05 academic year. Mobility of students and academic teachers. Study by the Foundation for the Development of the Education System, National Program Agency SOCRATES-Erasmus. MEiN, Warszawa (2006)

The Most Common Mathematical Mistakes in the Teaching of Scientific Subjects at Secondary Schools

Zuzana Václavíková

Abstract Mathematics is a subject with perhaps the greatest overlap with other fields, especially the natural sciences. As a secondary effect of a project focused on creating and piloting problem tasks in the field of chemistry and physics, utilizing inquiry-based learning, we observed the most common mathematical mistakes and conceptual errors that are made by students using mathematical knowledge in other areas of the natural sciences. A total of 40 problem tasks were created and verified in cooperation with secondary-school teachers of physics and chemistry and more than 650 solutions by student were qualitatively assessed. The paper will present the most common mistakes and errors repeated across all tasks and compare their occurrence between teachers who have and do not have mathematics as a secondary subject. The mistakes and errors will also be explained from a mathematical point of view and a proposal will be outlined on how to innovate the teaching of mathematics in secondary schools. This should lead to the correct understanding of the issues and the elimination of the errors found by this research.

1 Introduction

1.1 STEM Education in Czech Curriculum

Mathematics, as one of the disciplines in STEM education, requires the crossing of boundaries between it and others subjects for learners to develop a better understanding [1, 2]. Even so, the implementation of STEM education in the Czech curriculum is still not very noticeable and mathematics is often taught completely without any contextual relationship to other subjects. As a result, students do not transfer their mathematical knowledge to other subjects and are essentially unable to apply mathematics. There are many reasons why STEM education is so

Z. Václavíková (✉)
Faculty of Science, University of Ostrava, Ostrava, Czech Republic
e-mail: zuzana.vaclavikova@osu.cz

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_12

difficult to implement into Czech schools. One of the main reasons is probably the fact that teachers themselves do not often have any awareness of interdisciplinary relationships and teach individual topics without any connection to their application. Therefore, in the framework of the education of future teachers and in the framework of cooperation with active teachers, it is now crucial for us to work on activities that support interdisciplinarity [3].

1.2 Mathematical Errors: Research Overview

Error identification, error analysis and error handling are the most important starting points for researching teaching and the learning process, not only in mathematics, but also for any scientific discipline.

According to Radatz's [4, 5] historical survey, error analysis has been of interest to the mathematics education community for at least one hundred years. research interest focused on:

- Listing all potential error techniques;
- Determining the frequency distribution of these error techniques across age group;
- Analyzing special difficulties, particularly encountered when doing written division, and when operating with zero;
- Determining the persistence of individual error techniques;
- Attempting to classify and group errors.

Analyzing students' errors may reveal the faulty problem-solving process and provide information on the understanding of and the attitudes toward mathematical problems [5]. The purposes of error analysis are to

- Identify the patterns of errors or mistakes that students make in their work;
- Understand why students make the errors;
- Provide targeted instruction to correct the errors [6].

Much of the research, as well as more general studies on mathematical errors, focuses on understanding the underlying cognitive causes of these errors, either in order to understand the cause of specific errors, or more generally to identify the mechanisms underlying these errors.

In general, any research focused on the identification of mathematical errors is usually concerned only with mathematics itself, and not errors arising from the transfer of learning mathematical procedures taught in other subjects [6–11].

1.3 Mathematical Errors: Categorization

There are many possible approaches for error categorization and error taxonomy in mathematics [4–11].

The most cited categorization of mathematical errors is probably the Radatz categorization [4, 5]. Five categories of errors are identified:

- Errors due to language difficulties;
- Errors due to difficulties in obtaining spatial information;
- Errors due to a deficient mastery of prerequisite skills, facts, and concepts,
- Errors due to incorrect associations or rigidity in thinking;
- Errors due to the application of irrelevant rules or strategies.

Radatz argued that most mathematical errors are causally determined, and very often systematic [5].

Riccomini [12] outlined four types of errors in his research: procedural, factual, careless, and conceptual. Procedural errors occur when students working on the wrong order. Factual errors are computational errors and occur when students, for example, cannot identify sign, digit or use incorrect formula. Careless errors occur when students not paying attention (working too fast, making wrong count, writing the wrong number or not following the direction), and conceptual errors occur when students have misconceptions and poor understanding of mathematical concept, procedures, and applications.

The specific research directly focused on taxonomy of mathematical errors also exists. It is usually oriented towards a specific age group of pupils or students (elementary school, secondary school, high school) and errors in a specific mathematical area (algebraic operation, proportion, algebra, etc.). For example, Ford, Gillard and Pugh [10] developed a taxonomy of errors which undergraduate mathematics students may make when tackling mathematical problems. Each error is given a code to allow for quick reference to the error when providing feedback to students on their work.

Ben-Zeev [11] constructed a taxonomy of mathematical errors and attempted to identify the causes of these errors by integrating findings from different studies. The focus in this and other research is to understand why a student makes an error. A student may over-generalize an algorithm which holds in one context to a structurally similar context where the algorithm no longer works, something Ben-Zeev calls syntactic induction. However, this was done only in the context of mathematics teaching.

Research aimed at studying mathematical errors is driven, among other things, by the fact that in recent years, error analysis, incorrect exercises method and student-conducted error analysis aligns with the standards of mathematical practices and mathematical teaching practices.

1.4 Mathematical Errors: Our Object of Interest

One of the fundamental problems of the Czech curriculum is that the curricula of individual subjects are not coordinated. Related to this is the fact that if a mathematical apparatus is needed in chemistry for a selected topic, and students have not learnt it in their mathematics lessons, the relevant part of mathematics is taught by a chemistry teacher, and not a mathematician. Subsequently, when the topic is included in mathematics, students already have knowledge and skills in this area from a different subject, but students can also transfer some misconceptions. Of course, with the implementation of STEM education, the occurrence of errors due to incorrect associations or rigidities of thinking can arise. This is due to the fact that within a given field, applications from another field are taught from the point of their usefulness, and not always primarily with a professional approach, therefore, depriving students of a deeper theoretical understanding of the underlying principles.

2 The Research

The research was conducted using a qualitative method and used data obtained from the project "IBSE as a tool for acquiring pupils 'and teachers' abilities and attitudes to technical and scientific education with regard to market requirements", implemented as a cross-border cooperation with the University of Trnava. It was aimed at creating activities for teachers and students that would combine topics from physics, chemistry, and biology with the applied mathematical knowledge found within them.

2.1 Data Collection

During the project, research-oriented tasks utilising an inquiry-based educational approach [10, 11] for pupils aged 14–17 were prepared and verified. 8 experienced teachers with experience from 4 schools were involved in the task creation process. In addition, another 32 in-service teachers were involved, who were able to provide us with information on how they teach a given area of mathematics in their scientific subject. The tasks were thematically divided into 4 blocks: water, air, colours and temperature, and 10 tasks were prepared for each topic. The topics were deliberately chosen to relate to all-natural science subjects learnt at school, and to make use of the mathematical apparatus that pupils in a given age group have knowledge of. A methodological sheet by the author was prepared for each task, and this was intended for the teachers. It described the inclusion of the topic in the curriculum of the individual scientific subjects, the necessary tools for the experiment, the

specified target group, and the recommended age of the students. It described the research experiment in detail and provided hint questions in case the students did not know how to find the answer for the research task. Furthermore, a worksheet was created for each task, containing a motivational text and a research question—this was the goal of the research itself. The procedure for setting up the experiment, the tools needed, and the actual experiment itself, were left up to the students to complete. Students recorded the obtained or measured data on a worksheet, performed calculations, and formed conclusions from their research.

All created tasks were then piloted in a class of students who solved the research tasks in groups of three, but then completed the worksheet individually. After the implementation of all the pilot tasks, we evaluated the completed worksheets from the perspective of specific areas of science and evaluated the mistakes made by the students.

In total, we obtained over 650 solutions from the students from the 40 research-oriented tasks.

2.2 Methods and Tools

It was necessary to use mathematics in each worksheet. This mainly consisted of working with data (working with tables, creating and reading graphs, interpreting measured data, etc.), working with physical or chemical relationships (including working with units), descriptive statistics, working with percentages, interpolation, understanding the dependencies of quantities, and working with functions.

After obtaining all the data, and by cooperating with mathematicians and pedagogues of the natural sciences, the mathematical topics that appeared in the worksheets were divided into two groups, depending on whether it was first taught in mathematics or in another scientific subject.

Furthermore, we only dealt with the mathematical areas of the second group, i.e. those that were taught in a subject other than mathematics. Although we did not record the students' personal data, we did register the school and class they were from. This made it possible to organize the worksheets according to who has taught the specific topic. This combine with information regarding the specialization of individual teachers, made it possible to evaluate whether the subject was taught by a mathematician (i.e. the teacher had a specialization in combination with mathematics) or a non-mathematic teacher.

In our research, with regard to Riccomini [12], we chose not to focus on the procedural, factual and careless errors as although the students were working in groups, they completed the worksheets individually which eliminated these errors. Throughout they compared the notes and discussed their findings and conclusions. If someone made a procedural, factual or careless error, the others pointed out to it and he checked his calculations. Instead, we chose to focus on conceptual errors when using mathematics in other fields of the natural sciences. More specifically, this

examined the students' general misconceptions relating to their poor understanding of mathematical concepts, procedures and applications.

The evaluation of individual errors took place by comparing the worksheets completed by students with the solution provided by the teacher who had designed the research-oriented task, and with the teacher who had taught the topic to the students. The evaluation was also accompanied by interviews conducted with both students and teachers where a provided solution used a non-standard procedure but was still correct. This was done to determine where the solution had come from, either the student or the teacher, and was related to two situations; either mathematical procedures that are taught within the target group of students in mathematics, but without an applied context, or conversely, procedures used in mathematics but which had been taught for the first time in another subject with no obvious link to mathematics.

2.3 Research Questions

With regard to Radatz [7], the following research questions were used:

- Are there any errors that arise due to the fact that the application of mathematical procedures is not taught in mathematics, i.e. the mathematical apparatus is taught in a subject other than mathematics?
- If so, can misconceptions lead to errors in solving pure math problems?
- How often does the error occur?
- Is its occurrence affected by whether the mathematical topic was taught by a mathematician or a non-mathematician?

At the end of the research, we asked ourselves how to prevent such errors.

3 The Most Common Mathematical Mistakes

From the point of view of mathematics, we focused on monitoring the most common mistakes that students made. At the same time, we tried to track whether mistakes occur across the whole class, or in selected students, and whether they occur, or reoccur only in cases where the students were focused on another research area (chemistry or physics). We also observed whether the teacher of the scientific topic that the research question was focused on (chemistry, physics) also made the same mathematical errors and transmitted these mathematical inaccuracies or misunderstandings of concepts to the students.

3.1 Equality Relation

3.1.1 Students' Worksheets

The most common mathematical mistake that occurred in all worksheets was the incorrect interpretation of the equality relation, usually by writing numerical values of physical or chemical quantities into relations with/without units on the different sides of the equalities. If we write equality in mathematics, it means that the expression on the left side is identical to the expression on the right side. In lessons of mathematics with a general notation of expressions using x, y, $f(x)$, ..., it is absolutely clear. The problem occurs when students work with physical or chemical quantities that have a specific unit. Here it is necessary to realize that the unit actually indicates how many times the given measure is greater than the measure of the unit. For example, when we write the weight as $m = 25.3\,\mathrm{g}$, we say that the weight is 25.3 times greater than one gram. Simply, we work with a unit as if there was a mathematical operation "times" (i.e. multiplication) between the numerical data and the unit. We can write

$$m = 25.3\,\mathrm{g} = 25.3 \cdot 1\,\mathrm{g}.$$

The sign for multiplication, as everyone knows, does not have to be written in mathematics, for example $5x = 5 \cdot x$. In Fig. 1 we can see the student's solution for the calculation of density on a worksheet focused on the density of unknown substances. It is clear that

$$\frac{2.024}{0.002} = 1012$$

but

$$\frac{2.024}{0.002} \neq 1012\frac{\mathrm{kg}}{\mathrm{m}^3}.$$

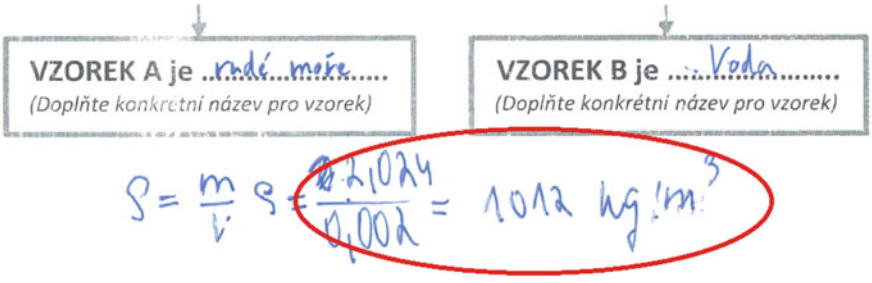


Fig. 1 Student's solution—incorrect equality relation

If we did not know (or if the students did not know) what the notation means and how it is written in mathematics, they would modify it by shortening and they would obtain

$$\frac{2.024}{0.002} = 1012\frac{\text{kg}}{\text{m}^3}$$
$$1 = \text{kg/m}^3$$
$$\text{m}^3 = \text{kg}$$

This is of course, nonsense, even though mathematically it is perfectly fine—the problem is that the "equality" does not hold and this is why the solution is completely incorrect. Therefore, we have to substitute either all the values without units (only a mathematical calculation), or write the units on both sides.

3.1.2 Teachers' Methodology Sheets

It is not only a problem of the students, because in some methodological sheets the same type of error also appeared. For example, as Fig. 2 shows, in the teacher's solution for the research task focused on the density of unknown substances, the same issue can be observed. In this case, the teacher was a chemistry teacher in combination with biology. After indicating the error, the teacher argued that such notation is common in chemistry and did not perceive it as a mistake. This is where the very problem of teaching without interdisciplinary connections can arise. The creation of mathematical mistakes when viewed from the perspective of other subjects, are transmitted through teachers to their students. The students and teachers will then both continue to make mistakes in this way. At the same time, the correct notion of concepts and procedures is harmed, because it seems that "what is true in mathematics works differently in chemistry".

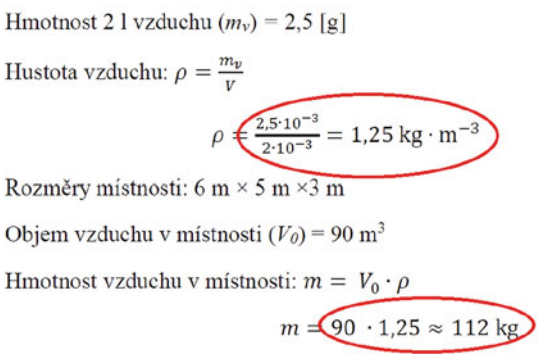


Fig. 2 Teacher's solution—incorrect equality relation

3.1.3 Using the Units: Why It Is the Best Approach

Thus, the question arises of whether to always teach students to use units when working with physical or chemical quantities, or to let them work on the calculation without units and then interpret the result with the units. The more useful way, of course, is to teach students to always use and write units. There are several reasons for this.

If we write the quantities with the units, we do not have to later remember which units to use with the calculation—it is clear from the calculation. In the above case, the student does not have to remember that the unit of density is $\frac{kg}{m^3}$.

If we use the mass and volume correctly with the units, the result will be correct. A student confusing the unit of density in their interpretation will not occur, this was shown by one student's worksheet, as shown in Fig. 3. For solving the problems, we do not need to memorize formulas but we need to understand the relationship between the variables. The relationship, the unit and the solution are very closely connected. The knowledge of one leads to the knowledge of another. When there is more mass in the same volume, the substance will be denser—knowing the unit makes the calculation of the formula clear.

Additionally, the opposite is true—if a student knows the unit, he/she does not have to remember the relationship for the calculation. In the case of density, if he/she remembers that density is given as $\frac{kg}{m^3}$, it is clear that it must be a ratio of weight in kilograms and of volume in cubic meters.

If we always use the units, we will not lose the point of what we are counting. This was shown for example, in the worksheet focused on minerals found in water. The student correctly calculated the percentages, but at the end, added a unit of grams to the result (how many percent). This further confused him in the task and it was obvious that it was not clear to him how he should handle the result of "0.01 g", see Fig. 4.

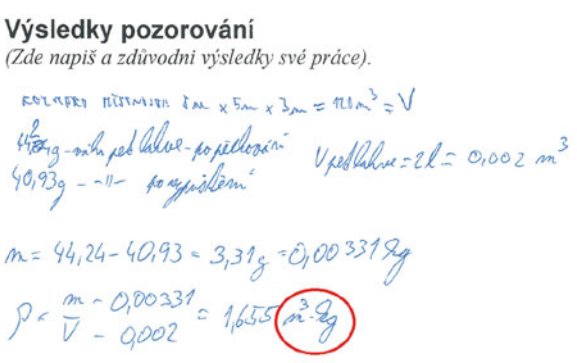


Fig. 3 Example of the student's worksheet

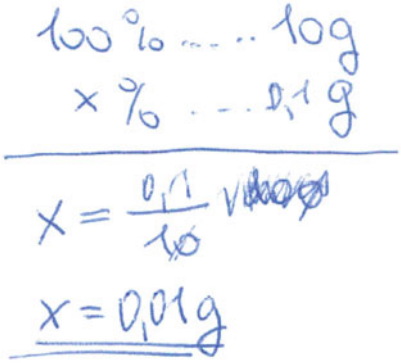


Fig. 4 Example of the student's incorrect interpretation of the calculation

3.2 Problems with Rounding Results

Another common mistake found, was with the rounding of the results, i.e. the interpretation of results with regard to the accuracy of the input data. If we have a measuring tool with a certain level of accuracy, or if we have data with fixed decimal places already appearing in the relation as an absolute term, the result cannot be rounded to higher decimals than the smallest number of decimals given in the relation. It is not possible to obtain by calculating, a result more accurate than the values entered into the relationship. This always results in a bigger error, so we can never round the result to more decimal places than the input data has. Thus, if the values given in the relationship are rounded to only three decimal places, we do not report the result more precisely than to three decimal places. However, for example, with the chemistry worksheets there was a direct requirement to round the result to 6 decimal places; although the input data was rounded to four decimal places as Fig. 5 shows (in this case other students even used seven decimal places).

3.3 Work with Decimal Notation

The last common mistake is a misinterpretation of decimal notation. Although it may not seem so with ordinary notation, in mathematics, when working with decimal numbers, we make a distinction between the numbers 5.2 and 5.20. In the first case, this means a number rounded to two decimal places, which may 'represent' the numbers 5.20, 5.21, 5.22, 5.23, 5.24 etc. The second example unambiguously indicates that the number is determined to two decimal places and the second digit after the decimal point is exactly zero. The "omission" of zeros at the end of decimal notation appeared in worksheets in all of the scientific subjects. It was then not clear how many decimal places in the relationship were actually being

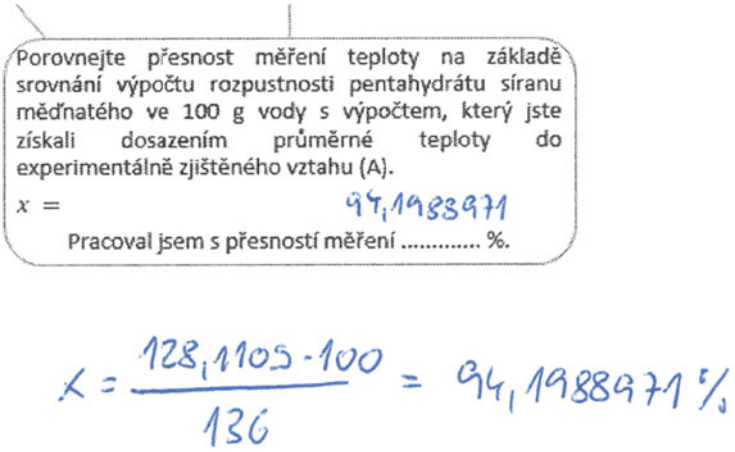


Fig. 5 Rounding the results of computation by students

worked on, and the students rounded up the results of their calculations to a different number of decimal places each time.

3.4 Other Inappropriate Mathematical Procedures

Other inappropriate mathematical procedures included, quantifying ongoing intermediate results when obtaining the overall result from multiple relationships. This is understandable and acceptable for younger students who cannot work with algebraic expressions in a general form; however, it should no longer occur at a high school level. The correct mathematical procedure is such that we first express the final relation before substituting numerical values, then we calculate the total result from the entered values. With partial calculations we increase the error of the result (we round it several times).

A good habit to form is, before measuring and according to the possibilities of the tools, to determine the accuracy with which we will calculate and to leave this for the entire solution of the problem. Then it would be clear throughout the measurement and calculation how the resulting values should be rounded. In Fig. 6 we can see that the students have not learned in this way. Despite using the same device during their measurement of one of the chemistry problem tasks, they still rounded the results completely differently.

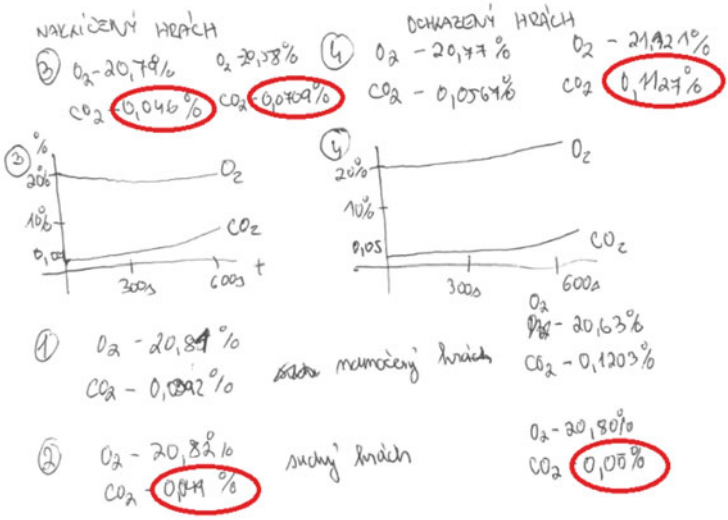


Fig. 6 Rounding the results of measurements by students

4 Conclusions

During this research, the answer to the question of whether there are the errors that arise due to the fact that the application of mathematical procedures is not taught in mathematics, was discovered. In the conducted interviews, mathematics' teachers confirmed that these errors, especially for weaker students, subsequently present problems and misunderstandings in mathematics. As mentioned in previous sections, it appears that the mistakes are indeed taken from the teachers. These mistakes did not occur in only one case, specifically, with the worksheets of a teacher who taught a combination of mathematics and physics. None of the mistakes described above appeared in either the methodological sheets or during the piloting of the tasks for the class of this teacher. For teachers who do not teach mathematics as a secondary subject, these errors occurred on almost all worksheets and with all students in their classes.

This brings us back to the crux of the whole issue, the solution to which could be brought about by the implementation of STEM education into the Czech curriculum together with coordinating the curricula of mathematics and the natural sciences. However, there is also the problem of needing to train active teachers who are not used to using such a model. Their reaction to highlighting their errors is often that it is irrelevant to their scientific subject and they do not perceive them as mistakes in their subject. Good communication between teachers of subjects, where the content overlaps or where knowledge from another subject is used in another, is also important. The first step in rectifying this, was organizing training for the teachers involved as well as their school colleagues. An expert outlined the mistakes that

appeared in the worksheets from other subjects and explained why it was incorrect and how it should be done correctly. Unfortunately, the way to improving the issue is a task for the long term, and the better use and integration of STEM education could provide only a partial solution in the near future.

Acknowledgments This research was funded by INTERREG V-A Slovak Republic—Czech Republic, EU—European Regional Development Fund, grant number NFP304010T963.

Conflicts of Interest The author declares no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Thibaut, L., Ceuppens, S., De Loof, H., et al.: Integrated STEM education: a systematic review of instructional practices in secondary education. Eur. J. STEM Edu. **3**(1), 02 (2018). https://doi.org/10.20897/ejsteme/85525.
2. Michael Shaughnessy, J.: Mathematics in a STEM context. Math. Teach. Middle School **18**(6), 324–324 (2013). https://doi.org/10.5951/mathteacmiddscho.18.6.0324
3. Treacy, P., O'Donoghue, J.: Authentic integration: a model for integrating mathematics and science in the classroom. Int. J. Math. Edu. Sci. Technol. **45**, 703–718 (2014). https://doi.org/10.1080/0020739X.2013.868543
4. Radatz, H.: Error analysis in mathematics education. J. Res. Math. Edu. **10**(3), 163–172 (1979). https://doi.org/10.2307/748804
5. Radatz, H.: Students' errors in the mathematical learning process: a survey. Learn. Math. **1**(1), 16–20 (1980)
6. Cohen, L.G., Spenciner, L.J.: Assessment of Children and Youth with Special Needs. Pearson Allyn Bacon Prentice Hall, Boston (2010)
7. Themane, K.M., Luneta, K.: Misconceptions and associated errors in the learning of mathematics place value in south african primary schools: a literature review (preprints 2021). https://doi.org/10.20944/preprints202105.0456.v1
8. Movshovitz-Hadar, N., Zaslavsky, O., Inbar, S.: An empirical classification model for errors in high school mathematics. J. Res. Math. Edu. **18**(1), 3–14 (1987). https://doi.org/10.2307/749532
9. Quinney, D.: So just what is conceptual understanding of mathematics? MSOR Connect. **8**(3), 2–7 (2008)
10. Ford, S., Gillard, J., Pugh, M.: Creating a taxonomy of mathematical errors for undergraduate mathematics. MSOR Connect. **18**(1), 37–45 (2019)
11. Ben-Zeev, T.: Rational errors and the mathematical mind. Rev. Gen. Psychol. **2**(4), 366–383 (1998). https://doi.org/10.1037%2F1089-2680.2.4.366
12. Riccomini, P.J.: How to use mathematical error analysis to improve instruction. In: Webinar Error Analysis to Inform Instruction. The Pennsylvania State University (2016)
13. Anderson, R.: Reforming science teaching: what research says about inquiry. J. Sci. Teacher Edu. **13**, 1–12 (2002). https://doi.org/10.1023/A:1015171124982
14. Barron, B., Darling-Hammond, L.: Prospects and challenges for inquiry-based approaches to learning. In: Dumont, H., D. Istance, F. Benavides (Eds.), The Nature of Learning: Using Research to Inspire Practice. OECD Publishing, Paris (2010). https://doi.org/10.1787/9789264086487-11-en

Challenges to the Development of Effective Creativity

Zhanat Zhunussova, Vladimir Mityushev, Yeskendyr Ashimov, Mohammad Rahmani, and Hamidullah Noori

Abstract One of the key problems for students, especially from developing countries during pandemic is a lack of the electronic materials. Obviously, there are a lot of problems arise during education in the online regime. First of all, it is a weak Internet connection combined with the lack of appropriate equipment. Even having a higher quality computer they could not setup a program. It is connected to the both their knowledge and the specialty. For example, the students biology, chemistry, philology and others are not taught to computational skills. In general, they are users as ordinary people. But the real situation concerning pandemic requires systematic changes procedures in computer education. All these changes should be unified for all the students in a group independently on their country. That is why the math teachers have to look for an alternative method to make a proper decision for a stated problem under supervision of works devoted to projects and diploma.

1 Introduction

We pay attention to the textbook [1] which may play the role of a guidance for teachers as well as for students. This book is for beginners and could be useful for higher qualified specialists in various domains not familiar with the main principles of modeling. It is worth to say that the textbook doesn't fit to a researcher like a PhD student in Mathematics or Physics who should concentrate her/his attention to a special problem and deeply go into the considered subject, because such a

Z. Zhunussova (✉)
Institute of Mathematics and Mathematical Modeling, Al-Farabi Kazakh National University, Almaty, Kazakhstan

V. Mityushev
Faculty of Computer Science and Telecommunications, Cracow University of Technology, Kraków, Poland

Y. Ashimov · M. Rahmani · H. Noori
Al-Farabi Kazakh National University, Almaty, Kazakhstan
e-mail: wladimir.mitiuszew@pk.edu.pl

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_13

PhD student understands the field of its work and knows the key notions and methods. In the same time, the book can be useful for a PhD student in Biology, Engineering and so forth who deeply knows special topics and needs to apply computer simulations. The textbook [1] is organized in such a way that it contains a minimum of information about modeling for a beginner who can find in the book the corresponding references and the proper key words in order to find the necessary sources in internet. For instance, in order to investigate the trajectory of the thrown stone one has to keep in mind the terms *gravitation, velocity, acceleration* to find the corresponding equation for the gravitational law and further to solve a problem. In order to model public traffic in a city one has to know the key notion of the *graph* theory. Traditionally, a teacher has the task to explain for a student *"what"*. A student shouldn't deeply know everything. Knowledge besides the fundamental theories and methods includes the option *how to find it in internet*. What is the best way to solve a problem? To apply a theory, to use a computer for fast computations? No, just to write the answer, maybe find it internet. If one needs a date or something like this, no problem. But if a student has to determine a force acting on an object, then usual way to put a button doesn't work. The student must know which button to put. The main question consists in the understanding what to look for. The conception of [1] concerns the study of the fundamental theory and its usage to quickly find the necessary information. In this sense, the textbook [1] is superficial, it doesn't contain all the theories. Roughly speaking, it answers the questions *"what"* and *"how"* simultaneously.

The textbook [1] contains a systematic description how to develop a mathematical model and explain the main steps. The strategy consists in the following steps:

- to introduce spatial variables (description of geometry) and time;
- to think about dimensional units; to introduce the units perhaps during solution of the mathematical problem;
- to introduce assumptions and conditions, first, as simple as possible;
- to formulate the law (physical, economical, biological, empirical etc.);
- to develop a mathematical description, first, as simple as possible;
- to try to solve the mathematical problem "by hand"; if that does not work, to try a numerical method; to compare the results if different methods are applied;
- verification of the model; if the results are suspect to get back to the previous steps.

These steps are illustrated by examples.

The textbook consists of three parts: general principles and methods, basic applications and advanced applications. The first part is divided into two sections which introduce to principles of mathematical modeling and numeric and symbolic computations. With considering of a simple example of the free falling object from a height h a reader can understand how to describe the trajectory of the object. The description of projectiles by means of an ordinary differential equation is presented. Next, the mathematical problem, more precisely Cauchy problem, is solved. Hence, the required particular solution is obtained. By this way a principle of hand calculations is described in detail. Further, some formulas from the example

are checked in the package $Mathematica^{®}$. The graph of the obtained trajectory is drawn with numerically given parameters. One can use acquired skills from the example in order to check equations by calculation of derivatives and integrals. In fact, it takes time by hand. Especially, for math teachers who must verify a lot of control works for the limited time. Nevertheless, it should be noted, that incorrect use of computer by someone yields incorrect results. Such kind result is shown on the example about calculating of few partial sums of the series. The series as harmonic series diverges, although direct observations of the results could lead to the conclusion that the series converges, see the fragment below

$$In[1] := S_{n\_} := \sum_{k=1}^{n} \frac{1}{K}$$

$$In[2] : Table[S_n, \{n, 1660, 1670\}]$$

$$Out[2] = \{7.99209, 7.99269, 7.99329, 7.99389, 7.9945,$$

$$7.9951, 7.9957, 7.9963, 7.9969, 7.9975, 7.99809\}$$

By this way, a principle of the stupid computer formulated as follow. Do not trust the computer and try to check the result by hand. Even possessing computational skills, one has to get a fundamental theoretical knowledge, simultaneously update your skills in computer sciences.

In such a way, the development of mathematical model is gradually discovered in the textbook [1]. The steps of the development are demonstrated in a simple example, the falling of an object in vacuum and in air. The classification of the mathematical modeling is proposed by types as deterministic and stochastic, discrete and continuous, linear and non-linear. These types are explained by examples.

2 Examples

The special attention is paid to

Principle of transition: continuous ↔ discrete. *To divide a continuous object into small parts, to apply a simple formula to every part, to calculate the sum for all the parts and get back to the continuous object through the limit operation.*

This principle is used in the standard course of calculus in introduction of Riemann integral as the area under the positive graph. Many equations of physics, biology, economy and other sciences are based on this principle. For instance, the radioactive decay satisfies the physical law $\delta m = -km\delta t$, i.e., the increment of mass is proportional to the total mass and time. Next, this rule leads to a differential equation and its solution $m(t) = m(0)exp(-kt)$. Such an approach plays the

fundamental role in mathematical physics and further can be used, for instance, in fluid mechanics when the consideration of a discrete liquid element yields the Navier-Stokes equations.

Stability of models is introduced by the principle, that a mathematical model must be stable. It is noted, in order to investigate deeply numerical stability, the reader should know the main mathematical approaches to stability should be found in the courses ODE and PDE.

There are many exercises useful for student laboratory and for independent student works in the form of projects. Below, we illustrate the Monte Carlo method to calculate the constant π. Let Ω be the square 2×2 and A be the disk enclosed to the square displayed in Figure. Let n points be randomly thrown on the square. Here, randomness means that each point is represented by its coordinates randomly chosen in the interval (-1, 1). A part of these points m goes onto the disk. For sufficiently large n one can expect that the ratio m/n will be close to the ratio of the areas of the disk to the area of the square $|A|/|\Omega|$. In the considered case, it is equal to $\pi/4$. Experimental computations of numbers n and m yields the value of π. Using this approach one can solve the following two examples from Chapter 5 [1].

Compute the area of the domain bounded by the ellipse $x^2/9+y^2/4 = 1$ *applying the Monte Carlo simulations.*

Compute the area $S(a, b)$ *of the domain bounded by the ellipse* $x^2/a^2+y^2/b^2 = 1$ *applying the Monte Carlo simulations.*

Hint1: Investigate numerically the dependence of $S(a, b)$ *on two parameters, i.e., on a with a fixed b and on b with a fixed a.*

Hint2: An alternative way is based on the formula $S(ka, kb) = k^2 S(a, b)$ *where* $k > 0$ *is a linear extension coefficient. Take* $k = (ab)^{-\frac{1}{2}}$. *Then,* $S(a, b) = ab\ S(d, d^{-1})$ *with* $d = a^{\frac{1}{2}} b^{-\frac{1}{2}}$. *Investigate numerically the dependence of* $S(d, d^{-1})$ *on d.*

Answer: $S(a, b) = \pi ab$.

Compute the surface area $S(a, b, c)$ *of the ellipsoid by fitting the parameter p in the approximate formula*

$$S(a, b, c) \approx 4\pi \left(\frac{a^p b^p + a^p c^p + c^p b^p}{3} \right)^{\frac{1}{p}}$$

Answer: $p \approx 1.6$.

Below, we present an exercise on statistics.

Consider data presented in the form of a set of dimensional numbers. Take only the first digits of these numbers. Generate a statistical distribution of the first digits $\{1, 2, \ldots, 9\}$ following Simon Newcomb (1881) for

1. *populations of countries,*
2. *areas of countries,*
3. *heights of mountains higher than 2 km.*

Investigate the same statistical distributions but in other bases of number systems. Investigate the same statistical distributions of first digits in stock exchanges taking prices of one stock in time.

3 *Mathematica*® Application for Numerical and Symbolic Computations

We have to note that *Mathematica*® contains on-line data on US and other stocks, mutual funds and other financial instruments. The special operators such as FinancialData combined with DateListPlot are used to study market on-line.

The number of exercises and their diversity (calculus, ODE, data fitting etc.) allows to look at the standard mathematical exercises from another computational point of view. It is worth noting that visualization including animation serves as a significant help to understand the considered problem. An electronic appendix with solutions with *Mathematica*® codes is applied.

The textbook establishes the general principles and methods of mathematical modeling. At the beginning a simple mathematical model is considered. The model is explained by hand calculations as well as applying the packages *Mathematica*® and MATLAB. Some numerical and symbolic computations are considered by iterative methods, Newton's method and a method of successive approximations.

In our case, the textbook has been used for students studying on specialty "Mathematics". The manner of explanation simultaneously by hand and a package are helpful for them. After getting a progress and understanding the basics of the package they have been able to develop their skills.

4 Weierstrass Functions and Their Invariants

As an example of such a project we consider the optimal random packing problem of disks on the plane torus, i.e., in a class of doubly periodic packing with three disks per the hexagonal fundamental domain [2]. The Weierstrass ϱ-function and its

derivatives are introduced by the following operators

$$\varrho[z\_] := WeierstrassP[z, g2g3];$$

$$\varrho[0, z\_] := WeierstrassP[z, g2g3];$$

$$\varrho[1, z\_] := WeierstrassPPrime[z, g2g3];$$

$$\varrho 1[z\_] := WeierstrassPPrime[z, g2g3];$$

$$\varrho[2, z\_] := 6(\varrho[0, z])^2 - 30S4/.S4- > 0;$$

$$\varrho[3, z\_] := 12\varrho[0, z]\varrho[1, z];$$

$$\varrho[k\_/; k > 3, z\_] := 12$$

$$\sum_{s=0}^{k} -3Binomial[k-3, s]\varrho[k-s-3, z]\varrho[s+1, z]$$

The roots of the Weierstrass function can be found by the operate NSolve. The optimal packing of three disks on the plane torus can be found by roots of the Weierstrass functions and their combinations [2]. The following function includes geometrical and analytical computations and illustrates the double periodic best packing of three disks

$$HC = Show[Graphics[\{\{Gray, Disk[\{-x[3], -y[3]\}, r],$$

$$Disk[\{\omega 1 - x[3], -y[3]\}, r],$$

$$Disk[\{\omega 1\omega 2 - x[3], \omega 3\omega 1\omega 2 - y[3]\}, r], Disk[\{\omega 1\omega 2 - x[3], \omega 3$$

$$\omega 1\omega 6 - y[3]\}, r], Disk[\{Re[\omega 1 + \omega 2 - x[3]], Im[\omega 1 + \omega 2] - y[3]\}, r],$$

$$Disk[\{\omega 1\omega 2 - Re[((\omega 1 + \omega 2))\omega 6], \omega 3$$

$$\omega 1\omega 6 - Im[((\omega 1 + \omega 2))\omega 6]\}, r]\},$$

$$\{Dashed, Line[\{\{-x[3], -y[3]\}, \{\omega 1 - x[3], -y[3]\},$$

$$\{Re[\omega 1 + \omega 2 - x[3]], Im[\omega 1 + \omega 2] - y[3]\}, \{Re[\omega 2] - x[3], Im[\omega 2] - y[3]\},$$

$$\{-x[3], -y[3]\}\}]\}, Arrow[\{\{-0.6, 0\}, \{0.6, 0\}\}],$$

$$Arrow[\{\{0, -0.6\}, \{0, 0.6\}\}]\}, AspectRatio\omega Automatic]]$$

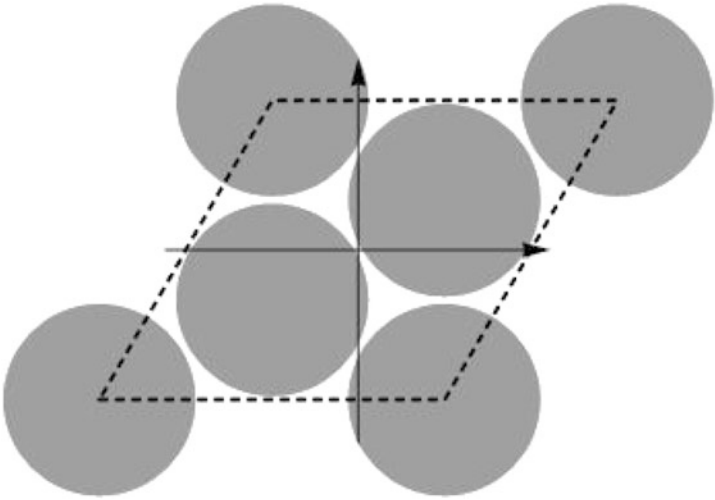


Fig. 1 The double periodic best packing of three disks in the hexagonal cell

The result is displayed in Fig. 1.

Concerning standard courses of high school, it is difficult to say what is hard to compute with the package Mathematica®. A beginners may start to use it at once on an elementary level and further improve her/his skills related to computations. This concerns pupils od secondary school too. Mathematica® contains thousands Explore thousands of free projects across science, engineering, technology, business, art, finance, social sciences of different levels and purposes [3].

Acknowledgments This research, V. Mityushev, Zh. Zhunussova and Ye. Ashimov, is funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP08856381).

References

1. Mityushev, V., Nawalaniec, W., Rylko, N.: Introduction to Mathematical Modeling and Computer Simulations. Taylor & Francis, Boca Raton (2018)
2. Mityushev, V., Rylko, N.: Optimal distribution of the non-overlapping conducting disks. Multiscale Model. Simul. **10**, 180–190 (2012)
3. Wolfram Demonstrations Project. https://demonstrations.wolfram.com/

Part IV
Complex Analysis and Partial Differential Equations

Universality of the Dirichlet Series in the Complex Plane

George Chelidze, George Giorgobiani, and Vaja Tarieladze

Abstract In this note we show that for any complex number s such that $0 < Re(s) \leq 1$ and $Im(s) \neq 0$, the convergent Dirichlet series

$$\sum_n (-1)^{n-1} \frac{1}{n^s},$$

as well as the divergent Dirichlet series

$$\sum_n \frac{1}{n^s}$$

are Riemann-like; i.e., the sum range under the rearrangements of each of these series is the whole complex plane.

1 Introduction

Let X be a Hausdorff topological abelian group. For a series $\sum_n x_n$ in X its *sum range* $\mathrm{SR}(\sum_n x_n)$ is defined as the set of all elements $s \in X$ for which there exist a permutation $\pi : \mathbb{N} \to \mathbb{N}$ such that the rearranged series $\sum_n x_{\pi(n)}$ converges in X and $s = \sum_{n=1}^{\infty} x_{\pi(n)}$. A series in X is *universal* in X if its sum range is the whole X. A series in X is *unconditionally* convergent if for every permutation $\pi : \mathbb{N} \to \mathbb{N}$ the rearranged series $\sum_n x_{\pi(n)}$ converges in X. It is known that for

G. Chelidze
Kutaisi International University, Kutaisi & Muskhelishvili Institute of Computational Mathematics of the Georgian Technical University, Tbilisi, Georgia
e-mail: giorgi.chelidze@kiu.edu.ge

G. Giorgobiani (✉) · V. Tarieladze
Muskhelishvili Institute of Computational Mathematics of the Georgian Technical University, Tbilisi, Georgia
e-mail: giorgobiani.g@gtu.ge; v.tarieladze@gtu.ge

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_14

each unconditionally convergent series $\sum_n x_n$ in X, the set $\mathrm{SR}(\sum_n x_n)$ consists of one element. The famous Riemann's theorem asserts that every convergent but not unconditionally convergent series of real numbers is universal in $\mathbb{R}$. However, in $\mathbb{R}^2$ not every series of this type is universal. As it is noted in [16], applying the Riemann's theorem it's not difficult to construct universal series in $\mathbb{R}^d$, $d = 2, 3, \ldots$.

When X is an infinite-dimensional normed space, the sum range has more diverse structure [11] (see also [5]). In the Banach space $X = C([0, 1])$ of all continuous real valued functions the existence of an universal series was first proved in [8]. Using the similar approach, this result was extended for an arbitrary infinite-dimensional separable Banach space in [15] and the existence of universal series with some additional properties was established in [7]. In [12] (cf. [3]) sufficient condition for the universality of the series in $L_2[0, 1]$ was included and a specific example of such a series was constructed as well.

In this note we consider the problem of universality of the signed Dirichlet series

$$\sum_n \theta_n \frac{1}{n^s}, \quad \theta_n = \pm 1, \ n = 1, 2, \ldots \tag{1}$$

for a fixed $s \in \mathbb{C}$ with $0 < Re(s) \leq 1$, $Im(s) \neq 0$.

We derive the universality of (1) in $\mathbb{C}$ in case when the series converges; in particular we get the universality of alternatively signed Dirichlet series

$$\sum_n (-1)^{n-1} \frac{1}{n^s}. \tag{2}$$

Recall that (2) converges for any $s \in \mathbb{C}, 0 < Re(s) < \infty$, and its limit is known as the Dirichlet η-function or the alternating ζ-function.

We also show that, rather unexpectedly, the divergent Dirichlet series

$$\sum_n \frac{1}{n^s} \tag{3}$$

is also universal in $\mathbb{C}$. Below we use the following characterizations of universal complex series.

Proposition 1 (cf. [6, Theorem 1.4]) *For a series of complex numbers $\sum_n z_n$ the following statements are equivalent:*

(*i*) $\mathrm{SR}(\sum_n z_n) = \mathbb{C}$.
(*ii*) $\mathrm{SR}(\sum_n z_n) \neq \emptyset$ *and*

$$\sum_{n=1}^{\infty} |Re(w z_n)| = \infty \quad \forall w \in \mathbb{C} \setminus \{0\}. \tag{4}$$

Proposition 2 (cf. [9, Theorem III]) *For a series of complex numbers* $\sum_n z_n$ *the following statements are equivalent:*

(i) $\mathrm{SR}(\sum_n z_n) = \mathbb{C}$.
(ii) $\lim_n z_n = 0$ *and*

$$\sum_{n=1}^{\infty} \max(0, Re(wz_n)) = \infty \quad \forall w \in \mathbb{C} \setminus \{0\} \tag{5}$$

Remark 1 Using the Proposition 1, it can be shown that for any fixed $z \in \mathbb{C}$ with $|z| = 1$ and $z \notin \{-1, 1\}$, we have: $\mathrm{SR}(\sum_n \frac{z^n}{n}) = \mathbb{C}$ (cf. [6]).

Remark 2 The result given in the previous remark fails for the division ring of quaternions $\mathbb{H}$: for any fixed $z \in \mathbb{H}$, $|z| = 1$, $z \notin \{-1, 1\}$, we have: $\mathrm{SR}(\sum_n \frac{z^n}{n}) \neq \mathbb{H}$ (cf. [2]).

Note finally that an analog of Proposition 1 fails if we replace $\mathbb{C}$ by an infinite-dimensional Hilbert space ([5], pg. 513). More general Proposition 2 (which is not mentioned neither in [14] nor in [11]) can be derived also from its Banach-space version contained in [13, Corollary 2].

2 Auxiliary Statements

In what follows, for a fixed $s \in \mathbb{C}$, let $\mathfrak{D}_s$ be the set of all sequences $(\xi_n)_{n\in\mathbb{N}}$ of complex numbers such that the series $\sum_n \xi_n n^{-s}$ converges in $\mathbb{C}$. It can be seen that $\mathfrak{D}_s$ is always a dense vector subspace and a Borel subset of $\mathbb{C}^{\mathbb{N}}$.

We recall the following particular case of Jensen-Cahen's theorem:

Lemma 1 *Let* $s \in \mathbb{C}$ *and* $Re(s) > 0$. *Then for any sequence* $(\xi_n)_{n\in\mathbb{N}} \in \mathbb{C}^{\mathbb{N}}$ *such that*

$$\sup_{n\in\mathbb{N}} |\sum_{k=1}^{n} \xi_k| < \infty, \tag{6}$$

Dirichlet series

$$\sum_n \xi_n n^{-s}$$

converges in $\mathbb{C}$, *i. e. we have*

$$(\xi_n)_{n\in\mathbb{N}} \in \mathfrak{D}_s. \tag{7}$$

In particular, $(\theta_n)_{n\in\mathbb{N}} \in \mathfrak{D}_s$, *where* $(\theta_n)_{n\in\mathbb{N}}$ *is the* **alternating sign sequence** $\theta_n = (-1)^{n-1}$, $n = 1, 2, \ldots$

Remark 3 In notations of Lemma 1

(a) In case $Re(s) > 1$, even from the weaker condition $\sup_{n\in\mathbb{N}} |\xi_n| < \infty$, we can conclude that the stronger conclusion

$$\sum_{n=1}^{\infty} |\xi_n n^{-s}| < \infty$$

holds instead of (7).

(b) In case $0 < Re(s) \leq 1$, the condition (6) implies (7) due to the observations

$$\lim_n n^{-s} = 0 \quad \text{and} \quad \sum_{n=1}^{\infty} |n^{-s} - (n+1)^{-s}| < \infty \tag{8}$$

and by use of Abel's identity (summation by parts; cf. [18, Theorem 1.2.4, (p.3)]). The second relation in (8) can be proved using the following inequality ([10, Ch. II.1, Lemma 2 (p.3)]):

$$|n^{-s} - (n+1)^{-s}| \leq \frac{|s|}{\sigma}(n^{-\sigma} - (n+1)^{-\sigma}), \ n = 1, 2, \ldots, \tag{9}$$

where $\sigma = Re(s)$. An "integral-free" proof of (9) is also possible.

Remark 4 In connection with Lemma 1 note that not only for the sequence n^{-s}, $n = 1, 2, \ldots$ with $s \in \mathbb{C}$, $Re(s) > 0$, but for *any* $(z_n)_{n\in\mathbb{N}} \in \mathbb{C}^{\mathbb{N}}$ *such that* $\lim_n z_n = 0$, the existence of the sign sequence $\theta_n \in \{1, -1\}$, $n = 1, 2, \ldots$ making the series $\sum_n \theta_n z_n$ convergent is guaranteed by the Dvoretzky-Hanani theorem [4] (or [11]).

Note also that:

(I) In terms of the theory of Dirichlet series Lemma 1 asserts that if a sequence $(\xi_n)_{n\in\mathbb{N}} \in \mathbb{C}^{\mathbb{N}}$ satisfies condition (6), then the abscissa of convergence of the Dirichlet series $\sum_n \xi_n n^{-s}$ equals zero.

(II) If $z \in \mathbb{C}$ and $0 < Re(z) < 1$, then

$$\sup_{n\in\mathbb{N}} |\sum_{k=1}^{n} \frac{1}{k^z}| = \infty. \tag{10}$$

In particular, Dirichlet series $\sum_n \frac{1}{n^z}$ does not converge. (In fact, suppose that

$$\sup_{n\in\mathbb{N}} |\sum_{k=1}^{n} \frac{1}{k^z}| < \infty. \tag{11}$$

Then taking in Lemma 1 $s = 1 - z$ and $\xi_n = \frac{1}{n^z}$, $n = 1, 2, \ldots$, we derive the convergence of $\sum_n \frac{1}{n}$.)

(III) If $t \in \mathbb{R} \setminus \{0\}$, then

$$\sup_{n\in\mathbb{N}} |\sum_{k=1}^{n} \frac{1}{k^{1+it}}| < \infty,$$

but again, Dirichlet series $\sum_n \frac{1}{n^{1+it}}$ diverges (cf. [10, Ch. II.1 (p.5)]; see also [1, (p. 247)]).

(IV) If $\frac{1}{2} < Re(s) \leq 1$, then $\mu(\mathfrak{D}_s \bigcap \{-1, 1\}^{\mathbb{N}}) = 1$, where μ stands for the canonical product probability measure on $\{-1, 1\}^{\mathbb{N}}$ (this follows from Rademacher theorem; cf. [17]).

Lemma 2 *Let $\varphi \in \mathbb{R}$, $t \in \mathbb{R} \setminus \{0\}$ and $0 < \sigma \leq 1$. Then*

$$\sum_{n=1}^{\infty} n^{-\sigma} \max(0, \cos(\varphi - t \ln n)) = \infty\,. \tag{12}$$

In particular,

$$\sum_{n=1}^{\infty} n^{-\sigma} |\cos(\varphi - t \ln n)| = \infty\,. \tag{13}$$

Proof Assume the contrary

$$\sum_{n=1}^{\infty} n^{-\sigma} \max(0, \cos(\varphi - t \ln n)) < \infty\,. \tag{14}$$

Fix $k \in \mathbb{N}$ and write

$$a_k = e^{\frac{1}{t}(2\pi k \frac{t}{|t|} + \varphi - \frac{\pi}{3}\frac{t}{|t|})},$$

$$b_k = e^{\frac{1}{t}(2\pi k \frac{t}{|t|} + \varphi + \frac{\pi}{3}\frac{t}{|t|})}.$$

Clearly $a_k < b_k < a_{k+1}$. Set:

$$\Lambda_k = \mathbb{N} \cap]a_k, b_k[\,.$$

As

$$\lim_k \min\{\Lambda_k\} = \infty\,,$$

from (14) we get

$$\lim_k \sum_{n\in\Lambda_k} n^{-\sigma} \max(0, \cos(\varphi - t\ln n)) = 0\,. \tag{15}$$

Let us see that (15) leads to a contradiction.

Choose an integer k_0 so that $a_k > 1$ and $b_k > a_k + 3$, for any $k > k_0$. Fix now $k \in \mathbb{N}$ with $k > k_0$. Observe that

$$\mathrm{card}(\Lambda_k) \geq b_k - a_k - 2 > 1\,.$$

As

$$n \in \Lambda_k \implies \cos(\varphi - t\ln n) \geq \frac{1}{2},$$

we can write

$$\sum_{n\in\Lambda_k} n^{-\sigma} \max(0, \cos(\varphi - t\ln n)) \geq \frac{1}{2}\sum_{n\in\Lambda_k} \frac{1}{n^\sigma} > \frac{1}{2}\sum_{n\in\Lambda_k} \frac{1}{n} \geq$$

$$\geq \frac{1}{2b_k}\mathrm{card}(\Lambda_k) \geq \frac{b_k - a_k - 2}{2b_k} = \frac{1}{2}(1 - e^{-\frac{2\pi}{3|t|}} - \frac{2}{b_k})\,.$$

So,

$$\sum_{n\in\Lambda_k} n^{-\sigma} \max(0, \cos(\varphi - t\ln n)) > \frac{1}{2}(1 - e^{-\frac{2\pi}{3|t|}} - \frac{2}{a_k + 3})\,. \tag{16}$$

From (16), as $\lim_k a_k = \infty$ we get:

$$\liminf_k \sum_{n\in\Lambda_k} n^{-\sigma} \max(0, \cos(\varphi - t\ln n)) \geq 1 - e^{-\frac{2\pi}{3|t|}} > 0\,. \tag{17}$$

Relation (17) contradicts (15). Hence, (14) doesn't hold and the lemma is proved. □

It would be interesting to describe the sequences (c_n) for which the following analogue of Lemma 2 is true:

$$c_n > 0,\ \sum_{n=1}^{\infty} c_n = \infty \implies \sum_{n=1}^{\infty} c_n \max(0, \cos(\ln n)) = \infty\,. \tag{18}$$

It can be shown that (18) is not true in general.

3 Universality Theorems

Our first universality result looks as follows.

Theorem 1 *Let s be a complex number with $0 < Re(s) \le 1$, $Im(s) \neq 0$, and let*

$$(\theta_n)_{n\in\mathbb{N}} \in \mathfrak{D}_s \bigcap \{-1, 1\}^{\mathbb{N}}. \tag{19}$$

Then $\mathrm{SR}(\sum_n \theta_n n^{-s}) = \mathbb{C}$*; i.e., the series (1) is universal in* $\mathbb{C}$.

Proof It suffices to show that the condition (ii) of Proposition 1 holds for $z_n = \theta_n n^{-s}, n = 1, 2, \ldots$. From (19) we have that $\mathrm{SR}(\sum_n z_n) \neq \emptyset$. So, it remains to show that

$$\sum_{n=1}^{\infty} |Re\left(wn^{-s}\right)| = \infty \quad \forall w \in \mathbb{C} \setminus \{0\}. \tag{20}$$

Fix $w \in \mathbb{C} \setminus \{0\}$ and write $w = re^{i\varphi}$ with some $r > 0$ and $\varphi \in \mathbb{R}$. Set also

$$\sigma := Re(s) \quad \text{and} \quad t := Im(s)\,.$$

Then we have

$$|Re\left(wn^{-s}\right)| = \frac{r}{n^{\sigma}} |\cos(\varphi - t\ln n)|, \;\; n = 1, 2, \ldots$$

and (20) is true by equality (13) of Lemma 2. □

It can be seen that Theorem 1 may fail without the assumption (19). Nevertheless, it is remarkable that the following statement is true.

Theorem 2 *Let s be a complex number such that $0 < Re(s) \le 1$ and $Im(s) \neq 0$. Then* $\mathrm{SR}(\sum_n n^{-s}) = \mathbb{C}$.

Proof It suffices to show that the condition (ii) of Proposition 2 holds for $z_n = n^{-s}, n = 1, 2, \ldots$. Clearly, $\lim_n z_n = 0$. So, it remains to show that

$$\sum_{n=1}^{\infty} \max(0, Re(wz_n)) = \infty \quad \forall w \in \mathbb{C} \setminus \{0\} \tag{21}$$

Fix $w \in \mathbb{C} \setminus \{0\}$ and write $w = re^{i\varphi}$ with some $r > 0$ and $\varphi \in \mathbb{R}$. Set also

$$\sigma := Re(s) \quad \text{and} \quad t := Im(s)\,.$$

Then we have

$$Re\left(wn^{-s}\right) = \frac{r}{n^{\sigma}}\cos(\varphi - t\ln n),\ \ n = 1, 2, \ldots$$

and (21) is true by equality (12) of Lemma 2. □

Note finally that in connection with Theorem 1 and Theorem 2 it would be interesting to know for which sequences $(\xi_n)_{n\in\mathbb{N}}$ of real or complex numbers the equality $\mathrm{SR}(\sum_n \xi_n n^{-s}) = \mathbb{C}$ holds.

References

1. Apostol, T.M.: Introduction to Analytic Number Theory. Springer, New York (1976)
2. Chelidze, G.: Giorgobiani, G.: Tarieladze, V.: Sum range of quaternion series. J. Math. Sci. **216**(4), 519–521 (2016)
3. Drobot, V.: Rearrangements of series of functions. Trans. Am. Math. Soc. **142**, 239–248 (1969)
4. Dvoretzky, A., Chojnacki, C.: Sur les changements des signes des termes d'une série á termes complexes. C.R. Acad. Sci. Paris **255**, 516–518 (1947)
5. Giorgobiani, G.: Rearrangements of series. J. Math. Sci. **239**, 437–548 (2019). https://doi.org/10.1007/s10958-019-04315-9
6. Giorgobiani, G., Tarieladze, V.: On complex universal series. Proc. A. Razmadze Math. Inst. **160**, 53–63 (2012)
7. Giorgobiani, G., Tarieladze, V.: Special universal series. In: I. Gorgidze et al. (Eds.), Several Problems of Applied Mathematics and Mechanics, pp. 125–130. Nova Science Publishers; Mathematics Research Developments, New York (2013)
8. Hadwiger, H.: Eine Bemerkung uber Umordnung von Reihen reeller Funktionen (in German). Tohoku Math. J. **46**, 22–25 (1939)
9. Halperin, I.: Sums of a series, permitting rearrangements. C.R. Math. Rep. Acad. Sci. Can. **8**(2), 87–102 (1986)
10. Hardy, G.H., Riesz, M.: The General Theory of Dirichlet's Series. Courier Corporation, Chelmsford (2013)
11. Kadets, M.I., Kadets, V.M.: Series in Banach Spaces. Birkhauser Verlag, Basel (1997)
12. Kashin, B.S., Saakyan, A.A.: Orthogonal series (in Russian). Nauka, Moscow (1984); translation of Mathematical Monographs, vol. 75, Amer. Math. Soc., Copyright, 2005
13. Pecherskii, D.V.: Rearrangements of series in Banach spaces and arrangements of signs. Matematicheskii Sbornik **177**(1), 24–35 (1988). English translation: Pecherskii, D. V. "Rearrangements of series in Banach spaces and arrangements of signs." Mathematics of the USSR-Sbornik 63.1 (1989): 23
14. Rosenthal, P.: The remarkable theorem of Levy and Steinitz. Am. Math. Month. **94**(4), 342–351 (1987)
15. Shklyarski, D.O.: Conditionally convergent series of vectors (in Russian). Uspehi Matem. Nauk **10**, 51–59 (1944)
16. Steinitz, E.: Bedingt convergente Reichen konvexe Systeme. (Teil I.) (in German). J. für Math. **143**, 128–175 (1913)
17. Vakhania, N.N., Tarieladze, V.I., Chobanyan, S.A.: Probability Distributions in Banach Spaces. D. Reidel Publishing, Dordreht (1987)
18. Zygmund, A.: Trigonometric Series, vol. 1. Cambridge University Press, Cambridge (2002)

On One Oscillation Problem of Zeroth Approximation of Hierarchical Model for Porous Elastic Plates with Variable Thickness

Natalia Chinchaladze

Abstract The aim of the paper is to investigate of homogeneous Dirichlet problem for the vibration problem of porous elastic prismatic shell-like bodies within the framework of known models of mathematical problems arising in connection of complicated geometry of bodies under consideration (see [18]). We consider cusped bodies for them BVPs and IBVPs are non-classical in general. The classical and weak setting of the problem are formulated. The weighted Sobolev spaces X^{κ} are introduced, which are crucial in our analysis. The coerciveness of the corresponding bilinear form is shown and uniqueness and existence results for the variational problem are proved.

1 Introduction

The development of science, industry and technologies on the one hand made the possibility of constructing such new composite materials with different physical properties (piezoelectric, piezomagnetic, multi-component mixtures, bio-materials, meta-materials etc.) that are not found naturally on Earth. On the other hand these new materials can be used for future development of the same fields. Several examples include piezoelectric sensors for vibration control [17], high precision actuators [1], materials with higher strength and stiffness [11] or ones that lower energy consumption [4, 16], production cost and size of sensors or actuators [1, 17].

In 1955 Ilia Vekua [18] published his models of elastic prismatic shells. In 1965 he offered analogous models for standard shells [19]. In both papers he considered a very important investigation of well-posedeness of boundary value problems (BVPs) of peculiar types which could arise in the case of cusped shells. Using I. Vekua dimension reduction method, complexity of the 3D domain, occupied by the body will be transformed into the degeneracy of the order of the 2D governing

N. Chinchaladze (✉)
Iv. Javakhishvili Tbilisi State University, I. Vekua Institute of Applied Mathematics & Faculty of Exact and Natural Sciences, Tbilisi, Georgia
e-mail: natalia.chinchaladze@tsu.ge

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_15

equations of the constructed hierarchy of 2D models on the boundary of the 2D projection of the 3D bodies under consideration.

Jaiani [9] is devoted to construction of hierarchical models for piezoelectric nonhomogeneous porous elastic and viscoelastic Kelvin-Voigt prismatic shells on the basis of linear theories [3, 6, 12, 14, 15]. Using I. Vekua [18] (see also [19]) dimension reduction method, governing systems are derived and in the Nth approximation of hierarchical models BVPs and IBVPs are set. In the $N = 0$ approximation, statical problem is investigated. The aim of this paper is to study analogous problem in the case of oscillation.

2 Field Equations for Kelvin-Voigt Materials

A Kelvin-Voigt material, also called a Voigt material, is a viscoelastic material having the properties both of elasticity and viscosity. The theories of viscoelasticity, which include the Maxwell model, the Kelvin-Voigt model, and the Standard Linear Solid model, are used to predict a material's response under different loading conditions. One of the simplest mathematical models constructed to describe the viscoelastic effects is the classical Kelvin-Voigt model (see Eringen [5]). The basic idea concerning this model is that the stress is dependent on the deformation tensor and deformation-rate tensor. This model consists of a Newtonian damper and Hooke elastic spring connected in parallel.

The field equations have the following form [6, 14]:

Motion Equations

$$X_{ji,j} + \Phi_i = \rho \ddot{u}_i(x_1, x_2, x_3, t), \quad (x_1, x_2, x_3) \in \Omega \subset \mathbb{R}^3, \quad t > t_0, \quad i, j = 1, 2, 3;$$

$$H_{j,j} + H_0 = \rho_0 \ddot{\varphi} - \mathcal{F},$$

where $X_{ij} \in C^1(\Omega)$ is the stress tensor; Φ_i are the volume force components; $\rho_0 := \rho k'$ (k' is equilibrated inertia), ρ is the reference mass density; $u_i \in C^2(\Omega)$ are the displacements; $H_j \in C^1(\Omega)$ is the component of the equilibrated stress vector, H_0 and F are the intrinsic and extrinsic equilibrated volume forces; Einstein's summation convention is used; indices after comma mean differentiation with respect to the corresponding variables of the Cartesian frame $Ox_1x_2x_3$ (throughout the paper we assume existence of the indicated (continuous) derivatives); dots as superscripts of the symbols mean derivatives with respect to time t.

Constitutive Equations (Isotropic Case)

$$X_{ij} = \lambda e_{kk}\delta_{ij} + 2\mu e_{ij} + \lambda^* \dot{e}_{kk}\delta_{ij} + 2\mu^* \dot{e}_{ij} + b\varphi\delta_{ij} + b^*\dot{\varphi}\delta_{ij}, \quad i, j = 1, 2, 3,$$
$$H_j = \tilde{\alpha}\varphi_{,j} + \alpha^*\dot{\varphi}_{,j}, \qquad j = 1, 2, 3,$$
$$H_0 = -be_{kk} - \xi\varphi - \nu^*\dot{e}_{kk} - \xi^*\dot{\varphi},$$

where $e_{ij} \in C^1(\Omega)$ is the strain tensor; $\varphi := \nu_0 - \nu \in C^2(\Omega)$ is the change in the volume fraction from the matrix reference volume fraction ν (clearly, the bulk reference density $\rho = \nu\gamma$, $0 < \nu \leq 1$, here γ is the matrix reference density); λ, λ^*, μ, μ^*, b, b^*, $\tilde{\alpha}$, α^*, ν^*, ξ, ξ^* are the constitutive coefficients, depending on x_1 and x_2;

Kinematic Relations

$$e_{ij} = \frac{1}{2}(u_{i,j} + u_{j,i}), \quad i, j = 1, 2, 3.$$

3 $N = 0$ Approximation for Porous Elastic Prismatic Shell-like Bodies

We consider prismatic shell-like bodies of Kelvin-Voigt material (see, e.g., [7]) which occupies 3D domain Ω with the projection ω (on the plane $x_3 = 0$) and the face surfaces

$$x_3 = \overset{(+)}{h}(x_1, x_2) \in C^2(\omega) \text{ and } x_3 = \overset{(-)}{h}(x_1, x_2) \in C^2(\omega), \quad (x_1, x_2) \in \omega,$$

where

$$2h(x_1, x_2) := \overset{(+)}{h}(x_1, x_2) - \overset{(-)}{h}(x_1, x_2), \quad (x_1, x_2) \in \omega,$$

is the thickness of the prismatic shell. Prismatic shells are called cusped shells if a set γ_0, consisting of $(x_1, x_2) \in \partial\omega$ for which $2h(x_1, x_2) = 0$, is not empty. If

$$\overset{(+)}{h}(x_1, x_2) = -\overset{(-)}{h}(x_1, x_2)$$

we have to do with plates of variable thickness.

Vekua's hierarchical models for elastic shells are the mathematical models (see [2, 7, 20]). Their construction is based on the multiplication of the basic equations of

linear elasticity by Legendre polynomials $P_r(ax_3 - b)$, $a := \frac{1}{h(x_1,x_2)}$), $b := \frac{\overset{(-)}{h} + \overset{(-)}{h}}{\overset{(+)}{h} - \overset{(-)}{h}}$, and then integration with respect to x_3 within the limits $\overset{(-)}{h}$ and $\overset{(+)}{h}$. By constructing Vekua's hierarchical models in first version on upper and lower surfaces stress-vectors are assumed to be known, while there the values of the displacements are calculated from their Fourier-Legendre series expansions on the segment $x_3 \in \left[\overset{(-)}{h}, \overset{(+)}{h}\right]$. Finally (see, e.g. [7]) we construct an equivalent infinite system with respect to the rth order moments u_{ir}, φ_r. After this, if we suppose that the moments whose subscripts, indicating moments' order, are greater than N equal zero and consider only the first $N + 1$ equations ($r = \overline{0, N}$) in the obtained infinite system of equations with respect to the r-th order moments u_{ir} and φ_r we obtain the $N-$th order approximation (hierarchical model) governing system consisting of $4N + 4$ equations with respect to $4N + 4$ unknown functions $\overset{N}{u}_{ir}$, $\overset{N}{\varphi}_r$ (roughly speaking $\overset{N}{u}_{ir}$, $\overset{N}{\varphi}_r$ is an "approximate value" of u_{ir}, φ_r, since $\overset{N}{u}_{ir}$, $\overset{N}{\varphi}_r$ are solutions of the derived finite system), $i = \overline{1, 3}$, $r = \overline{0, N}$.

In $N = 0$ approximation for viscoelastic Kelvin-Voigt prismatic shells the governing system has the following form (see [9]):

$$(\mu h v_{\alpha 0,\beta})_{,\alpha} + (\mu h v_{\beta 0,\alpha})_{,\alpha} + (\lambda h v_{\gamma 0,\gamma})_{,\beta} + (\mu^* h \dot{v}_{\alpha 0,\beta})_{,\alpha} + (\mu^* h \dot{v}_{\beta 0,\alpha})_{,\alpha}$$
$$+(\lambda^* h \dot{v}_{\gamma 0,\gamma})_{,\beta} + (bh\psi_0)_{,\beta} + (b^* h \dot{\psi}_0)_{,\beta} + \overset{0}{X}_\beta = \rho h \ddot{v}_{\beta 0}, \quad \beta = 1, 2; \tag{1}$$

$$(\mu h v_{30,\alpha})_{,\alpha} + (\mu^* h \dot{v}_{30,\alpha})_{,\alpha} + \overset{0}{X}_3 = \rho h \ddot{v}_{30}; \tag{2}$$

$$(\tilde{\alpha} h \psi_{0,\alpha})_{,\alpha} - bh v_{\gamma 0,\gamma} - \xi h \psi_0 + (\alpha^* h \dot{\psi}_{0,\alpha})_{,\alpha} - \nu^* h \dot{v}_{\gamma 0,\gamma} - \xi^* h \dot{\psi}_0 + \overset{0}{H}$$
$$= \rho h \ddot{\psi}_0 - \mathcal{F}_0, \tag{3}$$

where

$$v_{k0} := \frac{\overset{0}{u}_{i0}}{h}, \quad k = 1, 2, 3, \quad \psi_0 := \frac{\overset{0}{\varphi}_0}{h},$$

are so called the weighted displacements and the weighted volume fraction.

Note that, if we take

$$\lambda^* = 0, \quad \mu^* = 0, \quad b^* = 0, \quad \alpha^* = 0, \quad \nu^* = 0, \quad \xi^* = 0,$$

from the above obtained governing system, we get hierarchical models for porous elastic prismatic shells.

In case of $N = 0$ approximation for porous elastic prismatic shells, from (1)–(3), we get the following governing system

$$(\mu h v_{\alpha 0,\beta})_{,\alpha} + (\mu h v_{\beta 0,\alpha})_{,\alpha} + (\lambda h v_{\gamma 0,\gamma})_{,\beta} + (bh\psi_0)_{,\beta} + \overset{0}{X}_{\beta} = \rho h \ddot{v}_{\beta 0}, \tag{4}$$

$$(\mu h v_{30,\alpha})_{,\alpha} + \overset{0}{X}_3 = \rho h \ddot{v}_{30}; \tag{5}$$

$$(\tilde{\alpha} h \psi_{0,\alpha})_{,\alpha} - b h v_{\gamma 0,\gamma} - \xi h \psi_0 + \overset{0}{H} = \rho \ddot{\varphi}_0 - \mathcal{F}_0. \tag{6}$$

Dirichlet problem in the classical form looks like: find 4-dimensional vector

$$v = (v_{10}, v_{20}, v_{30}, \psi_0)^T,$$

in ω satisfying system (4)–(6) and the homogeneous Dirichlet boundary conditions

$$v_{i0} = 0, \quad i = 1, 2, 3; \quad \psi_0 = 0 \quad \text{on} \quad \partial\omega$$

Let denote by γ_0

$$\gamma_0 := \Big\{(x_1, x_2) \in \partial\omega \,:\, 2h(x_1, x_2) = 0\Big\}.$$

BCs for the weighted displacements and the weighted volume fraction are non-classical in the case of cusped prismatic shells. Namely, we are not always able to prescribe them at cusped edges.

We consider the body whose thickness is given by the following expression

$$2h(x_1, x_2) = h_0 x_2^{\kappa}, \quad x_2 \in [0, l] \quad h_0,\ \kappa,\ l = const > 0. \tag{7}$$

For the sake of simplicity we assume that

$$v_{\alpha 0} \equiv 0, \quad \alpha = 1, 2; \quad v_{30} \not\equiv 0.$$

In the case of harmonic vibration, taking into account (7), from (5), (6) we get

$$x_2 \overset{\circ}{v}_{30,\alpha\alpha} + \kappa \overset{\circ}{v}_{30,2} - 2\mu^{-1}\rho\vartheta \overset{\circ}{v}_{30} = 2(\mu h_0)^{-1} x_2^{1-\kappa} \overset{\circ}{X}_3, \tag{8}$$

$$x_2 \overset{\circ}{\psi}_{0,\alpha\alpha} + \kappa \overset{\circ}{\psi}_{0,2} - (\xi - \rho\vartheta^2)\tilde{\alpha} x_2 \overset{\circ}{\psi}_0 = -2(\tilde{\alpha} h_0)^{-1} x_2^{1-\kappa} F, \tag{9}$$

where

$$\overset{\circ}{H} + \mathcal{F}_0 = e^{\iota\vartheta t} F(x_1, x_2), \quad \overset{\circ}{X}_3 = e^{\iota\vartheta t} \overset{\circ}{X^0}_3(x_1, x_2),$$

$$v_{30} = e^{\iota\vartheta t} \overset{\circ}{v}_{30}(x_1, x_2), \quad \psi_0 = e^{\iota\vartheta t} \overset{\circ}{\psi}_0(x_1, x_2).$$

From the following theorem (see G. Jaiani, *On a generalization of the Keldysh theorem*, Georgian Mathematical Journal, **2**, 3 (1995), 291–297).

Theorem 1 *If the coefficients a_α, $\alpha = 1, 2$, and c of the equation*

$$x_2^{\kappa_\alpha} u_{,\alpha\alpha} + a_\alpha(x_1, x_2) u_{,\alpha} + c(x_1, x_2) u = 0, \quad c \le 0, \quad \kappa_\alpha = const \ge 0, \quad \alpha = 1, 2,$$

are analytic in $\overline{\omega}$, then

(i) if either $\kappa_2 < 1$, or $\kappa_2 \ge 1$,

$$a_2(x_1, x_2) < x_2^{\kappa_2 - 1} \tag{10}$$

in $\overline{\omega}_\delta$ for some $\delta = const > 0$, where

$$\omega_\delta := \{(x_1, x_2) \in \omega \ : \ 0 < x_2 < \delta\},$$

the Dirichlet problem (find $v_{30}, \psi_0 \in C^2(\omega) \cap C(\bar{\omega})$ by their values prescribed on $\partial\omega$) is well-posed;

(ii) if $\kappa_2 \ge 1$,

$$a_2(x_1, x_2) \ge x_2^{\kappa_2 - 1} \tag{11}$$

in ω_δ and $a_1(x_1, x_2) = O(x_2^{\kappa_1})$, $x_2 \to 0_+$ (O is the Landau symbol), the Keldysh problem ((Find bounded $v_{30}, \psi_0 \in C^2(\omega) \cap C(\omega \cup (\partial\omega \setminus \overline{\gamma}_0))$ by their values prescribed only on the arc $\partial\omega \setminus \overline{\gamma}_0$)) is well-posed.

it follows

Theorem 2 *If*

$$\xi - \rho\vartheta^2 \ge 0, \tag{12}$$

(i) $\kappa < 1$, the Dirichlet problem is well-posed;

(ii) if $\kappa \ge 1$, the Keldysh problem is well-posed.

Because of Eq. (8) mathematically coincide to the equation of $N = 0$ approximation of the classical linear theory of the elasticity (see [8]) we consider Eq. (9), which in our case can be rewritten as follows

$$-\left(h\overset{\circ}{\psi}_{0,\alpha}\right),_{\alpha} -(\rho\vartheta^2 - \xi)\tilde{\alpha}^{-1}h\overset{\circ}{\psi}_0 = \tilde{\alpha}^{-1}F, \tag{13}$$

$$\overset{\circ}{\psi}_0 = 0, \quad \text{on} \quad \partial\omega. \tag{14}$$

Let

$$\overset{\circ}{\psi}_0,\ \overset{\circ}{\psi}{}^*_0 \in C^2(\omega)\cap C^1(\bar{\omega}), \quad F \in C(\bar{\omega})$$

Using Green's formula and homogeneous BC we get

$$J(\overset{\circ}{\psi}_0, \overset{\circ}{\psi}{}^*_0) = \int\limits_{\omega} F\cdot\overset{\circ}{\psi}{}^*_0\, d\omega$$

$$J(\overset{\circ}{\psi}_0, \overset{\circ}{\psi}{}^*_0) = \int\limits_{\omega}\left[h\overset{\circ}{\psi}_{0,\alpha}(\overset{\circ}{\psi}{}^*_0),_{\alpha} -(\rho\vartheta^2 - \xi)\tilde{\alpha}^{-1}h\overset{\circ}{\psi}_0\overset{\circ}{\psi}{}^*_0\right] d\omega \tag{15}$$

Denote by $\mathcal{D}(\omega)$ a space of infinitely differentiable functions with compact support in ω. We introduce the bilinear form and norm by the following formulas:

$$(\overset{\circ}{\psi}_0, \overset{\circ}{\psi}{}^*_0)_{X^\kappa} := \int_{\omega} x_2^{\kappa}\left[\overset{\circ}{\psi}_{0,1}\overset{\circ}{\psi}{}^*_{0,1} + \overset{\circ}{\psi}_{0,2}\overset{\circ}{\psi}{}^*_{0,2}\right]d\omega$$

and

$$\|\overset{\circ}{\psi}_0\|^2_{X^\kappa} := \int_{\omega} x_2^{\kappa}\left[\overset{\circ}{\psi}{}^2_{0,1} + \overset{\circ}{\psi}{}^2_{0,2}\right]d\omega.$$

The last is the norm because of the well-known Hardy-type inequality (see [13], p. 69, [10]). So, X^κ is a Hilbert space.

The classical and weak setting of the homogeneous Dirichlet problem can be formulated as follows:

Problem 1 Find $v \in C^2(\omega)\cap C^1(\overline{\omega})$ satisfying Eq. (13) in ω and the homogeneous Dirichlet boundary condition (14).

Problem 2 Find $v \in X^\kappa$ satisfying the equality

$$J(\overset{\circ}{\psi}_0, \overset{\circ}{\psi}{}^*_0) = \langle F, \overset{\circ}{\psi}{}^*_0\rangle \quad \text{for all} \quad \overset{\circ}{\psi}{}^*_0 \in X^\kappa, \tag{16}$$

here F belongs to the adjoint space $[X^{\kappa}]^*$, and $\langle\cdot,\cdot\rangle$ denotes duality brackets between the spaces $[X^{\kappa}]^*$ and X^{κ}.

Lemma 1 *The bilinear form $J(\cdot,\cdot)$ is bounded and strictly coercive in the space $X^{\kappa}(\omega)$, i.e., there are positive constant C_0 and C_1 such that*

$$|J(\overset{\circ}{\psi}_0, \overset{\circ}{\psi}{}^*_0)| \le C_1 \|\overset{\circ}{\psi}_0\|_{X^{\kappa}} \|\overset{\circ}{\psi}{}^*_0\|_{X^{\kappa}}, \tag{17}$$

$$J(\overset{\circ}{\psi}_0, \overset{\circ}{\psi}_0) \ge C_0 \|\overset{\circ}{\psi}_0\|^2_{X^{\kappa}} \tag{18}$$

for all $\overset{\circ}{\psi}_0, \overset{\circ}{\psi}{}^*_0 \in X^{\kappa}$, because of (12).

Proof Equation (18) follows from (15) and Hardy's inequality (see [13], p. 69, [10])

$$J(\overset{\circ}{\psi}_0, \overset{0}{\psi}_0) = \int_{\omega} \Big[h \overset{\circ}{\psi}_{0,\alpha} \overset{\circ}{\psi}_{0,\alpha} - (\rho\theta^2 - \xi)\tilde{\alpha}^{-1} h \overset{\circ}{\psi}_0 \overset{\circ}{\psi}_0 \Big] d\omega$$

$$\ge \int_{\omega} h \Big[\overset{\circ}{\psi}_{0,\alpha} \overset{\circ}{\psi}_{0,\alpha} - 4(\rho\theta^2 - \xi)\tilde{\alpha}^{-1} x_2^2 \overset{\circ}{\psi}_{0,2} \overset{\circ}{\psi}_{0,2} \Big] d\omega$$

$$\ge \int_{\omega} h \Big[\overset{\circ}{\psi}_{0,\alpha} \overset{\circ}{\psi}_{0,\alpha} - 4(\rho\theta^2 - \xi)\tilde{\alpha}^{-1} l^2 \overset{\circ}{\psi}_{0,2} \overset{\circ}{\psi}_{0,2} \Big] d\omega$$

$$\ge C_0 \int_{\omega} h(\overset{\circ}{\psi}{}^2_{0,1} + \overset{0}{\psi}{}^2_{0,2}) d\omega = C_0 \|\overset{\circ}{\psi}_0\|^2_{X^{\kappa}}, \ C_0 := \min\{1, 1 - 4(\rho\theta^2 - \xi)\tilde{\alpha}^{-1} l^2\}.$$

Now, we have to prove (17).

$$|J(\overset{\circ}{\psi}_0, \overset{\circ}{\psi}{}^*_0)|^2 \le C_2 \Big| \int_{\omega} h \Big(\overset{\circ}{\psi}_{0,1} \overset{\circ}{\psi}{}^*_{0,1} + \overset{\circ}{\psi}_{0,2} \overset{\circ}{\psi}{}^*_{0,2} \Big) d\omega \Big|^2$$

$$+ \Big| \int_{\omega} (\rho\theta^2 - \xi)\tilde{\alpha}^{-1} h \overset{\circ}{\psi}_0 \overset{\circ}{\psi}{}^*_0 d\omega \Big|^2$$

$$+ 2C_3 \Big| \int_{\omega} h \Big(\overset{\circ}{\psi}_{0,1} \overset{\circ}{\psi}{}^*_{0,1} + \overset{\circ}{\psi}_{0,2} \overset{\circ}{\psi}{}^*_{0,2} \Big) d\omega \Big| \Big| \int_{\omega} (\rho\theta^2 - \xi)\tilde{\alpha}^{-1} h \overset{\circ}{\psi}_0 \overset{\circ}{\psi}{}^*_0 d\omega \Big|$$

$$\le C_2 \|\overset{\circ}{\psi}_0\|^2 \|\overset{\circ}{\psi}{}^*_0\|^2 + 16(\rho\theta^2 - \xi)^2 \tilde{\alpha}^{-2} l^{2\kappa} h_0^{2\kappa} \int_{\omega} x_2^{\kappa} \overset{\circ}{\psi}{}^2_{0,2} d\omega \int_{\omega} x_2^{\kappa} \overset{\circ}{\psi}{}^{*2}_{0,2} d\omega$$

$$+ 2C_3(\rho\theta^2 - \xi)\tilde{\alpha}^{-1} h_0^{\kappa} \Big[\Big| \int_{\omega} x_2^{\kappa} (\overset{\circ}{\psi}_{0,1} \overset{\circ}{\psi}{}^*_{0,1} + \overset{\circ}{\psi}_{0,2} \overset{\circ}{\psi}{}^*_{0,2}) d\omega \Big|^2 \Big| \int_{\omega} x_2^{\kappa} \overset{\circ}{\psi}_0 \overset{\circ}{\psi}{}^*_0 d\omega \Big|^2 \Big]^{1/2}$$

$$\le C_2 \|\overset{\circ}{\psi}_0\|^2 \|\overset{\circ}{\psi}{}^*_0\|^2$$

$$+ 16(\rho\theta^2 - \xi)^2\tilde{\alpha}^{-2}l^{2\kappa}h_0^{2\kappa}\int_\omega x_2^\kappa(\overset{\circ}{\psi}{}^2_{0,1} + \overset{\circ}{\psi}{}^2_{0,2})d\omega\int_\omega x_2^\kappa(\overset{\circ}{\psi}{}^{*2}_{0,1} + \overset{\circ}{\psi}{}^{*2}_{0,2})d\omega$$

$$+ 2C_3(\rho\theta^2 - \xi)\tilde{\alpha}^{-1}h_0^\kappa l^2\Big[\|\overset{\circ}{\psi}\|^2\|\overset{\circ}{\psi}{}^*_0\|^2\int_\omega x_2^\kappa\overset{\circ}{\psi}{}^2_{0,2}d\omega\int_\omega x_2^\kappa\overset{\circ}{\psi}{}^{*2}_{0,2}d\omega$$

$$\le C_4\|\overset{\circ}{\psi}{}_0\|^2\|\overset{\circ}{\psi}{}^*_0\|^2 + C_5\|\overset{\circ}{\psi}{}_0\|^2\|\overset{\circ}{\psi}{}^*_0\|^2 + C_6\|\overset{\circ}{\psi}{}_0\|^2\|\overset{\circ}{\psi}{}^*_0\|^2$$

this proves inequality (17).

Remark 1 If $J(\overset{\circ}{\psi}{}_0, \overset{\circ}{\psi}{}_0) = 0$, then $v \equiv 0$ by (18).

Theorem 3 *Let F be a bounded linear functional from $[X^\kappa]^*$. Then the variational problem (16) has a unique solution $v \in X^\kappa$ for an arbitrary value of the parameter κ and*

$$\|\overset{\circ}{\psi}{}_0\|_{X^\kappa} \le \frac{1}{C_0}\|F\|_{[X^\kappa]^*}.$$

Proof Taking into account Lemma 1.3, the proof immediately follows from the Lax-Milgram theorem. ■

Remark 2 It can be easily shown that if $F \in L(\omega)$ and $\operatorname{supp} F \cap \overline{\gamma}_0 = \emptyset$, then $F \in [X^\kappa]^*$ and

$$\langle F\,,\ \overset{\circ}{\psi}{}^*_0\rangle = \int\limits_\omega F(x)\,\overset{\circ}{\psi}{}^*_0(x)\,d\omega,$$

since $\overset{\circ}{\psi}{}^*_0 \in H^1(\omega_\varepsilon)$, where ε is sufficiently small positive number such that $\operatorname{supp} F \subset \omega_\varepsilon = \omega \cap \{x_2 > \varepsilon\}$. Therefore,

$$|\langle F\,,\ \overset{\circ}{\psi}{}^*_0\rangle| = \Big|\int\limits_\omega F(x)\,\overset{\circ}{\psi}{}^*_0(x)\,d\omega\Big| \le ||F||_{L_2(\omega)}\ ||\overset{\circ}{\psi}{}^*_0||_{L_2(\omega_\varepsilon)}$$

$$\le ||F||_{L_2(\omega)}\ ||\overset{\circ}{\psi}{}^*_0||_{H^1(\omega_\varepsilon)} \le C_\varepsilon\ ||F||_{L_2(\omega)}\ ||\overset{\circ}{\psi}{}^*_0||_{X^\kappa}\,.$$

In this case, we obtain the estimate

$$||\overset{\circ}{\psi}{}_0||_{X^\kappa} \le \frac{C_\varepsilon}{C_0}\,||F||_{L_2(\omega)}\,.$$

Remark 3 The space X^κ is a weighted Sobolev space.

Corollary 1 $\overset{\circ}{\psi}_0$ *has the zero trace on* $\partial\omega$ *if* $\kappa < 1$.

Remark 4 In case of full system

$$-\mu\Big[(h\overset{\circ}{u}_{\alpha,\beta})_{,\alpha}+(h\overset{\circ}{u}_{\beta,\alpha})_{,\alpha}\Big]-\lambda(h\overset{\circ}{u}_{\gamma,\gamma})_{,\beta}-b(h\overset{\circ}{u}_4)_{,\beta}-\rho h\vartheta^2\overset{\circ}{u}_\beta=F_\beta,$$

$$-\mu(h\overset{\circ}{u}_{3,\alpha})_{,\alpha}-\rho h\vartheta^2\overset{\circ}{u}_{30}=F_3,$$

$$-\tilde{\alpha}(h\overset{\circ}{u}_{4,\alpha})_{,\alpha}+bh\overset{\circ}{u}_{\gamma,\gamma}+\xi h\overset{\circ}{u}_4-\rho h\vartheta^2\overset{\circ}{u}_4=F_4,\quad \beta=1,2,$$

where

$$\overset{\circ}{u}_i:=\overset{\circ}{v}_{i0},\quad \overset{\circ}{u}_4:=\overset{\circ}{\psi}_0,\quad F_i:=\overset{\circ}{X}_i,\quad F_4:=\overset{\circ}{H}+\overset{\circ}{\mathcal{F}}_0,$$

we introduce the space Y^κ with inner product and a norm as follows

$$(\overset{\circ}{u},\overset{\circ}{u}{}^*)_{Y^\kappa}:=\int_\omega x_2^\kappa\Big[\Big(\overset{\circ}{u}_{1,1}+\overset{\circ}{u}_4\Big)\Big(\overset{\circ}{u}{}^*_{1,1}+\overset{\circ}{u}{}^*_4\Big)+\Big(\overset{\circ}{u}_{2,2}+\overset{\circ}{u}_4\Big)\Big(\overset{\circ}{u}{}^*_{2,2}+\overset{\circ}{u}{}^*_4\Big)$$
$$+\Big(\overset{\circ}{u}_{1,2}+\overset{\circ}{u}_{2,1}\Big)\Big(\overset{\circ}{u}{}^*_{1,2}+\overset{\circ}{u}{}^*_{2,1}\Big)+\overset{\circ}{u}_{3,\alpha}\overset{\circ}{u}{}^*_{3,\alpha}+\overset{\circ}{u}_{4,\alpha}\overset{\circ}{u}{}^*_{4,\alpha}\Big]d\omega,$$

$$\|\overset{\circ}{u}\|^2_{Y^\kappa}:=\int_\omega x_2^\kappa\Big[\Big(\overset{\circ}{u}_{1,1}+\overset{\circ}{u}_4\Big)^2+\Big(\overset{\circ}{u}_{2,2}+\overset{\circ}{u}_4\Big)^2+\Big(\overset{\circ}{u}_{1,2}+\overset{\circ}{u}_{2,1}\Big)^2$$
$$+\overset{\circ}{u}{}^2_{3,1}+\overset{\circ}{u}{}^2_{3,2}+\overset{\circ}{u}{}^2_{4,1}+\overset{\circ}{u}{}^2_{4,2}\Big]d\omega.$$

If

$$\vartheta^2\le\min\left\{\frac{\mu}{2\rho l^2};\frac{\xi}{\rho}\right\}\tag{19}$$

the coerciveness of the corresponding bilinear form and uniqueness and existence results for the variational problem can proved analogously.

In view of the homogeneous Dirichlet boundary condition, if $\kappa > 1$, the following Hardy inequality holds (see [13], p. 69, [10])

$$\int_\varepsilon^l x_2^{\kappa-2}v_{\alpha0}^2dx_2\ge\frac{4}{(\kappa-1)^2}\int_\varepsilon^l x_2^\kappa(v_{\alpha0,2})^2dx_2,\quad \kappa>1.$$

Replacing in last inequality κ by $\kappa+2$, we obtain

$$\int_\varepsilon^l x_2^\kappa v_{\alpha0}^2dx_2\ge\frac{4}{(\kappa-1)^2}\int_\varepsilon^l x_2^{\kappa+2}(v_{\alpha0,2})^2dx_2,\quad \textit{for any } \kappa>0.$$

Now, considering the limit procedure as $\varepsilon \to 0+$, since the limits of the last integrals exist for $v_{\alpha 0} \in Y^{\kappa}$, we immediately get the following

$$\int_0^l x_2^{\kappa} v_{\alpha 0}^2 dx_2 \geq \frac{4}{(\kappa-1)^2} \int_0^l x_2^{\kappa+2} (v_{\alpha 0,2})^2 dx_2, \quad for\ any\ \kappa > 0.$$

Integrating by x_1 over $]x_1^0, x_1^1[$, we get

$$\int_{\omega} x_2^{\kappa} v_{\alpha 0}^2 d\omega \geq \frac{4}{(\kappa-1)^2} \int_{\omega} x_2^{\kappa+2} (v_{\alpha 0,2})^2 d\omega, \quad for\ any\ \kappa > 0.$$

The linear spaces X^{κ} and Y^{κ} as sets of vector function as coincide and the norms $\|\cdot\|_{X^{\kappa}}$, $\|\cdot\|_{Y^{\kappa}}$ are equivalent if (19) is fulfilled.

References

1. Amelchenko, A.G., Bardin, V.A., VasiFev, V.A., Krevchick, V.D., Chernov, P.S., Shcherbakov, M.A.: Piezo actuators and piezo motors for driving systems. In: Dynamics of Systems, Mechanisms and Machines (Dynamics), pp. 1–4 (2016)
2. Chinchaladze, N., Gilbert, R., Jaiani, G., Kharibegashvili, S., Natroshvili, D.: Existence and uniqueness theorems for cusped prismatic shells in the N-th hierarchical model. Math. Methods Appl. Sci. **31**(11), 1345–1367 (2008)
3. Cowin, S.C., Nunziato, J.W.: Linear elastic materials with voids. J. Elasticity **13**, 125–147 (1983)
4. Cugat, O., Delamare, J., Reyne, G.: Magnetic micro-actuators systems (magmas). In: 2003 IEEE International Magnetics Conference (INTERMAG), pp. GB–04 (2003)
5. Eringen, A.C.: Mechanics of Continua. R.E. Krieger Publ. Com. Inc, Huntington, New York (1980)
6. Ieşan, D.: Classical and Generalized Models of Elastic Rods. CRC Press, A. Chapman Hall Book (2009)
7. Jaiani, G.: Cusped Shell-Like Structures. SpringerBriefs in Applied Science and Technology. Springer, Heidelberg (2011)
8. Jaiani, G.: Hierarchical models for viscoelastic Kelvin-Voigt prismatic shells with voids. Bull. TICMI **21**(1), 33–44 (2017)
9. Jaiani, G.: Piezoelectric Viscoelastic Kelvin-Voigt Cusped Prismatic Shells. Lecture Notes of TICMI, vol. 19 (2018)
10. Jaiani, G., Kufner, A.: Oscillation of cusped Euler-Bernoulli beams and Kirchhoff-Love plates. Hacettepe J. Math. Stat. **35**(1), 7–53 (2006)
11. Mittal, N., Ansari, F., Gowda, V.K., Brouzet, Ch., Chen, P., Larsson, P., Roth, S., Lundell, F., Wagberg, L., Kotov, N., Söderberg, D.: Multiscale control of nanocellulose assembly: transferring remarkable nanoscale fibril mechanics to macroscale fibers. ACS Nano **12**(5), 6378 (2018)
12. Natroshvili, D.: Mathematical Problems of Thermo-Electro-Magneto-Elasticity. Lecture Notes of TICMI, vol. 12. Tbilisi University Press (2011)
13. Opic, B., Kufner, A.: Hardy-type Inequality. Longman Sci. Tech, Harlow (1990)
14. Svanadze, M.M.: Steady vibrations problem in the theory of viscoelasticity for Kelvin-Voigt materials with voids. Proc. Appl. Math. Mech. **12**, 283–284 (2012)

15. Svanadze, M.M.: Potential method in the linear theory of viscoelastic materials with voids. J. Elasticity **114**(1), 101–126 (2014)
16. Taha, M., Walia, S., Ahmed, T., Headland, D., Withayachumnankul, W., Sriram, S., Bhaskaran, M.: Insulator-metal transition in substrate-independent vo2 thin film for phase-change devices. Sci. Rep. **7**, 17899 (2017)
17. Trindade, M.: Applications of piezoelectric sensors and actuators for active and passive vibration control. In: Conference Papers, Conference: 7th Brazilian Conference on Dynamics, Control and Applications, 05 (2008)
18. Vekua, I.N.: On a way of calculating of prismatic shells. Proc. A Razmadze Inst. Math. Georgian Acad. Sci. **21**, 191–259 (1955) (Russian)
19. Vekua, I.N.: The theory of thin shallow shells of variable thickness. Proc. A Razmadze Inst. Math. Georgian Acad. Sci. **30**, 5–103 (1965) (Russian)
20. Vekua, I.N.: Shell Theory: General Methods of Construction, Pitman Advanced Publishing Program. Boston, London (1985)

Solution of the Kirsch Problem for the Elastic Materials with Voids in the Case of Approximation $N = 1$ of Vekua's Theory

Bakur Gulua, Roman Jangava, Tamar Kasrashvili, and Miranda Narmania

Abstract In this paper we consider a boundary value problem for an infinite plate with a circular hole. The plate is the elastic material with voids. The hole is free from stresses, while unilateral tensile stresses act at infinity. The state of plate equilibrium is described by the system of differential equations that is derived from three-dimensional equations of equilibrium of an elastic material with voids (Cowin-Nunziato model) by Vekua's reduction method. its general solution is represented by means of analytic functions of a complex variable and solutions of Helmholtz equations. The problem is solved analytically by the method of the theory of functions of a complex variable.

1 Introduction

The nonlinear and linear theories for the behaviour of porous solids, in which the skeletal or matrix material is elastic and the interstices are voids of the material, was developed by Nunziato and Cowin [1, 2]. Such materials include, in particular,

B. Gulua (✉)
Sokhumi State University, Tbilisi, Georgia

I. Vekua Institute of Applied Mathematics of I. Javakhishvili Tbilisi State University, Tbilisi, Georgia

R. Janjgava
I. Vekua Institute of Applied Mathematics of I. Javakhishvili Tbilisi State University, Tbilisi, Georgia

Georgian National University SEU, Tbilisi, Georgia

T. Kasrashvili
Georgian Technical University, Tbilisi, Georgia

I. Vekua Institute of Applied Mathematics of I. Javakhishvili Tbilisi State University, Tbilisi, Georgia

M. Narmania
University of Georgia, Tbilisi, Georgia

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_16

rocks and soils, granulated and some other manufactured porous materials. This theory differs essentially form the classical theory of elasticity in that the volume fraction function corresponding to the void volume is considered as an independent variable. In spite of a great number of works on the theory of elastic materials with voids or empty pores, only a limited number of them focus on the study of plates and shells.

As is known, there exist many methods of reducing three-dimensional problems of equilibrium of elastic shells to two-dimensional problems. Some such general methods were proposed by famous mathematician and mechanician I. Vekua [3, 4]. He used the Cauchy−Poisson method, which is based on the expansion of displacements and stresses into series in terms of a system of functions with respect to the thickness coordinate. As basis functions Vekua used the Legendre polynomials, which make up a complete system on the considered interval, and for expansion coefficients he obtained a two-dimensional system of equilibrium equations for shells of variable thickness. According to Vekua's method, in the expansions of the sought functions we can preserve only the first member (approximation N = 0), the first two members (N = 1), the first three members (N = 2) and so on. Thus, the models obtained by the method under consideration are often called the hierarchical models of elastic shells.

2 Statement of the Problem

Let $Ox_1x_2x_3$ be the rectangular Cartesian coordinate system. Let $\Omega = \omega \times]-h,\ h[$ be an infinite plate with a circular hole of radius R centred at the origin O. The plate thickness is assumed to be constant and equal to $2h$. The plate is the isotropic material with voids.

The governing equations of the theory of elastic materials with voids can be expressed in the following form [2]:

- Equations of equilibrium

$$T_{ij,j} + \Phi_i = 0, \quad j = 1, 2, 3, \tag{1}$$

$$h_{i,i} + g + \Psi = 0, \tag{2}$$

  where T_{ij} is the symmetric stress tensor, Φ_i are the volume force components, h_i is the equilibrated stress vector, g is the intrinsic equilibrated body force and Ψ is the extrinsic equilibrated body force.
- Constitutive equations

$$\begin{aligned} &T_{ij} = \lambda e_{kk}\delta_{ij} + 2\mu e_{ij} + \beta\phi\delta_{ij}, \quad i, j = 1, 2, 3, \\ &h_i = \alpha\phi_{,i}, \quad i = 1, 2, 3, \\ &g = -\xi\phi - \beta e_{kk}, \end{aligned} \tag{3}$$

where λ and μ are the Lamé constants; α, β and ξ are the constants characterizing the body porosity; δ_{ij} is the Kronecker delta; $\phi := \nu - \nu_0$ is the change of the volume fraction function from the matrix reference volume fraction ν_0 (clearly, the bulk density $\rho = \nu\gamma$, $0 < \nu \leq 1$, here γ is the matrix density and ρ is the mass density); e_{ij} is the strain tensor and

$$e_{ij} = \tfrac{1}{2}\left(u_{i,j} + u_{j,i}\right), \tag{4}$$

where u_i, $i = 1, 2, 3$ are the components of the displacement vector..

The constitutive equations also meet some other conditions, following from physical considerations

$$\begin{aligned} &\mu > 0, \quad \alpha > 0, \quad \xi > 0, \\ &3\lambda + 2\mu > 0, \quad (3\lambda + 2\mu)\xi > 3\beta^2. \end{aligned} \tag{5}$$

In [5] using Vekua's dimension reduction method [3], linear two-dimensional (2D) governing equations were obtained from the above three-dimensional (3D) equations with respect to so-called r-th order moments of functions under consideration, where the zero order moments (which are averaged along the thickness of the plate) and the first order moments are defined as

$$\begin{pmatrix} \overset{(0)}{u_i}, \overset{(0)}{\phi} \end{pmatrix} = \frac{1}{2h}\int_{-h}^{h} (u_i, \phi)\, dx_3, \quad \begin{pmatrix} \overset{(1)}{u_i}, \overset{(1)}{\phi} \end{pmatrix} = \frac{3}{2h^2}\int_{-h}^{h} x_3 \cdot (u_i, \phi)\, dx_3.$$

Besides, In Section 7 under title $N = 0$ Approximation for Porous Isotropic Elastic Shells of [5] it is shown that in the case of cusped prismatic shells, depending on the character of vanishing of the thickness at the lateral boundary of the shell the boundary conditions of 2D problems for displacements and volume fraction functions are non-classical, in general, and the criteria's are given when the Dirichlet or the Keldysh type Boundary value problems are well-posed; The case of a plate of constant thickness is considered as well.

In particular, in the $N = 1$ approximation of I.Vekua's theory it is assumed that

$$u_i(x_1, x_2, x_3) = \overset{(0)}{u_i}(x_1, x_2) + \frac{x_3}{h}\overset{(1)}{u_i}(x_1, x_2),$$

$$\phi(x_1, x_2, x_3) = \overset{(0)}{\phi}(x_1, x_2) + \frac{x_3}{h}\overset{(1)}{\phi}(x_1, x_2),$$

Similarly, for the stress tensor components, the components of the equilibrated stress vector and the intrinsic equilibrated volume force we will have the following zero and first order moments

$$\begin{pmatrix} \overset{(0)}{T}_{ij}, \overset{(0)}{h}_i, \overset{(0)}{g} \end{pmatrix} = \frac{1}{2h} \int_{-h}^{h} \left(T_{ij}, h_i, g\right) dx_3,$$
$$\begin{pmatrix} \overset{(1)}{T}_{ij}, \overset{(1)}{h}_i, \overset{(1)}{g} \end{pmatrix} = \frac{3}{2h^2} \int_{-h}^{h} x_3 \cdot \left(T_{ij}, h_i, g\right) dx_3.$$

For $h = const$ the reduced system of equilibrium equations gets split into two independent systems: tension−compression equations with unknowns $\overset{(0)}{u}_1, \overset{(0)}{u}_2, \overset{(1)}{u}_3, \overset{(0)}{\phi}$ and bending equations with unknowns $\overset{(1)}{u}_1, \overset{(1)}{u}_2, \overset{(0)}{u}_3, \overset{(1)}{\phi}$. In this paper we consider the system of tension−compression equations.

In the case $N = 1$ approximation from [5] the basic relations of elastic isotropic plates with voids have the following form:

$$\begin{aligned} &\partial_\alpha \overset{(0)}{T}_{\alpha\gamma} = 0, \quad \alpha, \gamma = 1, 2 \\ &\partial_\alpha \overset{(1)}{T}_{\alpha 3} - \frac{3}{h} \overset{(1)}{T}_{33} = 0, \\ &\partial_\alpha \overset{(0)}{h}_\alpha + \overset{(0)}{g} = 0, \end{aligned} \tag{6}$$

where

$$\begin{aligned} &\overset{(0)}{T}_{\alpha\gamma} = \lambda \left(\overset{(0)}{\theta} + \overset{(1)}{u}_3 \right) \delta_{\alpha\gamma} + \mu \left(\partial_\alpha \overset{(0)}{u}_\gamma + \partial_\gamma \overset{(0)}{u}_\alpha \right) + \beta \overset{(0)}{\phi}\, \delta_{\alpha\gamma}, \\ &\overset{(0)}{T}_{33} = \lambda \left(\overset{(0)}{\theta} + \overset{(1)}{u}_3 \right) + 2\mu \overset{(1)}{u}_3 + \beta \overset{(0)}{\phi}, \\ &\overset{(1)}{T}_{\gamma 3} = \mu \partial_\gamma \overset{(1)}{u}_3, \\ &\overset{(0)}{h}_\gamma = \alpha \partial_\gamma \overset{(0)}{\phi}, \\ &\overset{(0)}{g} = -\xi \overset{(0)}{\phi} - \beta \left(\overset{(0)}{\theta} + \overset{(1)}{u}_3 \right), \end{aligned} \tag{7}$$

$$\overset{(0)}{\theta} = \partial_1 \overset{(0)}{u}_1 + \partial_2 \overset{(0)}{u}_2.$$

Substituting (7) into system (6), we obtain the following system of governing equations of statics with respect to the functions $\overset{(0)}{u}_1, \overset{(0)}{u}_2, \overset{(1)}{u}_3, \overset{(0)}{\phi}$

$$\begin{aligned}
&\mu\Delta\overset{(0)}{u}_1 + (\lambda+\mu)\partial_1\overset{(0)}{\theta} + \lambda\partial_1\overset{(1)}{u}_3 + \beta\partial_1\overset{(0)}{\phi} = 0,\\
&\mu\Delta\overset{(0)}{u}_2 + (\lambda+\mu)\partial_2\overset{(0)}{\theta} + \lambda\partial_2\overset{(1)}{u}_3 + \beta\partial_2\overset{(0)}{\phi} = 0,\\
&\mu\Delta\overset{(1)}{u}_3 - \frac{3}{h}\left[\lambda\overset{(0)}{\theta} + (\lambda+2\mu)\overset{(1)}{u}_3 + \beta\overset{(0)}{\phi}\right] = 0,\\
&(\alpha\Delta - \xi)\overset{(0)}{\phi} - \beta\left[\overset{(0)}{\theta} + \overset{(1)}{u}_3\right] = 0,
\end{aligned} \tag{8}$$

where $\Delta := \partial_{11} + \partial_{22}$ is the two-dimensional Laplace operator.

On the plane Ox_1x_2, we introduce the complex variable $z = x_1 + ix_2 = re^{i\vartheta}$, $(i^2 = -1)$ and the operators $\partial_z = 0.5(\partial_1 - i\partial_2)$, $\partial_{\bar{z}} = 0.5(\partial_1 + i\partial_2)$, $\bar{z} = x_1 - ix_2$, and $\Delta = 4\partial_z\partial_{\bar{z}}$.

In order to write system (8) in the complex form, we multiply the second equation of the system by i and sum the obtained with the first equation:

$$\begin{aligned}
&2\mu\partial_{\bar{z}}\partial_z\overset{(0)}{u}_+ + (\lambda+\mu)\partial_{\bar{z}}\overset{(0)}{\theta} + \lambda\partial_{\bar{z}}\overset{(1)}{u}_3 + \beta\partial_{\bar{z}}\overset{(0)}{\phi} = 0,\\
&\mu\Delta\overset{(1)}{u}_3 - \frac{3}{h}\left[\lambda\overset{(0)}{\theta} + (\lambda+2\mu)\overset{(1)}{u}_3 + \beta\overset{(0)}{\phi}\right] = 0,\\
&(\alpha\Delta - \xi)\overset{(0)}{\phi} - \beta\left[\overset{(0)}{\theta} + \overset{(1)}{u}_3\right] = 0,
\end{aligned} \tag{9}$$

where $\overset{(0)}{u}_+ = \overset{(0)}{u}_1 + i\overset{(0)}{u}_2$, $\overset{(0)}{\theta} = \partial_z\overset{(0)}{u}_+ + \partial_{\bar{z}}\overset{(0)}{\bar{u}}_+$.

As the analogues of the Kolosov-Muskhelishvili formulas [6] for system (9) we have

$$\begin{aligned}
&2\mu\overset{(0)}{u}_+ = \varkappa_1\varphi(z) - \varkappa_2 z\overline{\varphi'(z)} - \overline{\psi(z)} - p_1\partial_{\bar{z}}\chi_1(z,\bar{z}) - p_2\partial_{\bar{z}}\chi_2(z,\bar{z}),\\
&\overset{(1)}{u}_3 = l_{11}\chi_1(z,\bar{z}) + l_{12}\chi_2(z,\bar{z}) - E_1(\varphi'(z) + \overline{\varphi'(z)}),\\
&\overset{(0)}{\phi} = l_{21}\chi_1(z,\bar{z}) + l_{22}\chi_2(z,\bar{z}) - E_2(\varphi'(z) + \overline{\varphi'(z)}),
\end{aligned} \tag{10}$$

where $\varphi(z)$ and $\psi(z)$ are the arbitrary analytic functions of z, $\chi_1(z,\bar{z})$ and $\chi_2(z,\bar{z})$ are the general solutions of the Helmholtz equations

$$\Delta\chi - \kappa_1\chi = 0, \quad \Delta\chi - \kappa_2\chi = 0,$$

and κ_1, κ_2 are eigenvalues and l_{11}, l_{21}, l_{12}, l_{22} are eigenvectors of the matrix C. $E_1 = a_{11} + a_{12}$, $E_2 = a_{21} + a_{22}$ and a_{ij} are coefficients of the matrix $-C^1 D$:

$$C = \begin{pmatrix} \frac{12(\lambda+\mu)}{h(\lambda+2\mu)} & \frac{6\beta}{h(\lambda+2\mu)} \\ \frac{2\mu\beta}{\alpha(\lambda+2\mu)} & \frac{\xi}{\alpha} - \frac{\beta^2}{\alpha(\lambda+2\mu)} \end{pmatrix}, \quad D = \begin{pmatrix} \frac{3\lambda}{2h\mu(\lambda+2\mu)} & 0 \\ 0 & \frac{\beta}{2\alpha(\lambda+2\mu)} \end{pmatrix}.$$

Also $\varkappa_1 = \frac{1}{2} + \frac{(\lambda E_1+\beta E_2)\mu}{\lambda+2\mu}$, $\varkappa_2 = \frac{1}{2} - \frac{(\lambda E_1+\beta E_2)\mu}{\lambda+2\mu}$, $p_1 = \frac{4(\lambda l_{11}+\beta l_{21})\mu}{\kappa_1(\lambda+2\mu)}$, $p_2 = \frac{4(\lambda l_{12}+\beta l_{22})\mu}{\kappa_2(\lambda+2\mu)}$.

From (10) complex combinations of the stress tensor components are expressed by means of the formulas

$$\begin{aligned}
&\overset{(0)}{T}_{11} - \overset{(0)}{T}_{22} + 2i\overset{(0)}{T}_{12} = -2\varkappa_2 z\overline{\varphi''(z)} - \overline{\psi'(z)} - 2p_1\partial^2_{\bar z\bar z}\chi_1(z,\bar z) - 2p_2\partial^2_{\bar z\bar z}\chi_2(z,\bar z),\\
&\overset{(0)}{T}_{11} + \overset{(0)}{T}_{22} = E_3(\varphi'(z) + \overline{\varphi'(z)}) + E_4\chi_1(z,\bar z) + E_5\chi_2(z,\bar z),\\
&\overset{(1)}{T}_{+} = l_{11}\partial_{\bar z}\chi_1(z,\bar z) + l_{12}\partial_{\bar z}\chi_2(z,\bar z) - E_1\overline{\varphi''(z)}),\\
&\overset{(0)}{h}_{+} = l_{21}\partial_{\bar z}\chi_1(z,\bar z) + l_{22}\partial_{\bar z}\chi_2(z,\bar z) - E_2\overline{\varphi''(z)}),
\end{aligned} \tag{11}$$

where

$$E_3 = \frac{\lambda+\mu}{\mu}(\varkappa_1+\varkappa_2) - 2\lambda E_1 - 2\beta E_2,$$

$$E_4 = 2\lambda l_{11} + 2\beta l_{21} - \frac{\lambda+\mu}{\mu}8p_1\kappa_1,$$

$$E_5 = 2\lambda l_{12} + 2\beta l_{22} - \frac{\lambda+\mu}{\mu}8p_2\kappa_2.$$

Now consider a boundary value problem for an infinite plate with a circular hole Fig. 1. We formulate the boundary value problem: find such a solution of system (9) that on the hole contour and at infinity satisfies respectively the following boundary conditions

$$\begin{aligned}
&\overset{(0)}{T}_{rr} + i\overset{(0)}{T}_{r\vartheta} = 0, \quad \overset{(1)}{T}_{r3} = 0, \quad \overset{(0)}{h}_{r3} = 0,\\
&\overset{(0)}{T}{}^{\infty}_{11} = p, \quad \overset{(0)}{T}{}^{\infty}_{12} = \overset{(0)}{T}{}^{\infty}_{22} = \overset{(0)}{T}{}^{\infty}_{13} = \overset{(0)}{T}{}^{\infty}_{23} = 0,
\end{aligned} \tag{12}$$

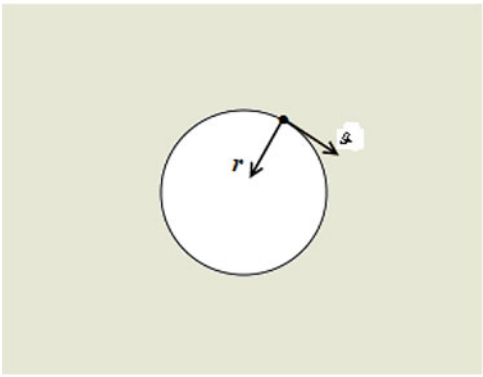


Fig. 1 An infinite plate with a circular hole

where

$$\begin{aligned}
&\overset{(0)}{T}_{rr} + i\overset{(0)}{T}_{r\vartheta} = \frac{1}{2}\left\{\overset{(0)}{T}_{11} + \overset{(0)}{T}_{22} + \left[\overset{(0)}{T}_{11} - \overset{(0)}{T}_{22} + 2i\overset{(0)}{T}_{12}\right] e^{-2i\vartheta}\right\},\\
&\overset{(1)}{T}_{r3} = \mathrm{Re}\left\{\overset{(1)}{T}_{+} e^{-i\vartheta}\right\},\\
&\overset{(1)}{h}_{r3} = \mathrm{Re}\left\{\overset{(0)}{h}_{+} e^{-i\vartheta}\right\}.
\end{aligned}$$

3 Solution of the Problem

Taking into account formula (11), the boundary conditions (12) take the form

$$\begin{aligned}
&\overset{(0)}{T}_{rr} + i\overset{(0)}{T}_{r\vartheta} = E_3(\varphi'(z) + \overline{\varphi'(z)}) + E_4\chi_1(z,\bar{z}) + E_5\chi_2(z,\bar{z})\\
&\left(-2\varkappa_2 z\overline{\varphi''(z)} - \overline{\psi'(z)} - 2p_1\partial^2_{\bar{z}\bar{z}}\chi_1(z,\bar{z}) - 2p_2\partial^2_{\bar{z}\bar{z}}\chi_2(z,\bar{z})\right) e^{-2i\vartheta},\\
&\overset{(1)}{T}_{r3} = \left(l_{11}\partial_{\bar{z}}\chi_1(z,\bar{z}) + l_{12}\partial_{\bar{z}}\chi_2(z,\bar{z}) - E_1\overline{\varphi''(z)}\right) e^{-i\vartheta}\\
&+ \left(l_{11}\partial_z\chi_1(z,\bar{z}) + l_{12}\partial_z\chi_2(z,\bar{z}) - E_1\varphi''(z)\right) e^{i\vartheta} = 0,\\
&\overset{(1)}{h}_{r3} = \left(l_{21}\partial_{\bar{z}}\chi_1(z,\bar{z}) + l_{22}\partial_{\bar{z}}\chi_2(z,\bar{z}) - E_2\overline{\varphi''(z)}\right) e^{-i\vartheta}\\
&+ \left(l_{21}\partial_z\chi_1(z,\bar{z}) + l_{22}\partial_z\chi_2(z,\bar{z}) - E_2\varphi''(z)\right) e^{-i\vartheta}.
\end{aligned} \tag{13}$$

The analytic functions $\varphi'(z)$, $\psi'(z)$ and the metaharmonic functions $\chi_1(z,\bar{z})$ and $\chi_2(z,\bar{z})$ are represented as the series

$$\begin{aligned}
&\varphi'(z)=\sum_{n=0}^{\infty}a_n z^{-n},\quad \psi'(z)=\sum_{n=0}^{\infty}b_n z^{-n},\\
&\chi_1(z,\bar{z})=\sum_{-\infty}^{+\infty}\alpha_n K_n(\sqrt{\kappa_1}r)e^{in\vartheta},\quad \chi_2(z,\bar{z})=\sum_{-\infty}^{+\infty}\beta_n K_n(\sqrt{\kappa_2}r)e^{in\vartheta}
\end{aligned}\tag{14}$$

where $K_n(\zeta r)$ is a modified Bessel function of n-th order. From (5) κ_1 and κ_2 are positive numbers.

Bearing in mind conditions at infinity, from formulas (11) we define the coefficients a_0 and b_0

$$a_0=\frac{E_3}{2}p,\quad b_0=-p.\tag{15}$$

Substituting (14) into (13), comparing the coefficients of same exponents and bearing in mind the condition of uniqueness of displacements, which follows from formulas (11)

$$\varkappa_1 a_1+\varkappa_2\bar{b}_1=0,$$

we obtain

$$\begin{aligned}
&b_2=2E_2R^2a_0,\\
&\frac{E_3}{R^2}\bar{a}_2+\left(E_4K_{-2}(\sqrt{\kappa_1}R)-\frac{p_1\kappa_1}{2}K_0(\sqrt{\kappa_1}R)\right)\alpha_2\\
&+\left(E_5K_{-2}(\sqrt{\kappa_2}R)-\frac{p_2\kappa_2}{2}K_0(\sqrt{\kappa_2}R)\right)\beta_2-\bar{b}_0=0,\\
&\frac{4E_1}{R^3}\bar{a}_2-l_{11}\sqrt{\kappa_1}\left(K_1(\sqrt{\kappa_1}R)+K_3(\sqrt{\kappa_1}R)\right)\alpha_2\\
&-l_{12}\sqrt{\kappa_2}\left(K_1(\sqrt{\kappa_2}R)+K_3(\sqrt{\kappa_2}R)\right)\beta_2=0,\\
&\frac{4E_2}{R^3}\bar{a}_2-l_{21}\sqrt{\kappa_1}\left(K_1(\sqrt{\kappa_1}R)+K_3(\sqrt{\kappa_1}R)\right)\alpha_2\\
&-l_{22}\sqrt{\kappa_2}\left(K_1(\sqrt{\kappa_2}R)+K_3(\sqrt{\kappa_2}R)\right)\beta_2=0,\\
&\frac{E_3-2\varkappa_2}{R^2}\bar{a}_2-frac14\bar{b}_4+\left(E_4K_2(\sqrt{\kappa_1}R)-\frac{p_1\kappa_1}{2}K_4(\sqrt{\kappa_1}R)\right)\alpha_2\\
&+\left(E_5K_2(\sqrt{\kappa_2}R)-\frac{p_2\kappa_2}{2}K_4(\sqrt{\kappa_2}R)\right)\beta_2=0.
\end{aligned}\tag{16}$$

So from (14) and (15) we find $a_0,\ a_2,\ b_0,\ b_2,\ b_4,\ \alpha_2,\ \beta_2$. All other coefficients in series (14) are equal to zero.

Thus, we have defined the sought functions $\varphi'(z)$, $\psi'(z)$, $\chi_1(z,\bar{z})$ and $\chi_2(z,\bar{z})$

$$\varphi'(z) = a_0 + \frac{a_2}{z^2}, \quad \psi'(z) = b_0 + \frac{b_2}{z^2} + \frac{b_4}{z^4},$$

$$\chi_1(z,\bar{z}) = 2K_2(\sqrt{\kappa_1}r)\alpha_2 \cos 2\vartheta, \quad \chi_2(z,\bar{z}) = 2K_2(\sqrt{\kappa_2}r)\beta_2 \cos 2\vartheta.$$

Thus, the problem stated is solved. As we see from the solution obtained, stresses depend on the materials of which the body consists.

References

1. Nunziato, J.W., Cowin, S.C.: A nonlinear theory of elastic materials with voids. Arch. Rat. Mech. Anal. **72**, 175–201 (1979)
2. Cowin, S.C., Nunziato, J.W.: Linear elastic materials with voids. J. Elasticity **13**, 125–147 (1983)
3. Vekua, I.: On a way of calculating of prismatic shells. Proceedings of A. Razmadze Institute of Mathematics of Georgian Academy of Sciences, **21**, 191–259 (1995)
4. Vekua, I.: Shell Theory: General Methods of Construction. Pitman Advanced Publishing Program, Boston (1985)
5. Jaiani, G.: Piezoelectric viscoelastic Kelvin-Voigt cusped prismatic shells. Lect. Notes TICMI **19**, 83 pp. (2018)
6. Muskhelishvili, N.I.: Some basic problems of the mathematical theory of elasticity. Nauka, Moscow (1966)

Analysis of BVP for Some Elliptic Systems on a Complex Plane

Giorgi Makatsaria and Nino Manjavidze

Abstract In the paper some special type elliptic systems of differential equations on the complex plane is studied. The correct BVP for these systems are considered. In some sense a unique class of solutions is effectively constructed for a sufficiently wide class of singular elliptic systems for which the Riemann-Hilbert problem can be correctly posed. The complete analysis of this problem is given.

1 Introduction

The boundary value problems for the first order elliptic systems on the complex plane were investigated extensively over the past years [4]. The first order linear system of partial differential equations

$$\frac{\partial u}{\partial x} = \mathcal{A}(x, y)\frac{\partial u}{\partial y} + \mathcal{B}(x, y)u(x, y) + \mathcal{F}(x, y),$$

where $u = (u_1, u_2, \ldots, u_n)$ is $2n$ desired vector, $\mathcal{A}, \mathcal{B}$ are given real $2n \times 2n$ matrices, depending on two variables x, y, and $\mathcal{F}$ is a given $2n$-vector. This system is elliptic in some plane domain $\mathbb{D}$ if and only if the matrix $\mathcal{A}$ has no real characteristic numbers in $\mathbb{D}$. When $n = 1$ in case of sufficient smoothness of the coefficients of the system, after corresponding change of variables this system can be reduced to one complex equation

$$\partial_{\bar{z}} w + Aw + B\bar{w} = F\left(\partial_{\bar{z}} = \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right)\right).$$

At present this equation is called Carleman-Vekua equation.

G. Makatsaria · N. Manjavidze (✉)
Ilia State University, Tbilisi, Georgia
e-mail: nino.manjavidze@iliauni.edu.ge

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_17

In this paper the Riemann-Hilbert boundary value problem

$$Re\{\overline{\lambda(t)}w(t)\} = \gamma(t), \quad t \in \Gamma, \tag{R-H}$$

for the following Carleman-Vekua equation

$$\frac{\partial w}{\partial \bar{z}} + A(z)w + B(z)\bar{w} = 0 \tag{C-V}$$

with the polar singularities in the domain G is investigated. G is a finite $m+1$-connected domain of the complex plane $z = x + iy$ with sufficiently smooth boundary provided that the given functions $\lambda(t)$ and $\gamma(t)$ are the Hölder continuous functions. It is well-known that the equation (C-V) in case of regular coefficients (i.e. $A(z), B(z) \in L_p(G)$ for some $p > 2$) the condition $\lambda(t) \neq 0, t \in \Gamma$ provides the Noetherity of the problem (R-H) in the class of continuous functions in $\bar{G}\setminus\{z_0\}$ satisfying the equation (C-V) in $G\setminus\{z_0\}$ and the asymptotic conditions $O\left(|z-z_0|^{\sigma}\right), z \to z_0$. Here z_0 is some point in the domain G and σ is some real number. For the Carleman-Vekua equation with the polar singularities the situation is essentially different. It is known that there exists a sufficiently wide class of equations permitting only trivial solutions in the domain $G\setminus\{z_0\}$ and satisfying the asymptotic conditions $O\left(|z-z_0|^{\sigma}\right), z \to z_0$, where z_0 is the point of polar singularity of the equation (C-V), σ is a real number. Therefore it makes no sense to consider the boundary value problems in this class. On the other hand if there are no restrictions on the solutions in the neighborhood of the singular point z_0 then it may occur that the homogeneous boundary problem has an infinite number of linearly independent solutions.

In this work for the Riemann-Hilbert problem the Noetherity conditions in particular like asymptotic conditions $O\left(\exp\left\{\delta_0 |z-z_0|^{-\sigma_0}\right\}\right), z \to z_0$ for a sufficiently wide class of the Carleman-Vekua equations are obtained. Here the constant parameters δ_0, σ_0 are uniquely defined by means of the coefficients of the equation, are independent from the given boundary functions and characterize the polar singularities of the coefficients. These asymptotic conditions are in some sense exact since if we seek the solution of the Riemann-Hilbert problem in the class satisfying the asymptotic condition $O\left(\exp\left\{\delta |z-z_0|^{-\sigma}\right\}\right), z \to z_0$, and if at least one from the equalities $\delta = \delta_0, \sigma = \sigma_0$ is not fulfilled then either the homogeneous problem has infinite number of linearly independent solutions or the non-homogeneous problem isn't solvable for any right-hand side.

In Sect. 2 the above mentioned asymptotic conditions are obtained; the general representation of the solutions of the Carleman-Vekua equations with the polar singularities satisfying these conditions are constructed. By means of these results the Riemann-Hilbert problem is correctly posed and is completely investigated in Sect. 3.

2 The Carlemann-Vekua Equations with the Polar Singularities

Let G be a bounded complex domain with the boundary Γ consisting from closed non-intersecting Liapunov smooth $\Gamma_0, \Gamma_1, \ldots, \Gamma_m$ contours and Γ_0 covers all the rest. Let G^* be some finite subset of the set G,

$$G^* = \{z_1, z_2, \ldots, z_N\}, N \geq 1.$$

Consider the Carlemann-Vekua equation

$$\frac{\partial w}{\partial \bar{z}} + A(z)w + B(z)\bar{w} = 0, \tag{1}$$

in the domain G, provided that the coefficient $B(z) \in L_p(G), p > 2$ and the coefficient $A(z)$ admits the following representation

$$A(z) = \overline{g(z)} + \sum_{k=1}^{N} \frac{A_k(z)}{|z - z_k|^{\nu_k}} \tag{2}$$

where the function $g(z)$ is holomorphic in $G \backslash G^*$ and has continuous boundary value on Γ; the function $A_k(z)$ admits the following representation

$$A_k(z) = a_k(z) \exp\{i n_k \arg(z - z_k)\}, \tag{3}$$

where

$$\frac{a_k(z) - \lambda_k}{|z - z_k|^{\nu_k}} \in L_p(G), p > 2;$$

the constants λ_k, ν_k, n_k are correspondingly complex, positive and entire numbers for every $k = 1, 2, \ldots, N$ (cf. [2, 3]).

Under the solution of Eq. (1) is understood the continuous generalized solution in $G \backslash G^*$; denote by $\mathcal{R}(A, B, G \backslash G^*)$ the set of all possible such solutions (cf. [6]).

Everywhere below the fulfillment of the following condition

$$\lambda_k \neq 0; |n_k - 1| > 2(\nu_k - 1) > 0, k = 1, 2, \ldots, N \tag{4}$$

is assumed.

From (2) we have that the coefficient $A(z)$ has the polar singularities of the form $|z - z_k|^{-\nu_k}$ and the singularities of the function $\overline{g(z)}$ in the points z_k. Below it will be established that the structure of the solutions of the Carleman-Vekua equation depends on the relationship between the parameters λ_k, ν_k, n_k. Generally speaking,

if the required conditions do not hold, then the assumptions proved below are not valid.

The following notations we need below

$$q_k = \mathrm{Res}_{z=z_k'}^{g(z)}, k = 1, 2, \ldots, N; \; Q_k = \frac{1}{2\pi i} \int_{\Gamma_k} g(t)dt, k = 1, 2, \ldots, m$$

Introduce an auxiliary function given by the formula

$$f(z) = \int_{\Gamma_{\zeta_0, z}} \overline{g(t)dt} - \sum_{k=1}^{m} \overline{Q_k} \log(\bar{z} - \overline{\tau_k}) - \sum_{k=1}^{N} \overline{q_k} \log(\bar{z} - \overline{z_k}),$$

in the domain $G \backslash G^*$, where ζ_0 is some fixed point in $G \backslash G^*$; $\Gamma_{\zeta_0, z}$ is a smooth contour connecting the points ς_0, z and lying in $G \backslash G^*$; τ_k is an arbitrary fixed point inside the contour Γ_k, $k = 1, 2, \ldots, m$. Consider also the function

$$F(z) = \Lambda(z) \exp\{2i \, Imf(z)\} \chi(z), z \in G \backslash G^*, \tag{5}$$

where

$$\Lambda(z) = \prod_{k=1}^{N} (z - z_k)^{-[2Req_k]} \tag{6}$$

$$\chi(z) = \exp\left\{ 2 \sum_{k=1}^{m} Q_k \log|z - \tau_k| + 2 \sum_{k=1}^{N} q_k \log|z - z_k| \right\}. \tag{7}$$

It follows from the conditions (4) that $2 - v_k - n_k \neq 0$, $k = 1, 2, \ldots, N$ and therefore by the formulas

$$\delta_k^* = \frac{2\lambda_k}{2 - v_k - n_k}, k = 1, 2, \ldots, N \tag{8}$$

the definite non-zero numbers are given. Assume

$$R(z) \equiv \sum_{k=1}^{N} \frac{\delta_k^*}{|z - z_k|^{v_{k-1}}} \cdot \exp\{i \, (n_k - 1) \cdot \arg(z - z_k)\}, \tag{9}$$

$$\Psi(z) \equiv F(z) \exp\{R(z)\}. \tag{10}$$

Consider the following Carleman-Vekua equation

$$\frac{\partial w_*}{\partial \bar{z}} + A_*(z)w_* + B_*(z)\overline{W_*} = 0, \tag{11}$$

where

$$A_*(z) = \sum_{k=1}^{N} \frac{a_k(z) - \lambda_k}{|z - z_k|^{\upsilon_k}} e^{in_k \arg(z - z_k)}, \quad B_*(z) = \frac{B(z)\Psi(z)}{\overline{\Psi(z)}}.$$

It is easy to see, that $A_*(z), B_*(z) \in L_p(G)$, $p > 2$ and hence (11) is the regular Carleman-Vekua equation.

The following theorem takes place.

Theorem 1 *By the following relation*

$$w_*(z) = \Psi(z)w(z), z \in G\backslash G^* \left(w_* \in \Re\left(A_*, B_*, G\backslash G^*\right), w \in \Re\left(A, B, G\backslash G^*\right)\right), \tag{12}$$

the bijective correspondence between the classes $\Re(A, B, G\backslash G^*)$ *and* $\Re(A_*, B_*, G\backslash G^*)$ *is established.*

Proof One can check directly the following equalities

$$\frac{\partial F(z)}{\partial \bar{z}} = F(z) \cdot \overline{g(z)},$$

$$\frac{\partial \exp\{R(z)\}}{\partial \bar{z}} = \exp\{R(z)\} \sum_{k=1}^{N} \frac{\lambda_k}{|z - z_k|^{\upsilon_k}} \cdot \exp\{in_k \cdot \arg(z - z_k)\},$$

$$\frac{\partial \Psi(z)}{\partial \bar{z}} = \Psi(z) \left[\sum_{k=1}^{N} \frac{\lambda_k}{|z - z_k|^{\upsilon_k}} \cdot \exp\{in_k \cdot \arg(z - z_k)\} + \overline{g(z)}\right].$$

□

It is clear that by means of the relation (12) the bijective correspondence is also established between the classes

$$\Re\left(A_*, B_*, G\backslash G^*\right) \cap C\left(\bar{G}\backslash G^*\right), \Re\left(A, B, G\backslash G^*\right) \cap C\left(\bar{G}\backslash G^*\right).$$

Let $\delta = (\delta_1, \delta_2, \ldots, \delta_N)$ and $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_N)$ are given N-dimensional vectors with the nonnegative components. Denote by $\Omega_0[\delta, \sigma]$ the class of all possible functions from the set $\Re(A, B, G\backslash G^*)$ satisfying the condition

$$w(z) = O\left(\exp\left\{\delta_k |z - z_k|^{-\sigma_k}\right\}\right), z \to z_k, k = 1, 2, \ldots, N. \tag{13}$$

Denote by $\overline{\Omega_0}[\delta, \sigma]$ the class of all possible functions from the set $\Omega_0[\delta, \sigma]$ admitting the continuous extension in $\left(\bar{G} \backslash G^*\right)$. By δ^* and v^* the following vectors are denoted

$$\delta^* \equiv \left(\left|\delta_1^*\right|, \left|\delta_2^*\right|, \ldots, \left|\delta_N^*\right|\right), v^* \equiv (v_1 - 1, v_2 - 1, \ldots, v_N - 1).$$

The class of the solutions $\Omega_0\left[\delta^*, v^*\right]$ is very important class in what follows.

The following theorem holds.

Theorem 2 *If for some k the inequality $\delta_k < \left|\delta_k^*\right|$ is fulfilled then the class $\Omega_0\left[\delta, v^*\right]$ is a trivial class (i.e., it contains only zero functions).*

The proof of the Theorem 2 follows from [5] and [1].

Theorem 2 directly implies that if for some k the inequality $\sigma_k < v_k - 1$ is valid, then for every vector δ the class $\Omega_0[\delta, \sigma]$ is a trivial class.

Therefore, if $\sigma_k = v_k - 1, k = 1, 2, \ldots, N$ and for some k_0 the inequality $\delta_{k_0} < \left|\delta_{k_0}^*\right|$ is fulfilled or if for some k_0 the inequality $\sigma_{k_0} < v_{k_0} - 1$ is fulfilled then the class $\Omega_0[\delta, \sigma]$ is a trivial class.

It is natural to investigate the class $\Omega_0[\delta, \sigma]$. The following theorem gives us the representation of the solution of this class.

Theorem 3 *By means of the relation (13) the bijective correspondence between the classes $\Omega_0\left[\delta^*, v^*\right]\left(\overline{\Omega_0}\left[\delta^*, v^*\right]\right), \Re\left(A_*, B_*, G\right)\left(\Re\left(A_*, B_*, G\right) \cap C(\bar{G})\right)$ is established.*

3 Investigation of the Riemann-Hilbert Boundary Value Problem

In order to pose correctly the Riemann-Hilbert boundary value problem it is clear from the above mentioned results that it is sufficient to require from the solution of Eq. (1) the fulfillment of the asymptotic condition of the form

$$w(z) = O\left(\exp\left\{\left|\delta_k^*\right| \cdot |z - z_k|^{-v_k+1}\right\}\right), z \to z_k, k = 1, 2 \ldots, N \qquad (*).$$

In the present section the proof of sufficiency of asymptotic condition $(*)$ is given and the boundary value problems are investigated. Consider the following boundary value problem: on the boundary Γ the Hölder continuous functions $\lambda(t)$ and $\gamma(t)$ are given, $\gamma(t)$ is a real function and $|\lambda(t)| = 1$; find the function $w(z) \in \overline{\Omega_0}[\delta, \sigma]$ satisfying the boundary equation

$$Re\{\overline{\lambda(t)} w(t)\} = \gamma(t), \quad t \in \Gamma. \qquad (14)$$

From the Theorem 2 it follows that in case $\gamma(t) \neq 0$ the problem (14) isn't solvable in the class $\overline{\Omega_0}[\delta, \sigma]$ if $\sigma_k = \upsilon_k - 1$ for some k or if $\sigma_k = \upsilon_k - 1$, $k = 1, 2, \ldots, N$, but $\delta_k < \left|\delta_k^*\right|$ for some k.

Let $\delta_k \neq 0, \sigma_k \geq \upsilon_k - 1, k = 1, 2, \ldots, N$. Denote by $\mathcal{H}$ the set of all possible values k of the index for which $\delta_k < \left|\delta_k^*\right|$.

The following theorem takes place.

Theorem 4 *The homogeneous boundary value problem (14), $(\gamma(t) = 0)$, in the class $\overline{\Omega_0}[\delta, \sigma]$ has infinite number of linearly independent solutions if and only if when one from the following conditions is fulfilled:*

1. $\mathcal{H} = \emptyset; \sum_{k=1}^{N} (\delta_k + \sigma_k) > \sum_{k=1}^{N} \left(\left|\delta_k^*\right| + \upsilon_k - 1\right)$;
2. $\mathcal{H} = \emptyset; \sigma_k > \upsilon_k - 1, k \in \mathcal{H}$.

Proof Let the first condition (1) be fulfilled. Then for all $k = 1, 2, \ldots, N$ the following inequalities hold

$$\delta_k \geq \left|\delta_k^*\right| + \upsilon_k - 1,$$

and there exists at least one $k = k_0$ for which the strict inequality is fulfilled

$$\delta_{k_0} + \sigma_{k_0} > \left|\delta_{k_0}^*\right| + \upsilon_{k_0} - 1, \tag{15}$$

From this last inequality follows that one from the inequalities $\delta_{k_0} \geq \left|\delta_{k_0}^*\right|, \sigma_{k_0} \geq \upsilon_{k_0} - 1$ is also strict. Let $\delta_{k_0} > \left|\delta_{k_0}^*\right|$ and let us fix an arbitrary number $S \geq 0$. Consider the solutions $w_*(z)$ from the class $\Re(A_*, B_*, G \backslash G^*)$ which are representable in the form

$$w_*(z) = \frac{\Phi(z)}{(z - z_{k_0})^s} \exp(\omega(z)), \tag{16}$$

where $\Phi(z)$ is a function, holomorphic in G and continuous in $\bar{G}$. It is easy to see that every function of the form (16) defines the solution $w(z)$ of Eq. (1) of the class $\overline{\Omega_0}[\delta, \sigma]$ by means of the relation (12). Further on, we can see that by the relation

$$w_*(z) = \frac{\omega_0(z)}{(z - z_{k_0})^s} \tag{17}$$

the bijective correspondence between the class of all functions of the form (16) and the class

$$\Re\left(A_*, B_*, \left(\frac{z - z_{k_0}}{z - \overline{z_{k_0}}}\right)^S, G\right) \cap C(\bar{G})$$

of the solutions of the equation

$$\frac{\partial \omega_0}{\partial z} + A_*(z)\omega_0 + B_* \left(\frac{z - z_{k_0}}{\bar{z} - \overline{z_{k_0}}} \right)^s \overline{\omega_0} = 0. \tag{18}$$

is established.

Together with the problem (14) consider the following boundary value problem: find the generalized solution of the problem (18) continuous in G and satisfying the boundary condition

$$\operatorname{Re} \left\{ \frac{\overline{\lambda(t)}}{\left(t - z_{k_0}\right)^s \Psi(t)} \omega_0(t) \right\} = 0, \quad t \in \Gamma. \tag{19}$$

It is clear that by the formulas (17), (12) every system of linearly independent solutions of the problem (19) defines the system of linearly independent solutions of homogeneous problem (14). On the other hand the number of linearly independent solutions of the problem (19) l satisfies the inequality

$$l \geq 2 \operatorname{ind} \left(\frac{\lambda(t)}{\left(\bar{t} - \overline{z_0}\right)^{s\top}(t)} \right) - m + 1, \tag{20}$$

by virtue of which we get

$$l \geq 2 \left(\operatorname{ind} \lambda(t) + S + \sum_{k=1}^{N} \left[2 \operatorname{Re} q_k \right] \right) - m + 1.$$

From here it follows that for an appropriate choice of S the number l will be arbitrarily large and therefore the homogeneous problem (14) has infinite number of linearly independent solutions. One can prove similarly that the set of linearly independent solutions of the homogeneous problem (14) is infinite in case $\sigma_{k_0} > v_{k_0} - 1, \delta_{k_0} = \left| \delta^*_{k_0} \right|$. Hence we obtain that if the condition (1) is fulfilled then the homogeneous problem (14) has the infinite number of linearly independent solutions.

Let now the condition (2) be fulfilled. Then on the basis of the relation

$$\overline{\Omega_0} \left[\delta^{(1)}, \sigma^{(1)} \right] \subset \overline{\Omega_0} \left[\delta^{(2)}, \sigma^{(2)} \right],$$

which follows directly from the conditions

$$\delta_k^{(2)} \neq 0, k = 1, 2, \ldots, N; \sigma_k^{(2)} > \sigma_k^{(1)}, k \in \mathcal{H},$$

we get that the homogeneous problem (14) has infinite number of linearly independent solutions. The sufficiency of one of the conditions (1), (2) is proved. Let us prove the necessity.

Let the homogeneous problem (14) has infinite number of linearly independent solutions and the set $\mathcal{H} \neq \emptyset$ then for every $k \in \mathcal{H}$ we have $\sigma_k > \upsilon_k - 1$. Indeed, if for at least one $k_0 \in \mathcal{H}$ we have $\sigma_{k_0} > \upsilon_{k_0} - 1$ then on the basis of the Theorem 2 the class $\overline{\Omega_0}[\delta, \sigma]$ would consist from only zero elements; we get a contradiction and so when $\mathcal{H} \neq \emptyset$ the condition (2) is fulfilled. Let now $\mathcal{H} = \emptyset$, then prove that

$$\sum_{k=1}^{N} (\delta_k + \sigma_k) > \sum_{k=1}^{N} \left(\left|\delta_k^*\right| + \upsilon_k - 1\right) \tag{21}$$

Indeed, otherwise it would be

$$\sum_{k=1}^{N} (\delta_k + \sigma_k) = \sum_{k=1}^{N} \left(\left|\delta_k^*\right| + \upsilon_k - 1\right) \tag{22}$$

and therefore $\delta_k = \left|\delta_k^*\right|, \sigma_k = \upsilon_k - 1, k = 1, 2, \ldots, N$.

Hence we will obtain that the homogeneous problem (14) has infinite number of linearly independent solutions in the class $\overline{\Omega_0}\left[\delta^*, \upsilon^*\right]$, but this problem has finite number of linearly independent solutions. Indeed, by virtue of the Theorem 3, using the relation (12), the bijective correspondence is established between the solutions of the problem (14) and the following boundary value problem: find the generalized solution of the equation

$$\frac{\partial w_*}{\partial \bar{z}} + A_*(z) w_*(t) + B_* \overline{w_*} = 0, \tag{23}$$

continuous in the domain $\bar{G}$ satisfying the boundary condition

$$Re\left\{\frac{\overline{\lambda(t)}}{\Psi(t)} w_*(t)\right\} = \gamma(t), t \in \Gamma. \tag{24}$$

The homogeneous problem (23) and (24), ($\gamma(t) = 0$), on the basis of [6] has finite number of linearly independent solutions. That is why the homogeneous problem (14) has finite number of linearly independent solutions in the class $\overline{\Omega_0}\left[\delta^*, \upsilon^*\right]$. Therefore the condition (22) isn't fulfilled and thus (21) is fulfilled. Theorem 14 is completely proved. □

Consider the boundary value problem: find the generalized solution continuous in the class $\bar{G}$ of the equation

$$\frac{\partial w_*'}{\partial \bar{z}} - A_*(z) w_*'(t) - \overline{B_*(z) w_*'} = 0, \tag{25}$$

satisfying the boundary condition

$$Re\left\{\frac{\lambda(t)}{\Psi(t)}t'(s)w'_*(t)\right\} = 0, t \in \Gamma. \tag{26}$$

It is easy to see that the number of linearly independent solutions of the problem (26), l_* is finite and it is clear that for the problem (23) to be solvable it is necessary and sufficient the fulfillment of the following equation

$$\int_\Gamma \frac{\lambda(t)}{\overline{\Psi(t)}}\gamma(t)w'_*(t)dt = 0 \tag{27}$$

for every solution of the problem (26).

On the basis of above obtained results the following theorem becomes evident.

Theorem 5 *The homogeneous problem (14) in the class* $\overline{\Omega_0}\left[\delta^*, v^*\right]$ *has finite number of linearly independent solutions and the non-homogeneous problem is solvable if and only if the condition (27) is fulfilled.*

Let l be a number of linearly independent solutions of the homogeneous problem (14). By means of the following evident equality

$$ind\left(\frac{1}{\overline{\Psi(t)}}\right) = \sum_{k=1}^{N}[2Req_k]$$

it follows the validity of the next theorem.

Theorem 6 *The following condition*

$$l - l_* = 2n + 2\sum_{k=1}^{N}[2Req_k] - m + 1$$

takes place, where q_k *is a residue of the function* $q(z)$ *at a point* z_k.

From the results obtained above in particular we get the theorem.

Theorem 7 *For the problem (14) to be Noetherian in the class* $\overline{\Omega_0}[\delta, \sigma]$ *it is necessary and sufficient the fulfillment of the condition*

$$\delta = \delta^*, \sigma = v^*.$$

Based on all the above, it can be said that for a sufficiently wide class of singular elliptic systems, singularity is significant for the correct setting of boundary problems as well as for their analysis.

References

1. Akhalaia, G., Giorgadze, G., Jikia, V., Makatsaria, G., Manjavidze, N.: Elliptic systems on Riemann surfaces. Tbilisi Int. Center Math. Inf. **13**, 1–154 (2012)
2. Begehr, H., Dai, D.-Q.: On the theory of a singular Vekua system. Oper. Theory Adv. Appl. **121**, 27–35 (2001)
3. Begehr, H., Dai, D.-Q.: On continuous solutions of a generalized Cauchy-Riemann system with more than one singularity. J. Differential Equations **196**, 67–90 (2004)
4. Gilbert, R.P., Buchanan, J.L.: First order elliptic systems. A function theoretic approach. In: Mathematics in Science and Engineering. Academic Press, Orlando (1983)
5. Makatsaria, G.: Singular points of solutions of some elliptic systems on the plane. J. Math.Sci. **160**(6), 737–744 (2009)
6. Vekua, I.: Generalized Analytic Functions. Pergamon, Oxford (1962)

Second Order Differential Operators Associated to the Space of Holomorphic Functions

Gian Rossodivita and Carmen Judith Vanegas

Abstract Let $\mathcal{F}$ be a given differential operator, then a function space $\mathcal{X}$ is called an associated space to $\mathcal{F}$ if $\mathcal{F}$ transforms $\mathcal{X}$ into itself. In this work we show the construction of all operators of second order with complex coefficients that are associated with the space of holomorphic functions. As an application the solvability of initial value problems involving these operators is shown.

1 Introduction

We say that a function space $\mathcal{X}$ is called an associated space to a given differential operator $\mathcal{F}$ if $\mathcal{F}$ transforms $\mathcal{X}$ into itself or we say that a pair $\mathcal{F}, \mathcal{G}$ of differential operators are associated in case $\mathcal{F}$ transforms solutions u of $\mathcal{G}u = 0$ again into solutions of this equation.

Associated spaces are used to solve initial value problems of the type

$$\partial_t u = \mathcal{F}(t, x, u, \partial_j u), \quad j = 0, \ldots, n, \tag{1}$$

$$u(0, x) = \varphi(x), \tag{2}$$

where $\varphi(x)$ satisfies the partial differential equation $\mathcal{G}(u) = 0$, provided that the associated space $\mathcal{X}$ of $\mathcal{F}$ contains all the solutions for $\mathcal{G}(u) = 0$, and that the elements of $\mathcal{X}$ satisfy an interior estimate, i.e., an estimate for the derivatives of the solutions near the boundary of a certain bounded domain (see [6]).

G. Rossodivita
Universidad Católica del Norte, Antofagasta, Chile
e-mail: gian.rossodivita@ucn.cl

C. J. Vanegas (✉)
Universidad Técnica de Manabí, Ave Urbina, Portoviejo, Ecuador

Universidad Yachay Tech, Urcuquí, Ecuador
e-mail: carmen.vanegas@utm.edu.ec; cvanegas@yachaytech.edu.ec

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_18

There are two basic problems in the theory of associated spaces. The first one is the direct problem, which consists on the construction of an associated space $\mathcal{X}$ to a given operator $\mathcal{F}$. In other words, $\mathcal{F}$ is given and one has to determine an associated equation $\mathcal{G}(u) = 0$ in which the initial value problem is solvable. The second one is the inverse problem, $\mathcal{G}$ is given and one determines all $\mathcal{F}$ for which solutions φ of $\mathcal{G}(u) = 0$ are admissible initial functions. In this article, we have worked with the inverse problem: we showed a characterization of all linear second order complex partial differential operators with complex coefficients that are associated with the space of holomorphic functions.

Necessary and sufficient conditions for evolution operators transforming holomorphic functions into themselves are given in [4], and [1] in the framework of complex analysis and elliptic complex analysis, respectively. Sufficient conditions for evolution operators transforming generalized analytic functions into themselves are given in [5]. Necessary and sufficient conditions for first order differential operators to be associated to the space of elliptic generalized analytic functions are given in [3]. In the framework of Clifford analysis we find in [7], necessary and sufficient conditions for linear first order partial differential operators $\mathcal{F}$ with coefficients of Clifford values, to be associated to the meta-q-monogenic operator:

$$\mathcal{D}_{(q,\lambda)} = \sum_{i=0}^{n} q_i \partial_i + \lambda \,, \tag{3}$$

where $q_0 = 1$, $q_i \in \mathcal{A}_n$, $i = 0, 1, 2, \ldots, n$ are constants and $\lambda \in \mathbb{R}$.

The results in this article are the first in the direction of considering associated operators of higher order. As an application, we show the solvability of initial value problems involving such operators associated to the space of holomorphic functions.

2 Associated Spaces

Definition 2.1 ([6]) Let $\mathcal{F}$ be a given differential operator depending on t, x, u and $\partial_i u$ for $i = 0, 1, \cdots, n$, while $\mathcal{G}$ is a differential operator with respect to the spacelike variables x_i with coefficients not depending on time t. $\mathcal{F}$ is said to be associated with $\mathcal{G}$ if $\mathcal{F}$ maps solutions for the differential equation $\mathcal{G}u = 0$ into solutions of the same equation for a fixedly chosen t, i.e.,

$$\mathcal{G}u = 0 \Rightarrow \mathcal{G}(\mathcal{F}u) = 0.$$

The function space $\mathcal{X}$ containing all the solutions for the differential equation $\mathcal{G}u = 0$ is called an associated space of $\mathcal{F}$. □

Next we will determine necessary and sufficient conditions such that an second order operator $\mathcal{F}$ be associated to the Cauchy-Riemann operator.

2.1 Necessary and Sufficient Conditions on the Coefficients of $\mathcal{F}$

We consider the following second order differential operator:

$$\begin{aligned}\mathcal{F}_2\,\omega = A_2(z)\partial_z^2\omega + B_2(z)\partial_{z\bar{z}}^2\omega + C_2(z)\partial_{\bar{z}}^2\omega + D_2(z)\overline{\partial_z^2\omega}\\ +E_2(z)\overline{\partial_{z\bar{z}}^2\omega} + F_2(z)\overline{\partial_{\bar{z}}^2\omega} + \mathcal{F}_1\omega,\end{aligned}$$

where $\mathcal{F}_1$ is defined by

$$\begin{aligned}\mathcal{F}_1\,\omega = A_1(z)\partial_z\omega + B_1(z)\overline{\partial_z\omega} + C_1(z)\partial_{\bar{z}}\omega + D_1(z)\overline{\partial_{\bar{z}}\omega}\\ +E_1(z)\,\omega + F_1(z)\,\overline{\omega} + G_1(z),\end{aligned}$$

and all coefficients of $\mathcal{F}_2$ and $\mathcal{F}_1$ are complex valued.

We will determine conditions over the coefficients of $\mathcal{F}_2$ such that

$$\partial_{\bar{z}}\omega = 0 \;\Rightarrow\; \partial_{\bar{z}}(\mathcal{F}_2\omega) = 0,$$

So for an arbitrary holomorphic function ω we have:

$$\begin{aligned}\mathcal{F}_2\,\omega = A_2(z)\partial_z^2\omega + D_2(z)\overline{\partial_z^2\omega}\\ +A_1(z)\partial_z\omega + B_1(z)\overline{\partial_z\omega} + E_1(z)\,\omega + F_1(z)\,\overline{\omega} + G_1(z).\end{aligned} \tag{1}$$

Then assuming that the coefficients of (1) are continuously differentiable with respect to z and $\bar{z}$ and applying the Cauchy-Riemann operator to $\mathcal{F}_2\,\omega$, we get

$$\begin{aligned}\partial_{\bar{z}}(\mathcal{F}_2\omega) = D_2(z)\overline{\partial_z{}^3\omega} + \partial_{\bar{z}}A_2(z)\partial_z{}^2\omega + (\partial_{\bar{z}}D_2(z) + B_1(z))\,\overline{\partial_z{}^2\omega}\\ +\partial_{\bar{z}}A_1(z)\partial_z\omega + (F_1(z) + \partial_{\bar{z}}B_1(z))\,\overline{\partial_z\omega} + \partial_{\bar{z}}E_1(z)\,\omega\\ +\partial_{\bar{z}}F_1(z)\,\overline{\omega} + \partial_{\bar{z}}G_1(z).\end{aligned} \tag{2}$$

Therefore $\partial_{\bar{z}}(\mathcal{F}_2\omega) = 0$ if the following sufficient conditions are satisfied:

$$\begin{aligned}D_2(z) = 0,\;\; B_1(z) = 0,\;\; F_1(z) = 0,\\ \partial_{\bar{z}}A_2(z) = 0,\;\; \partial_{\bar{z}}A_1(z) = 0,\, \partial_{\bar{z}}E_1(z) = 0,\;\text{and}\;\partial_{\bar{z}}G_1(z) = 0.\end{aligned} \tag{3}$$

Thus second-order operators of the form

$$\mathcal{F}\omega = A_2(z)\partial_z^2\omega + A_1(z)\partial_z\omega + E_1(z)\,\omega + G_1(z), \tag{4}$$

with the coefficients $A_2(z),\ A_1(z),\ E_1(z)\ and\ G_1(z)$ as holomorphic functions, are associated to the Cauchy-Riemann operator.

Now we assume that $(\mathcal{F}_2, \partial_{\overline{z}})$ is an associated pair, i.e., $\partial_{\overline{z}}(\mathcal{F}_2\omega) = 0$ if only ω is a holomorphic function.

In order to obtain the conditions on the coefficients of operator $\mathcal{F}_2$ given by (1) we will start by choosing special functions of the associated space, in this case special holomorphic functions, and we will write out the relations assuming that $\mathcal{F}_2$ is holomorphic for those functions.

Then choosing the function $\omega = 0$ in

$$\begin{aligned}\partial_{\overline{z}}(\mathcal{F}_2\omega) = {} & D_2(z)\overline{\partial_z{}^3\omega} + \partial_{\overline{z}}A_2(z)\partial_z{}^2\omega + T_1(z)\overline{\partial_z{}^2\omega} \\ & + \partial_{\overline{z}}A_1(z)\partial_z\omega + T_2(z)\,\overline{\partial_z\omega} + \partial_{\overline{z}}E_1(z)\,\omega \\ & + \partial_{\overline{z}}F_1(z)\,\overline{\omega} + \partial_{\overline{z}}G_1,\end{aligned}$$

where $T_1(z) = \partial_{\overline{z}}D_2(z) + B_1(z)$ and $T_2(z) = F_1(z) + \partial_{\overline{z}}B_1(z)$, we obtain $\partial_{\overline{z}}(\mathcal{F}_2\omega) = \partial_{\overline{z}}G_1 = 0$ and so G_1 is holomorphic and the term $\partial_{\overline{z}}G_1$ can be omitted from $\partial_{\overline{z}}(\mathcal{F}_2\omega)$.

We now choose the functions $\omega = 1$ and $\omega = i$ in

$$\begin{aligned}\partial_{\overline{z}}(\mathcal{F}_2\omega) = {} & D_2(z)\overline{\partial_z{}^3\omega} + \partial_{\overline{z}}A_2(z)\partial_z{}^2\omega + T_1(z)\overline{\partial_z{}^2\omega} \\ & + \partial_{\overline{z}}A_1(z)\partial_z\omega + T_2(z)\,\overline{\partial_z\omega} + \partial_{\overline{z}}E_1(z)\,\omega \\ & + \partial_{\overline{z}}F_1(z)\,\overline{\omega},\end{aligned}$$

to get

$$\partial_{\overline{z}}E_1(z) + \partial_{\overline{z}}F_1(z) = 0, \quad \partial_{\overline{z}}E_1(z) - \partial_{\overline{z}}F_1(z) = 0,$$

which implies $\partial_{\overline{z}}E_1(z) = \partial_{\overline{z}}F_1(z) = 0$ and $\partial_{\overline{z}}\mathcal{F}_2(\omega)$ reduces to

$$\begin{aligned}\partial_{\overline{z}}(\mathcal{F}_2\omega) = {} & D_2(z)\overline{\partial_z{}^3\omega} + \partial_{\overline{z}}A_2(z)\partial_z{}^2\omega + T_1(z)\overline{\partial_z{}^2\omega} \\ & + \partial_{\overline{z}}A_1(z)\partial_z\omega + T_2(z)\,\overline{\partial_z\omega}.\end{aligned}$$

Next we choose the holomorphic functions $\omega = z$, and $\omega = iz$. For these functions we obtain from

$$\begin{aligned}\partial_{\overline{z}}(\mathcal{F}_2\omega) = {} & D_2(z)\overline{\partial_z{}^3\omega} + \partial_{\overline{z}}A_2(z)\partial_z{}^2\omega + T_1(z)\overline{\partial_z{}^2\omega} \\ & + \partial_{\overline{z}}A_1(z)\partial_z\omega + T_2(z)\,\overline{\partial_z\omega}\end{aligned}$$

the equations

$$\partial_{\overline{z}}A_1(z) + T_2(z) = 0,\ \ \partial_{\overline{z}}A_1(z) - T_2(z) = 0$$

implying $\partial_{\bar{z}} A_1(z) = T_2(z) = 0$ and so

$$\partial_{\bar{z}}(\mathcal{F}_2\omega) = D_2(z)\overline{\partial_z{}^3\omega} + \partial_{\bar{z}} A_2(z)\partial_z{}^2\omega + T_1(z)\overline{\partial_z{}^2\omega}$$

Choosing the holomorphic functions $\omega = z^2$, and $\omega = iz^2$, we have that

$$\partial_{\bar{z}}(\mathcal{F}_2\omega) = D_2(z)\overline{\partial_z{}^3\omega} + \partial_{\bar{z}} A_2(z)\partial_z{}^2\omega + T_1(z)\overline{\partial_z{}^2\omega}$$

implies

$$\partial_{\bar{z}} A_2(z) + T_1(z) = 0 \quad \text{and} \quad \partial_{\bar{z}} A_2(z) - T_1(z) = 0,$$

which in turn implies $\partial_{\bar{z}} A_2(z) = T_1(z) = 0$ and so $\partial_{\bar{z}}(\mathcal{F}_2\omega) = D_2(z)\overline{\partial_z{}^3\omega}$.

Finally taking $\omega = z^3$ in the above equation, we have $D_2(z) = 0$.

Since $T_1(z) = \partial_{\bar{z}} D_2(z) + B_1(z)$ and $T_2(z) = F_1(z) + \partial_{\bar{z}} B_1(z)$, we get $B_1(z) = 0$ and then $F_1(z) = 0$.

Therefore the following statement is true:

Theorem *Suppose D_2, B_1, F_1, A_2, A_1, E_1 and G_1 are continuously differentiable with respect to z and $\bar{z}$. Then second order partial differential operators given by*

$$\begin{aligned}\mathcal{F}_2\,\omega = {} & A_2(z)\partial_z^2\omega + B_2(z)\partial_{z\bar{z}}^2\omega + C_2(z)\partial_{\bar{z}}^2\omega + D_2(z)\overline{\partial_z^2\omega} \\ & + E_2(z)\overline{\partial_{z\bar{z}}^2\omega} + F_2(z)\overline{\partial_{\bar{z}}^2\omega} \\ & + A_1(z)\partial_z\omega + B_1(z)\overline{\partial_z\omega} + C_1(z)\partial_{\bar{z}}\omega + D_1(z)\overline{\partial_{\bar{z}}\omega} \\ & + E_1(z)\,\omega + F_1(z)\,\overline{\omega} + G_1(z),\end{aligned}$$

are associated with the Cauchy-Riemann operator if and only if the following conditions are satisfied:

$$\begin{aligned} D_2(z) = 0,\ B_1(z) = 0,\ F_1(z) = 0, \\ \partial_{\bar{z}} A_2(z) = 0,\ \partial_{\bar{z}} A_1(z) = 0, \partial_{\bar{z}} E_1(z) = 0,\ \textit{and}\ \partial_{\bar{z}} G_1(z) = 0.\end{aligned}$$

3 Solution of Initial Value Problems via Associated Spaces

We consider the initial value problem

$$\partial_t \omega(t, z) = \mathcal{F}\omega(t, z) \tag{1}$$

$$\omega(0, z) = \varphi(z), \tag{2}$$

where $t \in [0, T]$ is the variable time, $z \in \mathbb{C}$, $\omega(t, z)$ is a complex-valued function and $\mathcal{F}\omega(t, z)$ is as in (4).

This problem can be rewritten as (see [2])

$$\omega(t, z) = \varphi(z) + \int_0^t \mathcal{F}\omega(\tau, z)d\tau. \tag{3}$$

Consequently, the solution of the initial value problem (1), (2) is a fixed point of the operator

$$T\omega(t, z) = \varphi(z) + \int_0^t \mathcal{F}\omega(\tau, z)d\tau. \tag{4}$$

and vice versa.

The existence and uniqueness of this problem can be showed using the contraction mapping principle. To apply such a principle, the operator (4) should map a certain Banach space B of holomorphic functions into itself. Since the operator $\mathcal{F}$ also depends on the derivatives with respect to z of ω, this map exists in case when the derivatives with respect to z of $(T\omega(t, z))$ do exist and can be estimated accordingly. Therefore, one has to restrict the operator to a space of holomorphic functions for which the derivatives with respect to z of a holomorphic function ω can be estimated by ω itself. Then the Lipschitz condition with respect to the function ω and their derivatives on $\mathcal{F}$ is necessary. This sought space is the so-called associated space and the estimates for the derivatives with respect to z of ω can be attained by using the so-called interior estimate.

Interior estimates can be obtained via integral representations using the Cauchy kernel.

In consequence, the method of associated operators is applied to solve initial value problems with initial holomorphic functions and we have the following theorem:

Theorem *Let $\mathcal{F}$ be the operator defined in Theorem 2.1. Suppose $\mathcal{F}$ and the operator $\partial_{\bar{z}}$ form an associated pair of operators, for each fixed $t \in [0, T]$, and the solutions of the corresponding equation $\partial_{\bar{z}}\omega = 0$ satisfy an interior estimate of first order. Then the initial value problem (1), (2) is solvable provided that the initial function is a holomorphic function.* □

4 Conclusions

We have given necessary and sufficient conditions on the coefficients of the operator $\mathcal{F}$ under which $\mathcal{F}$ is associated with the Cauchy-Riemann operator $\partial_{\bar{z}}$. It means that $\mathcal{F}$ transforms holomorphic functions into holomorphic functions, for a fixedly chosen t.

Using the equation $\partial_{\bar{z}}\omega = 0$ some derivatives could have been discarded from $\partial_{\bar{z}}(\mathcal{F}\omega)$, and the sufficient conditions for $\partial_{\bar{z}}(\mathcal{F}\omega) = 0$ could have been obtained by comparison of the coefficients. On the other side, by substituting special holomorphic functions, we showed that these conditions are also necessary. Therefore, we have found all linear second order operators of the given form, which are associated to $\partial_{\bar{z}}$ in $\mathbb{C}$.

Theorem 3 implies that each initial value problem (1) and (2) is solvable provided that the initial function is a holomorphic function, i.e. if it belongs to an associated space to $\mathcal{F}$. The technique of associated spaces allowed us to solve the initial value problem of type (1) and (2).

References

1. Alayón-Solarz, D., Vanegas, C.J.: Operators associated to the Cauchy-Riemann operator in elliptic complex numbers. Adv. Appl. Clifford Algebras **22**, 257–270 (2012)
2. Nagumo, M.: Über das Anfangswertproblem Partieller Differentialgleichungen. Jpn. J. Math. **18**, 41–47 (1941)
3. Rossodivita, G., Vanegas, C.J.: Associated Operators to the Space of Elliptic Generalized-Analytic Functions. New Trends in Analysis and Interdisciplinary Applications. Birkhäuser, Basel, pp. 135–141 (2017)
4. Son, L.H., Tutschke, W.: First Order differential operators associated to the Cauchy-Riemann equations in the plane. Complex Variables Elliptic Equations **48**, 797–801 (2003)
5. Tutschke, W.: Solution of Initial Value Problems in Classes of Generalized Analytic Functions. Teubner Leipzig and Springer-Verlag (1989)
6. Tutschke, W.: Associated spaces - a new tool of real and complex analysis. In: Function Spaces in Complex and Clifford Analysis. National University Publishers, Hanoi, pp. 253–268 (2008)
7. Vanegas, C.J., Vargas, F.A.: Associated Operators to the Space of Meta-q-Monogenic Functions. Clifford Analysis and Related Topics. CART 2014, Springer Proceedings in Mathematics & Statistics, vol. 260 (2018)

Constructional Method for a Non-local Boundary and Initial Problem Raised from a Free Boundary Model of Cancer

Jian-Rong Zhou, Heng Li, and Yongzhi Xu

Abstract In this paper we investigate a parabolic partial differential equation with non-local boundary condition motivated by ductal carcinoma in situ (DCIS) model. Approximation solution of the present problem is implemented by Ritz-Galerkin method. Numerical experiment shows that the method is effective and accurate.

1 Introduction

Ductal carcinoma in situ (DCIS) refers to a special diagnosis of breast cancer. In our earlier papers [11, 14], we introduced a free boundary problem model of DCIS and four kinds of inverse problems related to different diagnosis methods. For more information, see [9–14, 19]. Among them, clinical data of the third model is obtained by a sequence of tomograph, which is related to a boundary and initial problem of parabolic equation.

In this paper, we consider the following parabolic equation

$$\frac{\partial v}{\partial t} = \frac{\partial^2 v}{\partial x^2} + p(x,t), \quad 0 < x < 1, \quad 0 < t < 1, \tag{1}$$

with initial condition

$$v(x,0) = f(x), \quad 0 < x < 1, \tag{2}$$

J.-R. Zhou
Department of Mathematics, Foshan University, Foshan, Guangdong, China

H. Li
Department of Mathematics, Governors State University, University Park, IL, USA
e-mail: hli@govst.edu

Y. Xu (✉)
Department of Mathematics, University of Louisville, Louisville, KY, USA
e-mail: ysxu0001@louisville.edu

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_19

non-local boundary conditions

$$v(1,t) = g(t), \quad 0 < t < 1, \tag{3}$$

$$\int_0^{b(t)} v(x,t)dx = m(t), \quad 0 < t < 1, \quad 0 < b(t) < 1, \tag{4}$$

and compatibility conditions

$$f(1) = v(1,0) = g(0), \tag{5}$$

$$\int_0^{b(0)} f(x)dx = m(0) \tag{6}$$

The problem is to determine $v(x,t)$ for given $p(x,t)$, $f(x)$, $g(t)$, $b(t)$ and $m(t)$.

The presence of non-local boundary conditions can make the application of standard numerical methods complicated and affect the accuracy of result. Thus in this paper, we convert non-local boundary value problems to a desirable equivalent problem and solve it by using the Ritz-Galerkin method. This method is a powerful tool to solve differential equation by converting the non-linear problem into a set of linear equations. It has been widely used in many areas of mathematics, especially in the field of numerical analysis [2–4, 6, 15–18].

The remainder of this paper is organized as follows: In Sect. 2, we present equivalent forms of original problem. Then we introduce the properties of Bernstein polynomials in Sect. 3. The numerical schemes for the solution of equations are described in Sect. 4. Finally, one numerical experiment is exhibited in Sect. 5 to verify the accuracy and efficiency of the novel method.

2 Equivalent Problems

In this section, we introduce one transformation and an transition function $G(x,t)$ to convert our problem (1)–(6) to two equivalent forms.

Introduce the first transformation:

$$w(x,t) = v(x,t) - F(x,t), \tag{7}$$

where

$$F(x,t) = \frac{2x - b(t)}{2 - b(t)} g(t) + \frac{2m(t)(1-x)}{b(t)(2-b(t))}, \tag{8}$$

Under the first transformations (7), we obtain the first equivalent form of original problem(1)–(6) as following:

$$\frac{\partial w}{\partial t} = \frac{\partial^2 w}{\partial x^2} + K(x,t), \quad 0 < x < 1, \quad 0 < t < 1, \tag{9}$$

with initial condition

$$w(x,0) = \widetilde{f}(x), \quad 0 < x < 1, \tag{10}$$

boundary conditions

$$w(1,t) = 0, \quad 0 < t < 1, \tag{11}$$

$$\int_0^{b(t)} w(x,t)dx = 0, \quad 0 < t < 1, \tag{12}$$

and compatibility conditions

$$w(1,0) = \widetilde{f}(1) = 0, \tag{13}$$

$$\int_0^{b(0)} w(x,0)dx = \int_0^{b(0)} \widetilde{f}(x)dx = 0, \tag{14}$$

where

$$\begin{aligned} K(x,t) &= p(x,t) + \frac{\partial^2 F(x,t)}{\partial x^2} - \frac{\partial F(x,t)}{\partial t} \\ &= p(x,t) - \frac{(2x - b(t))(2 - b(t))g'(t) + 2g(t)b'(t)(x-1)}{(2-b(t))^2} \\ &\quad - \frac{2m'(t)b(t)(2-b(t))(1-x) - 4m(t)b'(t)(1-b(t))(1-x)}{(b(t))^2(2-b(t))^2}, \end{aligned} \tag{15}$$

$$\widetilde{f}(x) = f(x) - \frac{2x - b(0)}{2 - b(0)} g(0) - \frac{2m(0)(1-x)}{b(0)(2-b(0))}. \tag{16}$$

In order to convert non-local boundary condition to desirable form and apply Ritz-Galerkin method to it, we introduce transition function

$$G(x,t) = \int_{b(t)}^{x} w(s,t)ds + (x^2 - 2x) \cdot \int_0^1 w(s,t)ds, \tag{17}$$

then we have

$$\frac{\partial G(x,t)}{\partial x} = w(x,t) + 2(x-1)\int_0^1 w(s,t)ds, \tag{18}$$

$$\frac{\partial^2 G(x,t)}{\partial x^2} = \frac{\partial w(x,t)}{\partial x} + 2\int_0^1 w(s,t)ds, \tag{19}$$

$$\frac{\partial^3 G(x,t)}{\partial x^3} = \frac{\partial^2 w(x,t)}{\partial x^2}, \tag{20}$$

$$\frac{\partial w(x,t)}{\partial t} = \frac{\partial^2 G(x,t)}{\partial x \partial t} - 2(x-1)\int_0^1 \frac{\partial w(s,t)}{\partial t}ds. \tag{21}$$

$$\int_0^1 w(s,t)ds = \frac{G(b(t),t)}{b(t)(b(t)-2)}, \tag{22}$$

thus the second equivalent form of original problem is as follows:

$$\frac{\partial G^2}{\partial x \partial t} = \frac{\partial^3 G}{\partial x^3} + K(x,t) + 2(x-1)\cdot\frac{d}{dt}\left(\frac{G(b(t),t)}{b(t)(b(t)-2)}\right),\ 0<x<1,\ 0<t<1, \tag{23}$$

with initial condition

$$G(x,0) = \int_{b(0)}^{x} \widetilde{f}(s)ds + (x^2-2x)\int_0^1 \widetilde{f}(s)ds, \quad 0<x<1, \tag{24}$$

boundary conditions

$$\frac{\partial G}{\partial x}(1,t) = w(1,t) = 0, \quad 0<t<1, \tag{25}$$

$$G(0,t) = 0, \quad 0<t<1, \tag{26}$$

and compatibility conditions

$$\frac{\partial G}{\partial x}(1,0) = w(1,0) = 0, \tag{27}$$

$$G(0,0) = 0, \tag{28}$$

Furthermore, we can obtain the relationship between $G(x,t)$ and $v(x,t)$ as following:

$$\begin{aligned} G(x,t) &= \int_{b(t)}^{x} v(s,t)ds + (x^2-2x)\int_0^1 v(s,t)ds \\ &\quad - \int_{b(t)}^{x} F(s,t)ds - (x^2-2x)\int_0^1 F(s,t)ds \qquad (29) \\ &= \int_{b(t)}^{x} v(s,t)ds + (x^2-2x)\int_0^1 v(s,t)ds - \frac{m(t)(2x-x^2-2b(t)+b^2(t))}{b(t)(2-b(t))} \\ &\quad - \frac{g(t)x(x-b(t))}{2-b(t)} - (x^2-2x)\left(\frac{g(t)(1-b(t))}{2-b(t)} + \frac{m(t)}{b(t)(2-b(t))}\right), \end{aligned}$$

and

$$\begin{aligned} v(x,t) &= \frac{\partial G}{\partial x}(x,t) - 2(x-1)\frac{G(b(t),t)}{b(t)(b(t)-2)} + F(x,t) \qquad (30) \\ &= \frac{\partial G}{\partial x}(x,t) - 2(x-1)\frac{G(b(t),t)}{b(t)(b(t)-2)} + \frac{2x-b(t)}{2-b(t)}g(t) + \frac{2m(t)(1-x)}{b(t)(2-b(t))}. \end{aligned}$$

3 Bernstein Polynomials and Their Properties

The general form of the Bernstein polynomials of mth degree proposed by Bhatti and Bracken [1] is defined on the interval [0, 1] as

$$B_{i,m}(x) = \frac{m!}{i!(m-i)!}x^i(1-x)^{m-i}, \quad 0 \le i \le m. \qquad (31)$$

It can easily be shown that each of the Bernstein polynomials is positive and also the sum of all the Bernstein polynomials is unity for all real $x \in [0,1]$, that is,

$$\sum_{i=0}^{m} B_{i,m}(x) = 1, \quad x \in [0,1]. \qquad (32)$$

Moreover, the Bernstein polynomials have the following properties:

$$B_{i,m}(x) = (1-x)B_{i,m-1}(x) + xB_{i-1,m-1}(x), \qquad (33)$$

$$B_{i,m-1}(x) = \frac{m-i}{m}B_{i,m}(x) + \frac{i+1}{m}B_{i+1,m}(x), \qquad (34)$$

$$B'_{i,m}(x) = m(B_{i-1,m-1}(x) - B_{i,m-1}(x)), \tag{35}$$

$$\int_0^1 B_{i,m}(x)dx = \frac{1}{m+1}, \quad i = 0, 1, \cdots m. \tag{36}$$

Each kth degree Bernstein basis function can be expressed in the mth degree Bernstein basis as (see [7])

$$B_{i,k}(x) = \sum_{j=i}^{m-k+i} \frac{k!(m-k)!j!(m-j)!}{i!(k-i)!(j-i)!(m-k-j+i)!m!} B_{j,m}(x),$$
$$(i = 0, 1, \cdots k), \text{ as } k \le m. \tag{37}$$

A set of Legendre polynomials, denoted by $\{L_k(x)\}$ for $k = 0, 1, \cdots$, is orthogonal with respect to the weighting function $\omega(x) = 1$ over the interval [0, 1]. These polynomials satisfy the recurrence relation [5]

$$(k+1)L_{k+1}(x) = (2k+1)(2x-1)L_k(x) - kL_{k-1}(x), \quad k = 1, 2, \cdots, \tag{38}$$

with

$$L_0(x) = 1, \quad L_1(x) = 2x - 1. \tag{39}$$

It can be shown [8] that the Legendre polynomial $L_m(x)$ can be expressed in the mth degree Bernstein basis $B_{0,m}(x), B_{1,m}(x), \cdots, B_{m,m}(x)$ as

$$L_m(x) = \sum_{i=0}^{m} (-1)^{m+i} \frac{m!}{i!(m-i)!} B_{i,m}(x). \tag{40}$$

Thus, from (37) and (40), we can obtain that any given polynomial $P_m(x)$ of degree m can be expanded in the mth degree Legendre and Bernstein base on $x \in [0, 1]$

$$P_m(x) = \sum_{k=0}^{m} l_k L_k(x) = \sum_{i=0}^{m} c_i B_{i,m}(x). \tag{41}$$

Let $V = L^2[0, 1]$ is the vector space of real functions whose domain is the close interval [0, 1] and all functions in $V = L^2[0, 1]$ are assumed to be square integrable. We define the inner product of $f(x)$ and $g(x)$ as follows

$$< f(x), g(x) >= \int_0^1 f(x)g(x)dx. \tag{42}$$

Remarks

(1) Space $Span\{L_0(x), L_1(x), \cdots, L_m(x)\} = Span\{B_{0,m}(x), B_{1,m}(x), \cdots, B_{m,m}(x)\} := Y \subset V$ and $B_{1,m}(x), B_{2,m}(x), \cdots, B_{m,m}(x)$ are basis of subspace Y of V.

(2) Suppose $f(x) \in V = L^2[0, 1]$, then there exist a unique best approximation to $f(x)$ out of Y such as $y_0(x) \in Y$; that is, if $y(x) \in Y$,

$$\| y_0(x) - f(x) \| \leq \| y(x) - f(x) \|, \tag{43}$$

moreover

$$\begin{aligned} y_0(x) &= \sum_{k=0}^{m} c_k B_{k,m} \\ &= (c_0, c_1, \cdots, c_m)(B_{0,m}(x), B_{1,m}(x), \cdots, B_{m,m}(x))^T := C^T \phi, \end{aligned} \tag{44}$$

where coefficient matrix C^T can be obtained by

$$C^T = < f, \phi^T > < \phi, \phi^T >^{-1}. \tag{45}$$

4 Bernstein Ritz-Galerkin Method for Our Problem

In this section, we apply Ritz-Galerkin method to the second equivalent problem (23)–(28) in Sect. 2, then the approximate solution of original problem can be obtained easily by (29) and (30).

Consider the second equivalent form as following:

$$\frac{\partial G^2}{\partial x \partial t} = \frac{\partial^3 G}{\partial x^3} + K(x,t) + 2(x-1) \cdot \frac{d}{dt}\left(\frac{G(b(t),t)}{b(t)(b(t)-2)}\right), \quad 0 < x < 1, \ 0 < t < 1, \tag{46}$$

with initial condition

$$G(x,0) = \int_{b(0)}^{x} \widetilde{f}(s)ds + (x^2 - 2x)\int_0^1 \widetilde{f}(s)ds, \quad 0 < x < 1, \tag{47}$$

boundary conditions

$$\frac{\partial G}{\partial x}(1,t) = w(1,t) = 0, \quad 0 < t < 1, \tag{48}$$

$$G(0,t) = 0, \quad 0 < t < 1, \tag{49}$$

and compatibility conditions

$$\frac{\partial G}{\partial x}(1,0) = w(1,0) = 0, \tag{50}$$

$$G(0,0) = 0, \tag{51}$$

where

$$K(x,t) = p(x,t) - \frac{(2x - b(t))(2 - b(t))g'(t) + 2g(t)b'(t)(x-1)}{(2-b(t))^2} \tag{52}$$
$$- \frac{2m'(t)b(t)(2-b(t))(1-x) - 4m(t)b'(t)(1-b(t))(1-x)}{(b(t))^2(2-b(t))^2},$$

$$\widetilde{f}(x) = f(x) - \frac{2x - b(0)}{2 - b(0)} g(0) - \frac{2m(0)(1-x)}{b(0)(2-b(0))}. \tag{53}$$

Let

$$W(G) = \frac{\partial G^2}{\partial x \partial t} - \frac{\partial^3 G}{\partial x^3} - K(x,t) - 2(x-1) \cdot \frac{d}{dt}\left(\frac{G(b(t),t)}{b(t)(b(t)-2)}\right) = 0, \tag{54}$$

A Ritz-Galerkin approximation to (54) is constructed as follows. The approximation solution $\widetilde{G}(x,t)$ is sought in the form of the truncated series

$$\widetilde{G}(x,t) = G(x,0) \cdot \left(\sum_{i=0}^{N} \sum_{j=0}^{M} c_{i,j}\ t\ B_{i,N}(x) B_{j,M}(t) + 1 \right), \tag{55}$$

where $B_{i,N}(x)$, $B_{j,M}(t)$ are Bernstein polynomials. From compatibility conditions (50) and (51), it is easy to see that the approximation solution $\widetilde{G}(x,t)$ satisfies the initial condition (47) and the boundary conditions (48) and (49).

Now the expansion coefficients $c_{i,j}$ are determined by the Galerkin equations

$$< W(\widetilde{G}(x,t)), B_{i,N}(x) B_{j,M}(t) >= 0, \quad (i = 0, 1, \cdots, N,\ \ j = 0, 1, \cdots, M), \tag{56}$$

where $< . >$ denotes the inner product defined by

$$< W(\widetilde{G}(x,t)), B_{i,N}(x) B_{j,M}(t) >= \int_0^1 \int_0^1 W(\widetilde{G}(x,t)) B_{i,N}(x) B_{j,M}(t) dt dx. \tag{57}$$

Galerkin equations (56) gives a system of $(N+1)(M+1)$ linear equations which can be solved for the elements $c_{i,j}$ using mathematical software.

5 Numerical Application

In this section, a numerical example is exhibited to verify the efficiency and accuracy of our scheme.

Example 1 Consider (1)–(6) with

$$p(x,t) = \frac{x+t+2}{(x+t+1)^2}, \quad 0 \le x \le 1,\ 0 \le t \le 1, \tag{58}$$

$$f(x) = \ln(x+1), \quad 0 \le x \le 1, \tag{59}$$

$$g(t) = \ln(t+2), \quad 0 \le t \le 1, \tag{60}$$

$$b(t) = \frac{t+1}{2}, \quad 0 \le t \le 1, \tag{61}$$

$$m(t) = \frac{t+1}{2}\left(\ln(t+1) + 3\ln 3 - 3\ln 2 - 1\right), \quad 0 \le t \le 1, \tag{62}$$

which has the exact solution

$$v(x,t) = \ln(x+t+1), \tag{63}$$

From (23)–(28), we can obtain the following equivalent problem

$$\frac{\partial G^2}{\partial x \partial t} = \frac{\partial^3 G}{\partial x^3} + K(x,t) + 8(x-1)\cdot\frac{d}{dt}\left(\frac{G(\frac{t+1}{2},t)}{(t+1)(t-3)}\right),\ 0 < x < 1,\ 0 < t < 1, \tag{64}$$

with initial condition

$$G(x,0) = (1+x)\ln(1+x) - (1-\ln 2)x^2 + (1-3\ln 2)x, \quad 0 < x < 1, \tag{65}$$

boundary conditions

$$\frac{\partial G}{\partial x}(1,t) = 0, \quad 0 < t < 1, \tag{66}$$

$$G(0,t) = 0, \quad 0 < t < 1, \tag{67}$$

where

$$K(x,t) = \frac{x+t+2}{(x+t+1)^2} - \frac{(4x-t-1)(3-t)+4(x-1)(t+2)\ln(t+2)}{(3-t)^2(t+2)} - \frac{4(1-x)(t+1)\left(\ln(t+1)+3\ln\frac{3}{2}\right)+8(1-t)(1-x)}{(t+1)(3-t)^2}, \tag{68}$$

From (7), (17) and (63), we can deduce that the problem (64)–(67) has the exact solution

$$G(x,t) = (x+t+1)\ln(x+t+1) - (t+1)\ln(t+1) + x(x-2)(t+1)\ln\frac{t+2}{t+1} - x^2 + x(1-\ln(t+2)). \tag{69}$$

We applied the method presented in this paper with $N = 2$, $M = 4$ and solved Eq. (64).

From Galerkin equations (56), we have

$$\begin{cases} c_{0,0} = 1.0600,\ c_{0,1} = 1.5023,\ c_{0,2} = 0.5983,\ c_{0,3} = 0.8410,\ c_{0,4} = 0.6442, \\ c_{1,0} = 1.0200,\ c_{1,1} = 1.1670,\ c_{1,2} = 0.7445,\ c_{1,3} = 0.7004,\ c_{1,4} = 0.6289, \\ c_{2,0} = 1.0267,\ c_{2,1} = 1.1064,\ c_{2,2} = 0.7069,\ c_{2,3} = 0.7022,\ c_{2,4} = 0.6049. \end{cases} \tag{70}$$

From Eqs. (69), we can obtain the approximate solution $\widetilde{G}(x,t)$ of the problem (64)–(67) as following

$$\widetilde{G}(x,t) = G(x,0)\cdot\left(\sum_{i=0}^{N=2}\sum_{j=0}^{M=4} c_{i,j}\ t\ B_{i,N}(x)B_{j,M}(t) + 1\right), \tag{71}$$

According to (30), we can get following corresponding approximate solution $\widetilde{v}(x,t)$ of the problem (1)–(6).

$$\widetilde{v}(x,t) = \frac{\partial\widetilde{G}}{\partial x}(x,t) - \frac{8(x-1)\widetilde{G}(\frac{t+1}{2},t)}{(t+1)(t-3)} + \frac{(4x-t-1)\ln(t+2)}{3-t} + \frac{4(1-x)(\ln(t+1)+3\ln\frac{3}{2}-1)}{3-t}. \tag{72}$$

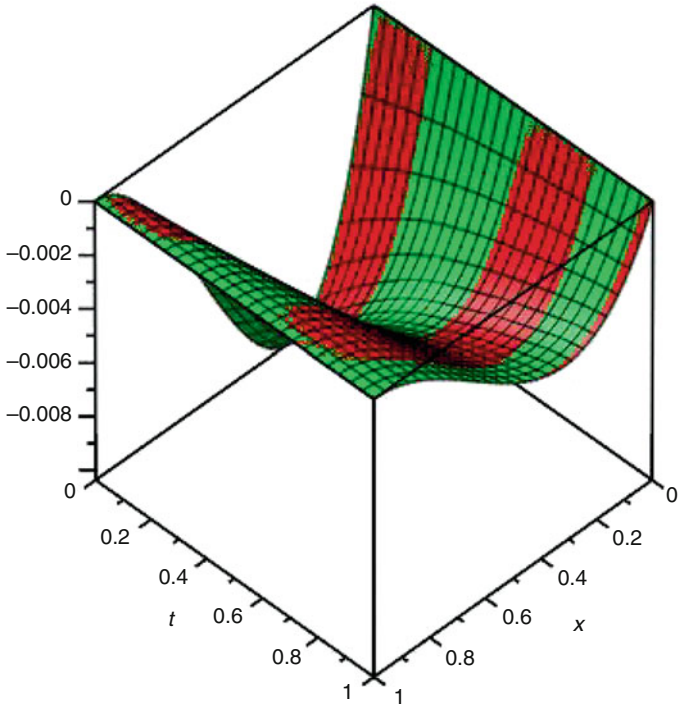


Fig. 1 Exact (red) and approximate (green) solutions of $G(x, t)$ in Example 1

In Fig. 1, the exact and approximate solutions of $G(x, t)$ with $N = 2$, $M = 4$ are plotted.

In Fig. 2, the exact and approximate solutions of $v(x, t)$ with $N = 2$, $M = 4$ are plotted.

Tables 1 and 2 present respectively absolute error for $G(x, t)$ and $v(x, t)$ with $N = 2$ and $M = 4$ in example one.

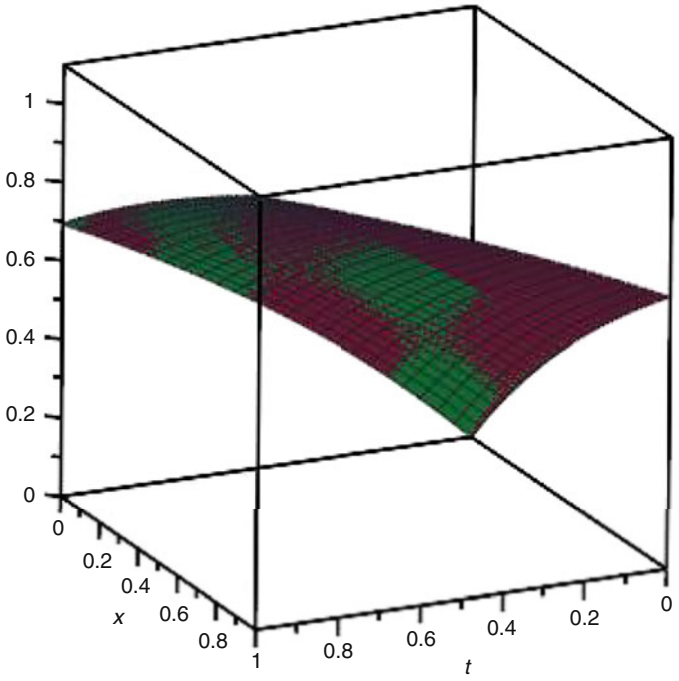


Fig. 2 Exact (red) and approximate (green) solutions of $v(x, t)$ in Example 1

Table 1 The absolute error for $G(x, t)$ in Example 1

(x, t)	Absolute error for $G(x, t)$
(0,0)	0
(0.1,0.1)	-7.17×10^{-5}
(0.2,0.2)	-1.98×10^{-5}
(0.3,0.3)	6.08×10^{-5}
(0.4,0.4)	6.85×10^{-5}
(0.5,0.5)	2.68×10^{-5}
(0.6,0.6)	-8.84×10^{-6}
(0.7,0.7)	-1.53×10^{-5}
(0.8,0.8)	-5.30×10^{-6}
(0.9,0.9)	2.78×10^{-7}
(1,1)	0

Table 2 The absolute error for $v(x,t)$ in Example 1

(x,t)	Absolute error for $v(x,t)$
(0,0)	0
(0.1,0.1)	6.17×10^{-6}
(0.2,0.2)	-1.63×10^{-5}
(0.3,0.3)	-2.31×10^{-5}
(0.4,0.4)	-7.55×10^{-6}
(0.5,0.5)	-1.43×10^{-5}
(0.6,0.6)	2.84×10^{-5}
(0.7,0.7)	3.03×10^{-5}
(0.8,0.8)	2.23×10^{-5}
(0.9,0.9)	9.23×10^{-6}
(1,1)	0

Acknowledgments The first author's research was supported in part by the *National Natural Science Foundation of the People's Republic of China* under grant numbers 11201070, the *Science Research Fund of Department of Guangdong Province of the People's Republic of China* under grant numbers Yq2013161.

References

1. Bhatti, M.I., Bracken, P.: Solution of differential equations in a Bernstein polynomial basis. J. Comput. Appl. Math. **205**, 272–280 (2007)
2. Bouziani, A., Merazga, N., Benamira, S.: Galerkin method applied to a parabolic evolution problem with nonlocal boundary conditions. Nonlinear Anal. Theory Methods Appl. **69**, 1515–1524 (2008)
3. Cannon, J.R., Lin, Y.: A Galerkin procedure for diffusion equation with boundary integral conditions. Int. J. Eng. Sci. **28**, 579–587 (1990)
4. Cannon, J.R., Matheson, A.L.: A numerical procedure for diffusion subject to the specification of mass. Int. J. Eng. Sci. **31**, 347–355 (1993)
5. Datta, K.B., Mohan, B.M.: Orthogonal Functions in Systems and Control. World Scientific, River Edge, NJ (1995)
6. Dehghan, M., Yousefi, S.A., Rashedi, K.: Ritz-Galerkin method for solving an inverse heat conduction problem with a nonlinear source term via Bernstein multi-scaling functions and cubic B-spline functions. Inverse Prob. Sci. Eng. **21**, 500–523 (2013)
7. Farouki, R.T.: Legendre-Bernstein basis transformations. J. Comput. Appl. Math. **119**(1–2), 145–160 (2000)
8. Li, Y.M., Zhang, X.Y.: Basis conversion among Bezier, Tchebyshev and Legendre. Comput. Aided Geom. Des. **15**, 637–642 (1998)
9. Li, H., Zhou, J.R.: Direct and inverse problem for the parabolic equation with initial value and time-dependent boundaries. Appl. Anal. **95**(6), 1307–1326 (2016)
10. Li, H., Xu, Y., Zhou, J.R.: A free boundary problem arising from DCIS mathematical model. Math. Methods Appl. Sci. **40**(10), 3566–3579 (2017)
11. Xu, Y.: A free boundary problem model of ductal carcinoma in situ. Discrete Contin. Dyn. Syst. Ser. B **4**(1), 337–348 (2004)
12. Xu, Y.: A mathematical model of ductal carcinoma in situ and its characteristic stationary solutions. In: Begehr, H., et al. (eds.) Advances in Analysis. World Scientific (2005)

13. Xu, Y.: An inverse problem for the free boundary model of ductal carcinoma in situ. In: Begehr, H., Nicolosi, F. (eds.) More Progresses in Analysis. World Science Publisher, pp. 1429–1438 (2008)
14. Xu, Y., Gilbert, R.: Some inverse problems raised from a mathematical model of ductal carcinoma in situ. Math. Comput. Model. **49**, 814–828 (2009)
15. Yousefi, S.A., Barikbin, Z.: Ritz-Galerkin method with Bernstein polynomial basis for finding the product solution form of heat equation with non-classic boundary conditions. Int. J. Numer. Methods Heat Fluid Flow **22**, 39–48 (2012)
16. Zhou, J.R., Li, H.: A Ritz-Galerkin approximation to the solution of parabolic equation with moving boundaries. Boundary Value Probl. **2015**, 236 (2015)
17. Zhou, J.R., Li, H., Xu, Y.: Ritz-Galerkin method for solving a parabolic equation with non-local and time-dependent boundary conditions. Math. Methods Appl. Sci. **39**, 1241–1253 (2016)
18. Zhou, J.R., Li, H., Xu, Y.: Ritz-Galerkin method for solving an inverse problem of parabolic equation with moving boundaries and integral condition. Appl. Anal. **98**(10), 1741–1755 (2019)
19. Zhou, J.R., Xu, Y., Li, H.: Another way of solving a free boundary problem related to DCIS model. Appl. Anal. **100**(15) 3244–3258 (2021)

Part V
Complex Variables and Potential Theory

A Perturbation Result for a Neumann Problem in a Periodic Domain

Matteo Dalla Riva, Paolo Luzzini, and Paolo Musolino

Abstract We consider a Neumann problem for the Laplace equation in a periodic domain. We prove that the solution depends real analytically on the shape of the domain, on the periodicity parameters, on the Neumann datum, and on its boundary integral.

1 Introduction

The aim of this paper is to prove the analytic dependence of the solution of a periodic Neumann problem for the Laplace equation, upon joint perturbation of the domain, the periodicity parameters, the Neumann datum, and its integral on the boundary. The domain is obtained as the union of congruent copies of a periodicity cell of edges of length $q_{11}, \dots, q_{nn}$ with a hole whose shape is the image of a reference domain through a diffeomorphism ϕ. As Neumann datum we take the projection of a function g, defined on the boundary of the reference domain and suitably rescaled, on the space of functions with zero integral on the boundary. As it happens for non-periodic Neumann problems, in order to identify one solution, we impose that the integral of the solution on the boundary is equal to a given real constant k. By means of a periodic version of potential theory, we prove that the solution of the problem depends real analytically on the 'periodicity-domain-Neumann datum-integral' quadruple $((q_{11}, \dots, q_{nn}), \phi, g, k)$.

M. Dalla Riva (✉)
Dipartimento di Ingegneria, Università degli Studi di Palermo, Palermo, Italy
e-mail: matteo.dallariva@unipa.it

P. Luzzini
Dipartimento di Matematica 'Tullio Levi-Civita', Università degli Studi di Padova, Padova, Italy
e-mail: pluzzini@math.unipd.it

P. Musolino
Dipartimento di Scienze Molecolari e Nanosistemi, Università Ca' Foscari Venezia, Venezia Mestre, Italy
e-mail: paolo.musolino@unive.it

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_20

Many authors have investigated the behavior of the solutions to boundary value problems upon domain perturbations. We mention, e.g., Henry [7] and Sokolowski and Zolésio [18] for elliptic domain perturbation problems. Lanza de Cristoforis [10, 11] has exploited potential theory in order to prove that the solutions of boundary value problems for the Laplace and Poisson equations depend real analytically upon domain perturbation. Moreover, analyticity results for domain perturbation problems for eigenvalues have been obtained for example for the Laplace equation by Lanza de Cristoforis and Lamberti [8], for the biharmonic operator by Buoso and Provenzano [2], and for the Maxwell's equations by Lamberti and Zaccaron [9].

In order to introduce our problem, we fix once for all a natural number

$$n \in \mathbb{N} \setminus \{0, 1\}$$

that represents the dimension of the space. If $(q_{11}, \dots, q_{nn}) \in]0, +\infty[^n$ we define a periodicity cell Q and a matrix $q \in \mathbb{D}_n^+(\mathbb{R})$ as

$$Q \equiv \prod_{j=1}^{n}]0, q_{jj}[, \qquad q \equiv \begin{pmatrix} q_{11} & 0 & \cdots & 0 \\ 0 & q_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & q_{nn} \end{pmatrix},$$

where $\mathbb{D}_n(\mathbb{R})$ is the space of $n \times n$ diagonal matrices with real entries and $\mathbb{D}_n^+(\mathbb{R})$ is the set of elements of $\mathbb{D}_n(\mathbb{R})$ with diagonal entries in $]0, +\infty[$. Here we note that we can identify $\mathbb{D}_n^+(\mathbb{R})$ and $]0, +\infty[^n$. We denote by $|Q|_n$ the n-dimensional measure of the cell Q, by ν_Q the outward unit normal to ∂Q, where it exists, and by q^{-1} the inverse matrix of q. We find convenient to set

$$\widetilde{Q} \equiv]0, 1[^n, \qquad \tilde{q} \equiv I_n,$$

where I_n denotes the identity $n \times n$ matrix. Then we introduce the reference domain: we take

$$\begin{aligned} &\alpha \in]0, 1[\text{ and a bounded open connected subset } \Omega \text{ of } \mathbb{R}^n \\ &\text{of class } C^{1,\alpha} \text{ such that } \mathbb{R}^n \setminus \overline{\Omega} \text{ is connected}, \end{aligned} \tag{1}$$

where the symbol '$\overline{\ }$' denotes the closure of a set. For the definition of sets and functions of the Schauder class $C^{1,\alpha}$ we refer, e.g., to Gilbarg and Trudinger [6]. In order to model our variable domain we consider a class of diffeomorphisms $\mathcal{A}_{\partial\Omega}^{\widetilde{Q}}$ from $\partial\Omega$ into their images contained in $\widetilde{Q}$ (see (3) below). By the Jordan-Leray separation theorem, if $\phi \in \mathcal{A}_{\partial\Omega}^{\widetilde{Q}}$, the set $\mathbb{R}^n \setminus \phi(\partial\Omega)$ has exactly two open

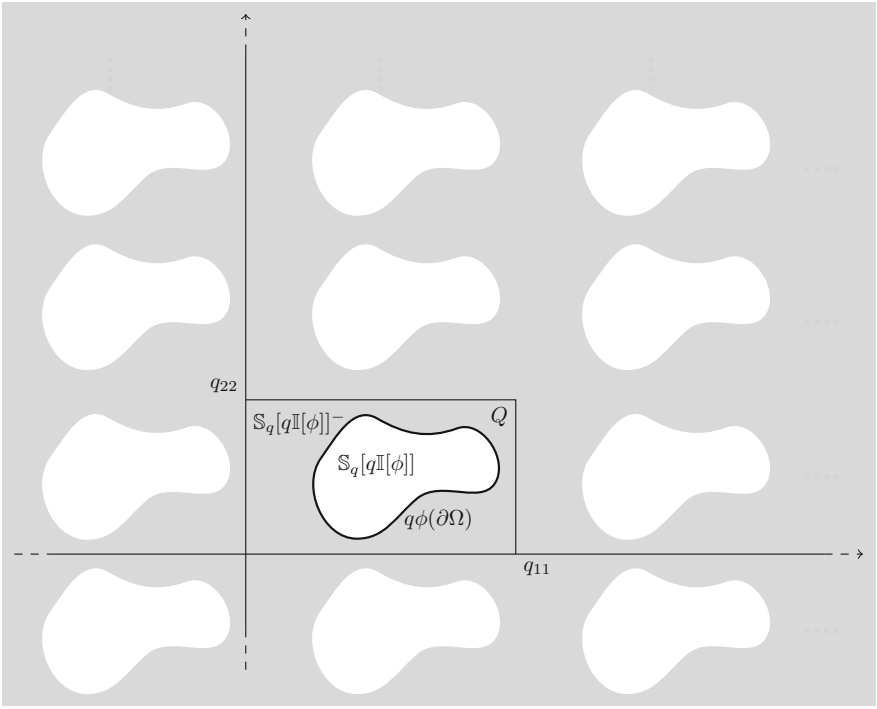


Fig. 1 The sets $\mathbb{S}_q[q\mathbb{I}[\phi]]^-$ (in gray), $\mathbb{S}_q[q\mathbb{I}[\phi]]$ (in white), and $q\phi(\partial\Omega)$ (in black) in case $n = 2$

connected components (see, e.g., Deimling [5, Thm. 5.2, p. 26]). We denote by $\mathbb{I}[\phi]$ the bounded open connected component of $\mathbb{R}^n \setminus \phi(\partial\Omega)$. Since $\phi(\partial\Omega) \subseteq \widetilde{Q}$, a topological argument shows that $\widetilde{Q} \setminus \overline{\mathbb{I}[\phi]}$ is also connected (cf., e.g., [3, Theorem A.10]). We are now in the position to introduce the following two periodic domains (see Fig. 1):

$$\mathbb{S}_q[q\mathbb{I}[\phi]] \equiv \bigcup_{z\in\mathbb{Z}^n} (qz + q\mathbb{I}[\phi]), \qquad \mathbb{S}_q[q\mathbb{I}[\phi]]^- \equiv \mathbb{R}^n \setminus \overline{\mathbb{S}_q[q\mathbb{I}[\phi]]}.$$

The set $\mathbb{S}_q[q\mathbb{I}[\phi]]^-$ will be the one where we shall set our Neumann problem. Clearly, a perturbation of q produces a modification of the whole periodicity structure of $\mathbb{S}_q[q\mathbb{I}[\phi]]^-$, while a perturbation of ϕ induces a change in the shape of the holes $\mathbb{S}_q[q\mathbb{I}[\phi]]$.

If $q \in \mathbb{D}_n^+(\mathbb{R})$, $\phi \in C^{1,\alpha}(\partial\Omega, \mathbb{R}^n) \cap \mathcal{A}^{\widetilde{Q}}_{\partial\Omega}$, $g \in C^{0,\alpha}(\partial\Omega)$ and $k \in \mathbb{R}$, we consider the following periodic Neumann problem for the Laplace equation:

$$\begin{cases} \Delta u = 0 & \text{in } \mathbb{S}_q[q\mathbb{I}[\phi]]^-, \\ u(x + qz) = u(x) & \forall x \in \overline{\mathbb{S}_q[q\mathbb{I}[\phi]]^-}, \forall z \in \mathbb{Z}^n, \\ \frac{\partial}{\partial \nu_{q\mathbb{I}[\phi]}} u(x) = g\big(\phi^{(-1)}(q^{-1}x)\big) & \\ \qquad - \frac{1}{\int_{\partial q\mathbb{I}[\phi]} d\sigma} \int_{\partial q\mathbb{I}[\phi]} g\big(\phi^{(-1)}(q^{-1}y)\big)\, d\sigma_y & \forall x \in \partial q\mathbb{I}[\phi], \\ \int_{\partial q\mathbb{I}[\phi]} u\, d\sigma = k. & \end{cases} \tag{2}$$

We note that the function

$$g\big(\phi^{(-1)}(q^{-1}\cdot)\big) - \frac{1}{\int_{\partial q\mathbb{I}[\phi]} d\sigma} \int_{\partial q\mathbb{I}[\phi]} g\big(\phi^{(-1)}(q^{-1}y)\big)\, d\sigma_y$$

clearly belongs to the space

$$C^{0,\alpha}(\partial q\mathbb{I}[\phi])_0 \equiv \left\{ \mu \in C^{0,\alpha}(\partial q\mathbb{I}[\phi]) : \int_{\partial q\mathbb{I}[\phi]} \mu\, d\sigma = 0 \right\}.$$

As a consequence, the solution of problem (2) in the space $C^{1,\alpha}_q(\overline{\mathbb{S}_q[q\mathbb{I}[\phi]]^-})$ of q-periodic functions in $\overline{\mathbb{S}_q[q\mathbb{I}[\phi]]^-}$ of class $C^{1,\alpha}$ exists and is unique and we denote it by $u[q, \phi, g, k]$ (see [3, Thm. 12.23]). Our aim is to prove that $u[q, \phi, g, k]$ depends, in a sense that we will clarify, analytically on (q, ϕ, g, k) (see Theorem 1). Our work originates from Lanza de Cristoforis [10, 11] on the real analytic dependence of the solution of the Dirichlet problem for the Laplace and Poisson equations upon domain perturbations. Moreover, this paper can be seen as the Neumann counterpart of [15], where the authors have proved analyticity properties for the solution of a periodic Dirichlet problem. An analysis similar to the one of the present paper was also carried out for periodic problems related to physical quantities arising in fluid mechanics and in material science (see [4, 14, 16]).

2 Preliminary Results

In order to consider shape perturbations, we introduce a class of diffeomorphisms. Let Ω be as in (1). Let $\mathcal{A}_{\partial\Omega}$ be the set of functions of class $C^1(\partial\Omega, \mathbb{R}^n)$ which are injective and whose differential is injective at all points of $\partial\Omega$. The set $\mathcal{A}_{\partial\Omega}$ is well-known to be open in $C^1(\partial\Omega, \mathbb{R}^n)$ (see, e.g., Lanza de Cristoforis and Rossi [13, Lem. 2.5, p. 143]). Then we set

$$\mathcal{A}^{\widetilde{Q}}_{\partial\Omega} \equiv \left\{ \phi \in \mathcal{A}_{\partial\Omega} : \phi(\partial\Omega) \subseteq \widetilde{Q} \right\}. \tag{3}$$

In order to analyze our boundary value problem, we are going to exploit periodic layer potentials. To define these operators, it is enough to replace the fundamental solution of the Laplace operator by a q-periodic tempered distribution $S_{q,n}$ such that $\Delta S_{q,n} = \sum_{z\in\mathbb{Z}^n} \delta_{qz} - \frac{1}{|Q|_n}$, where δ_{qz} is the Dirac measure with mass in qz (see e.g., [3, Chapter 12]). We can take

$$S_{q,n}(x) = -\sum_{z\in\mathbb{Z}^n\setminus\{0\}} \frac{1}{|Q|_n 4\pi^2 |q^{-1}z|^2} e^{2\pi i (q^{-1}z)\cdot x}$$

in the sense of distributions in $\mathbb{R}^n$ (see e.g., Ammari and Kang [1, p. 53], [3, §12.1]). Moreover, $S_{q,n}$ is even, real analytic in $\mathbb{R}^n \setminus q\mathbb{Z}^n$, and locally integrable in $\mathbb{R}^n$ (see e.g., [3, Thm. 12.4]). We now introduce the periodic single layer potential. Let Ω_Q be a bounded open subset of $\mathbb{R}^n$ of class $C^{1,\alpha}$ for some $\alpha \in]0, 1[$ such that $\overline{\Omega_Q} \subseteq Q$. We define the following two periodic domains:

$$\mathbb{S}_q[\Omega_Q] \equiv \bigcup_{z\in\mathbb{Z}^n} \left(qz + \Omega_Q\right), \qquad \mathbb{S}_q[\Omega_Q]^- \equiv \mathbb{R}^n \setminus \overline{\mathbb{S}_q[\Omega_Q]}$$

and we set

$$v_q[\partial\Omega_Q, \mu](x) \equiv \int_{\partial\Omega_Q} S_{q,n}(x-y)\mu(y)\, d\sigma_y \qquad \forall x \in \mathbb{R}^n$$

and

$$W_q^*[\partial\Omega_Q, \mu](x) \equiv \int_{\partial\Omega_Q} \nu_{\Omega_Q}(x) \cdot DS_{q,n}(x-y)\mu(y)\, d\sigma_y \qquad \forall x \in \partial\Omega_Q$$

for all $\mu \in L^2(\partial\Omega_Q)$. The symbol ν_{Ω_Q} denotes the outward unit normal field to $\partial\Omega_Q$, $d\sigma$ denotes the area element on $\partial\Omega_Q$ and $DS_{q,n}$ denotes the gradient of $S_{q,n}$. The function $v_q[\partial\Omega_Q, \mu]$ is called the q-periodic single layer potential. Now let $\mu \in C^{0,\alpha}(\partial\Omega_Q)$. As is well known, $v_q^+[\partial\Omega_Q, \mu] \equiv v_q[\partial\Omega_Q, \mu]_{|\overline{\mathbb{S}_q[\Omega_Q]}}$ belongs to $C_q^{1,\alpha}(\overline{\mathbb{S}_q[\Omega_Q]})$ and $v_q^-[\partial\Omega_Q, \mu] \equiv v_q[\partial\Omega_Q, \mu]_{|\overline{\mathbb{S}_q[\Omega_Q]^-}}$ belongs to $C_q^{1,\alpha}(\overline{\mathbb{S}_q[\Omega_Q]^-})$ (see [3, Thm. 12.8]). Moreover, the following jump formula for the normal derivative of the q-periodic single layer potential $v_q[\partial\Omega_Q, \mu]$ holds:

$$\frac{\partial}{\partial\nu_{\Omega_Q}} v_q^{\pm}[\partial\Omega_Q, \mu] = \mp\frac{1}{2}\mu + W_q^*[\partial\Omega_Q, \mu] \qquad \text{on } \partial\Omega_Q.$$

For a proof of the above formula we refer to [3, Thm. 12.11].

Since our approach will be based on integral operators, we need to understand how integrals behave when we perturb the domain of integration. Moreover, we need also to understand the regularity of the normal vector upon domain perturbations. For such reasons, we collect those results in the lemma below (for a proof, see Lanza de Cristoforis and Rossi [13, p. 166]).

Lemma 1 *Let α, Ω be as in (1). Then the following statements hold.*

(i) For each $\psi \in C^{1,\alpha}(\partial\Omega, \mathbb{R}^n) \cap \mathcal{A}_{\partial\Omega}$, there exists a unique $\tilde{\sigma}[\psi] \in C^{0,\alpha}(\partial\Omega)$ such that $\tilde{\sigma}[\psi] > 0$ and

$$\int_{\psi(\partial\Omega)} \omega(s)\, d\sigma_s = \int_{\partial\Omega} \omega \circ \psi(y) \tilde{\sigma}[\psi](y)\, d\sigma_y, \qquad \forall \omega \in L^1(\psi(\partial\Omega)).$$

Moreover, the map $\tilde{\sigma}[\cdot]$ from $C^{1,\alpha}(\partial\Omega, \mathbb{R}^n) \cap \mathcal{A}_{\partial\Omega}$ to $C^{0,\alpha}(\partial\Omega)$ is real analytic.

(ii) The map from $C^{1,\alpha}(\partial\Omega, \mathbb{R}^n) \cap \mathcal{A}_{\partial\Omega}$ to $C^{0,\alpha}(\partial\Omega, \mathbb{R}^n)$ which takes ψ to $\nu_{\mathbb{I}[\psi]} \circ \psi$ is real analytic.

3 Analyticity of the Solution

Our first goal is to transform problem (2) into an integral equation. In order to analyze the solvability of the obtained integral equation, we need the following lemma.

Lemma 2 *Let α, Ω be as in (1). Let $q \in \mathbb{D}_n^+(\mathbb{R})$. Let $\phi \in C^{1,\alpha}(\partial\Omega, \mathbb{R}^n) \cap \mathcal{A}^{\widetilde{Q}}_{\partial\Omega}$. Let N be the map from $C^{0,\alpha}(\partial q\mathbb{I}[\phi])$ to itself, defined by*

$$N[\mu] \equiv \frac{1}{2}\mu + W_q^*[\partial q\mathbb{I}[\phi], \mu] \qquad \forall \mu \in C^{0,\alpha}(\partial q\mathbb{I}[\phi]).$$

Then N is a linear homeomorphism from $C^{0,\alpha}(\partial q\mathbb{I}[\phi])$ to itself. Moreover, N restricts to a linear homeomorphism from $C^{0,\alpha}(\partial q\mathbb{I}[\phi])_0$ to itself.

Proof By Dalla Riva et al. [3, Thm. 12.20], we deduce that N is a linear homeomorphism from $C^{0,\alpha}(\partial q\mathbb{I}[\phi])$ to itself. By Dalla Riva et al. [3, Prop. 12.15], we have that $\frac{1}{2}\mu + W_q^*[\partial q\mathbb{I}[\phi], \mu]$ belongs to $C^{0,\alpha}(\partial q\mathbb{I}[\phi])_0$ if and only if μ belongs to $C^{0,\alpha}(\partial q\mathbb{I}[\phi])_0$. As a consequence, we also have that N restricts to a linear homeomorphism from $C^{0,\alpha}(\partial q\mathbb{I}[\phi])_0$ to itself. □

Then, in the following proposition, we show how to convert the Neumann problem into an equivalent integral equation.

Proposition 1 *Let α, Ω be as in (1). Let $q \in \mathbb{D}_n^+(\mathbb{R})$. Let $\phi \in C^{1,\alpha}(\partial\Omega, \mathbb{R}^n) \cap \mathcal{A}^{\widetilde{Q}}_{\partial\Omega}$. Let $g \in C^{0,\alpha}(\partial\Omega)$. Let $k \in \mathbb{R}$. Then the boundary value problem*

$$\begin{cases} \Delta u = 0 & \text{in } \mathbb{S}_q[q\mathbb{I}[\phi]]^-, \\ u(x + qz) = u(x) & \forall x \in \overline{\mathbb{S}_q[q\mathbb{I}[\phi]]^-}, \forall z \in \mathbb{Z}^n, \\ \frac{\partial}{\partial \nu_{q\mathbb{I}[\phi]}} u(x) = g\big(\phi^{(-1)}(q^{-1}x)\big) & \\ \qquad - \frac{1}{\int_{\partial q\mathbb{I}[\phi]} d\sigma} \int_{\partial q\mathbb{I}[\phi]} g\big(\phi^{(-1)}(q^{-1}y)\big)\, d\sigma_y & \forall x \in \partial q\mathbb{I}[\phi], \\ \int_{\partial q\mathbb{I}[\phi]} u\, d\sigma = k & \end{cases} \tag{4}$$

has a unique solution $u[q, \phi, g, k]$ in $C^{1,\alpha}_q(\overline{\mathbb{S}_q[q\mathbb{I}[\phi]]^-})$. Moreover,

$$\begin{aligned} u[q, \phi, g, k](x) &= v_q^-[\partial q\mathbb{I}[\phi], \mu](x) \\ &\quad + \frac{1}{\int_{\partial q\mathbb{I}[\phi]} d\sigma} \left(k - \int_{\partial q\mathbb{I}[\phi]} v_q^-[\partial q\mathbb{I}[\phi], \mu]\, d\sigma \right) \qquad \forall x \in \overline{\mathbb{S}_q[q\mathbb{I}[\phi]]^-}, \end{aligned} \tag{5}$$

where μ is the unique solution in $C^{0,\alpha}(\partial q\mathbb{I}[\phi])_0$ of the integral equation

$$\begin{aligned} \frac{1}{2}\mu(x) + W_q^*[\partial q\mathbb{I}[\phi], \mu](x) &= g\big(\phi^{(-1)}(q^{-1}x)\big) \\ &\quad - \frac{1}{\int_{\partial q\mathbb{I}[\phi]} d\sigma} \int_{\partial q\mathbb{I}[\phi]} g\big(\phi^{(-1)}(q^{-1}y)\big)\, d\sigma_y \qquad \forall x \in \partial q\mathbb{I}[\phi]. \end{aligned} \tag{6}$$

Proof By Dalla Riva et al. [3, Thm. 12.23] we know that problem (4) has a unique solution. Moreover, by Lemma 2, equation (6) has a unique solution μ which belongs to $C^{0,\alpha}(\partial q\mathbb{I}[\phi])_0$. Then by the properties of the periodic single layer potential (see, e.g., [3, Thm. 12.8]), we deduce that the right hand side of (5) solves problem (4). □

In Proposition 1, we have seen an integral equation on $\partial q\mathbb{I}[\phi]$, namely equation (6), equivalent to problem (2). However, if we want to study the dependence of the solution of the integral equation on the parameters (q, ϕ, g, k), it may be convenient to transform the equation on the (q, ϕ)-dependent set $\partial q\mathbb{I}[\phi]$ into an equation on a fixed domain. We do so in the lemma below.

Lemma 3 *Let α, Ω be as in (1). Let $q \in \mathbb{D}_n^+(\mathbb{R})$. Let $\phi \in C^{1,\alpha}(\partial\Omega, \mathbb{R}^n) \cap \mathcal{A}^{\widetilde{Q}}_{\partial\Omega}$. Let $g \in C^{0,\alpha}(\partial\Omega)$. Then the function $\theta \in C^{0,\alpha}(\partial\Omega)$ solves the equation*

$$\begin{aligned}\frac{1}{2}\theta(t) + \int_{q\phi(\partial\Omega)} \nu_{q\mathbb{I}[\phi]}(q\phi(t)) \cdot DS_{q,n}(q\phi(t) - y)\theta\big(\phi^{(-1)}(q^{-1}y)\big)d\sigma_y \\ = g(t) - \frac{1}{\int_{\partial\Omega} \tilde{\sigma}[q\phi]\, d\sigma} \int_{\partial\Omega} g\tilde{\sigma}[q\phi]\, d\sigma \qquad \forall t \in \partial\Omega\,,\end{aligned} \tag{7}$$

if and only if the function $\mu \in C^{0,\alpha}(\partial q\mathbb{I}[\phi])$, with μ delivered by

$$\mu(x) = \theta\big(\phi^{(-1)}(q^{-1}x)\big) \qquad \forall x \in \partial q\mathbb{I}[\phi], \tag{8}$$

solves the equation

$$\begin{aligned}&\frac{1}{2}\mu(x) + W_q^*[\partial q\mathbb{I}[\phi], \mu](x) \\ &= g\big(\phi^{(-1)}(q^{-1}x)\big) - \frac{1}{\int_{\partial q\mathbb{I}[\phi]} d\sigma} \int_{\partial q\mathbb{I}[\phi]} g\big(\phi^{(-1)}(q^{-1}y)\big)\, d\sigma_y \quad \forall x \in \partial q\mathbb{I}[\phi]\,.\end{aligned}$$

Moreover, Eq. (7) has a unique solution θ in $C^{0,\alpha}(\partial\Omega)$ and the function μ delivered by (8) belongs to $C^{0,\alpha}(\partial q\mathbb{I}[\phi])_0$.

Proof It is a direct consequence of the theorem of change of variable in integrals, of Lemma 2, and of the obvious equality

$$\int_{\partial q\mathbb{I}[\phi]} \left(g\big(\phi^{(-1)}(q^{-1}x)\big) - \frac{1}{\int_{\partial q\mathbb{I}[\phi]} d\sigma} \int_{\partial q\mathbb{I}[\phi]} g\big(\phi^{(-1)}(q^{-1}y)\big)\, d\sigma_y \right) d\sigma_x = 0\,,$$

which implies that

$$g\big(\phi^{(-1)}(q^{-1}\cdot)\big) - \frac{1}{\int_{\partial q\mathbb{I}[\phi]} d\sigma} \int_{\partial q\mathbb{I}[\phi]} g\big(\phi^{(-1)}(q^{-1}y)\big)\, d\sigma_y$$

is in $C^{0,\alpha}(\partial q\mathbb{I}[\phi])_0$. □

Our next goal is to study the dependence of the solution of the integral equation (7) upon (q, ϕ, g). We wish to apply the implicit function theorem in Banach spaces. Therefore, having in mind equation (7), we introduce the map Λ from $\mathbb{D}_n^+(\mathbb{R}) \times$

$\left(C^{1,\alpha}(\partial\Omega,\mathbb{R}^n)\cap\mathcal{A}^{\widetilde{Q}}_{\partial\Omega}\right)\times\left(C^{0,\alpha}(\partial\Omega)\right)^2$ to $C^{0,\alpha}(\partial\Omega)$ by setting

$$\begin{aligned}\Lambda[q,\phi,g,\theta](t)\equiv{}&\frac{1}{2}\theta(t)\\&+\int_{q\phi(\partial\Omega)}\nu_{q\mathbb{I}[\phi]}(q\phi(t))\cdot DS_{q,n}(q\phi(t)-y)\theta\left(\phi^{(-1)}(q^{-1}y)\right)d\sigma_y\\&-g(t)+\frac{1}{\int_{\partial\Omega}\tilde{\sigma}[q\phi]\,d\sigma}\int_{\partial\Omega}g\tilde{\sigma}[q\phi]\,d\sigma\quad\forall t\in\partial\Omega,\end{aligned}$$

for all $(q,\phi,g,\theta)\in\mathbb{D}^+_n(\mathbb{R})\times\left(C^{1,\alpha}(\partial\Omega,\mathbb{R}^n)\cap\mathcal{A}^{\widetilde{Q}}_{\partial\Omega}\right)\times\left(C^{0,\alpha}(\partial\Omega)\right)^2$.

We are now ready to apply the implicit function theorem for real analytic maps in Banach spaces to equation $\Lambda[q,\phi,g,\theta]=0$ and prove that the solution θ depends analytically on (q,ϕ,g).

Proposition 2 *Let α, Ω be as in (1). Then the following statements hold.*

(i) Λ *is real analytic.*

(ii) For each $(q,\phi,g)\in\mathbb{D}^+_n(\mathbb{R})\times\left(C^{1,\alpha}(\partial\Omega,\mathbb{R}^n)\cap\mathcal{A}^{\widetilde{Q}}_{\partial\Omega}\right)\times C^{0,\alpha}(\partial\Omega)$, *there exists a unique* θ *in* $C^{0,\alpha}(\partial\Omega)$ *such that*

$$\Lambda[q,\phi,g,\theta]=0\qquad\text{on }\partial\Omega,$$

and we denote such a function by $\theta[q,\phi,g]$.

(iii) The map $\theta[\cdot,\cdot,\cdot]$ *from* $\mathbb{D}^+_n(\mathbb{R})\times\left(C^{1,\alpha}(\partial\Omega,\mathbb{R}^n)\cap\mathcal{A}^{\widetilde{Q}}_{\partial\Omega}\right)\times C^{0,\alpha}(\partial\Omega)$ *to* $C^{0,\alpha}(\partial\Omega)$ *that takes* (q,ϕ,g) *to* $\theta[q,\phi,g]$ *is real analytic.*

Proof By Luzzini et al. [17, Thm. 3.2 (ii)], Lemma 1, and standard calculus in Banach spaces, we deduce the validity of statement (i). Statement (ii) follows by Lemmas 2 and 3. In order to prove (iii), since the analyticity is a local property, it suffices to fix (q_0,ϕ_0,g_0) in $\mathbb{D}^+_n(\mathbb{R})\times\left(C^{1,\alpha}(\partial\Omega,\mathbb{R}^n)\cap\mathcal{A}^{\widetilde{Q}}_{\partial\Omega}\right)\times C^{0,\alpha}(\partial\Omega)$ and to show that $\theta[\cdot,\cdot,\cdot]$ is real analytic in a neighborhood of (q_0,ϕ_0,g_0) in the product space $\mathbb{D}^+_n(\mathbb{R})\times\left(C^{1,\alpha}(\partial\Omega,\mathbb{R}^n)\cap\mathcal{A}^{\widetilde{Q}}_{\partial\Omega}\right)\times C^{0,\alpha}(\partial\Omega)$. By standard calculus in normed spaces, the partial differential $\partial_\theta\Lambda[q_0,\phi_0,g_0,\theta[q_0,\phi_0,g_0]]$ of Λ at $(q_0,\phi_0,g_0,\theta[q_0,\phi_0,g_0])$ with respect to the variable θ is delivered by

$$\begin{aligned}&\partial_\theta\Lambda[q_0,\phi_0,g_0,\theta[q_0,\phi_0,g_0]](\psi)(t)\\&=\frac{1}{2}\psi(t)+\int_{q_0\phi_0(\partial\Omega)}\nu_{q_0\mathbb{I}[\phi_0]}(q_0\phi_0(t))\cdot DS_{q_0,n}(q_0\phi_0(t)-y)\psi\left(\phi_0^{(-1)}(q_0^{-1}y)\right)d\sigma_y\\&\qquad\qquad\forall t\in\partial\Omega,\end{aligned}$$

for all $\psi \in C^{0,\alpha}(\partial\Omega)$. Lemma 2 together with a change of variable implies that $\partial_\theta \Lambda[q_0, \phi_0, g_0, \theta[q_0, \phi_0, g_0]]$ is a linear homeomorphism from $C^{0,\alpha}(\partial\Omega)$ onto $C^{0,\alpha}(\partial\Omega)$. Finally, by the implicit function theorem for real analytic maps in Banach spaces (see, e.g., Deimling [5, Thm. 15.3]) we deduce that $\theta[\cdot, \cdot, \cdot]$ is real analytic in a neighborhood of (q_0, ϕ_0, g_0) in $\mathbb{D}_n^+(\mathbb{R}) \times \left(C^{1,\alpha}(\partial\Omega, \mathbb{R}^n) \cap \mathcal{A}^{\widetilde{Q}}_{\partial\Omega}\right) \times C^{0,\alpha}(\partial\Omega)$. □

Remark 1 By Lemma 1, Propositions 1 and 2, we have the following representation formula for the solution $u[q, \phi, g, k]$ of problem (2):

$$u[q, \phi, g, k](x) = \int_{\partial\Omega} S_{q,n}(x - q\phi(s))\theta[q, \phi, g](s)\tilde{\sigma}[q\phi](s)\, d\sigma_s$$
$$+ \frac{\left(k - \int_{\partial\Omega}\int_{\partial\Omega} S_{q,n}(q(\phi(t) - \phi(s)))\theta[q, \phi, g](s)\tilde{\sigma}[q\phi](s)d\sigma_s\tilde{\sigma}[q\phi](t)d\sigma_t\right)}{\int_{\partial\Omega}\tilde{\sigma}[q\phi]d\sigma}$$
$$\forall x \in \overline{\mathbb{S}_q[q\mathbb{I}[\phi]]^-},$$

for all $(q, \phi, g, k) \in \mathbb{D}_n^+(\mathbb{R}) \times \left(C^{1,\alpha}(\partial\Omega, \mathbb{R}^n) \cap \mathcal{A}^{\widetilde{Q}}_{\partial\Omega}\right) \times C^{0,\alpha}(\partial\Omega) \times \mathbb{R}$.

By exploiting the representation formula of Remark 1 and the analyticity result for $(q, \phi, g) \mapsto \theta[q, \phi, g]$ of Proposition 2, we are ready to prove our main result on the analyticity of $u[q, \phi, g, k]$ as a map of the variable (q, ϕ, g, k).

Theorem 1 *Let α, Ω be as in (1). Let*

$$(q_0, \phi_0, g_0, k_0) \in \mathbb{D}_n^+(\mathbb{R}) \times \left(C^{1,\alpha}(\partial\Omega, \mathbb{R}^n) \cap \mathcal{A}^{\widetilde{Q}}_{\partial\Omega}\right) \times C^{0,\alpha}(\partial\Omega) \times \mathbb{R}.$$

Let U be a bounded open subset of $\mathbb{R}^n$ such that $\overline{U} \subseteq \mathbb{S}_{q_0}[q_0\mathbb{I}[\phi_0]]^-$. Then there exists an open neighborhood $\mathcal{U}$ of (q_0, ϕ_0, g_0, k_0) in

$$\mathbb{D}_n^+(\mathbb{R}) \times \left(C^{1,\alpha}(\partial\Omega, \mathbb{R}^n) \cap \mathcal{A}^{\widetilde{Q}}_{\partial\Omega}\right) \times C^{0,\alpha}(\partial\Omega) \times \mathbb{R}$$

such that the following statements hold.

(i) $\overline{U} \subseteq \mathbb{S}_q[q\mathbb{I}[\phi]]^-$ for all $(q, \phi, g, k) \in \mathcal{U}$.
(ii) Let $m \in \mathbb{N}$. Then the map from $\mathcal{U}$ to $C^m(\overline{U})$ which takes (q, ϕ, g, k) to the restriction $u[q, \phi, g, k]_{|\overline{U}}$ of $u[q, \phi, g, k]$ to $\overline{U}$ is real analytic.

Proof We first note that, by taking $\mathcal{U}$ small enough, we can deduce the validity of (i). The validity of (ii) follows by the representation formula of Remark 1, by Lemma 1, by Proposition 2, by the regularity results of [12] on the analyticity of integral operators with real analytic kernels, and by standard calculus in Banach spaces. □

Acknowledgments 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). P.L. and P.M. acknowledge the support of the Project BIRD191739/19 'Sensitivity analysis of partial differential equations in the mathematical theory of electromagnetism' of the University of Padova. P.M. acknowledges the support of the grant 'Challenges in Asymptotic and Shape Analysis - CASA' of the Ca' Foscari University of Venice. P.M. also acknowledges the support from EU through the H2020-MSCA-RISE-2020 project EffectFact, Grant agreement ID: 101008140.

References

1. Ammari, H., Kang H.: Polarization and Moment Tensors, With Applications to Inverse Problems and Effective Medium Theory. Springer (2007)
2. Buoso, D., Provenzano, L.: A few shape optimization results for a biharmonic Steklov problem. J. Differential Equations **259**(5), 1778–1818 (2015)
3. Dalla Riva, M., Lanza de Cristoforis, M., Musolino, P.: Singularly Perturbed Boundary Value Problems: A Functional Analytic Approach. Springer Nature, Cham (2021)
4. Dalla Riva, M., Luzzini, P., Musolino, P., Pukhtaievych, R.: Dependence of effective properties upon regular perturbations. In: Andrianov, I., Gluzman, S., Mityushev, V. (eds.) Mechanics and Physics of Structured Media, pp. 271–301. Elsevier (2022)
5. Deimling, D.: Nonlinear Functional Analysis. Springer-Verlag (1985)
6. Gilbarg, D., Trudinger, N.S.: Elliptic Partial Differential Equations of Second Order, 2nd edn. Springer (1983)
7. Henry, D.: Topics in nonlinear analysis. In: Trabalho de Matematica, vol. 192. Universidade de Brasilia (1982)
8. Lamberti, P.D., Lanza de Cristoforis, M.: A real analyticity result for symmetric functions of the eigenvalues of a domain dependent Dirichlet problem for the Laplace operator. J. Nonlinear Convex Anal. **5**(1), 19–42 (2004)
9. Lamberti, P.D., Zaccaron, M.: Shape sensitivity analysis for electromagnetic cavities. Math. Methods Appl. Sci. **44**(13), 10477–10500 (2021)
10. Lanza de Cristoforis, M.: A domain perturbation problem for the Poisson equation. Complex Var. Theory Appl. **50**(7–11), 851–867 (2005)
11. Lanza de Cristoforis, M.: Perturbation problems in potential theory, a functional analytic approach. J. Appl. Funct. Anal. **2**(3), 197–222 (2007)
12. Lanza de Cristoforis, M., Musolino, P.: A real analyticity result for a nonlinear integral operator. J. Integral Equations Appl. **25**(1), 21–46 (2013)
13. Lanza de Cristoforis, M., Rossi, L.: Real analytic dependence of simple and double layer potentials upon perturbation of the support and of the density. J. Integral Equations Appl. **16**, 137–174 (2004)
14. Luzzini, P., Musolino, P.: Perturbation analysis of the effective conductivity of a periodic composite. Netw. Heterog. Media **15**(4), 581–603 (2020)
15. Luzzini, P., Musolino, P.: Domain perturbation for the solution of a periodic Dirichlet problem. In: Cerejeiras, P., Reissig, M., Sabadini, I., Toft, J. (eds.) Current Trends in Analysis, its Applications and Computation, Proceedings of the 12th ISAAC Congress (Aveiro, 2019), Research Perspectives. Birkhäuser (2021)
16. Luzzini, P., Musolino, P., Pukhtaievych, R.: Shape analysis of the longitudinal flow along a periodic array of cylinders. J. Math. Anal. Appl. **477**(2), 1369–1395 (2019)

17. Luzzini, P., Musolino, P., Pukhtaievych, R.: Real analyticity of periodic layer potentials upon perturbation of the periodicity parameters and of the support. In: Proceedings of the 12th ISAAC Congress (Aveiro, 2019), Research Perspectives. Birkhäuser (2021)
18. Sokolowski, J., Zolésio, J.P.: Introduction to Shape Optimization. Shape Sensitivity Analysis. Springer (1992)

On One Inequality for Non-overlapping Domains

Iryna Denega

Abstract In this paper, an approach which allowed to obtain new estimates of the products of the inner radii of mutually non-overlapping domains is proposed. Problem of the maximum of the product of inner radii of two non-overlapping multiconnected domains is considered.

1 Preliminaries

Let $\mathbb{N}$, $\mathbb{R}$ be the sets of natural and real numbers, respectively, $\mathbb{C}$ be the complex plane, $\overline{\mathbb{C}} = \mathbb{C} \bigcup \{\infty\}$ be its one point compactification, U be the open unit disk, $\mathbb{R}^+ = (0, \infty)$.

Definition 1 Let $B \subset \overline{\mathbb{C}}$ be a simply connected domain and $a \in B$. According to the Riemann theorem on mapping, there exists a unique conformal mapping of the domain B onto the unit disk at which $f(a) = 0 \in U$, $f'(a) \in \mathbb{R}^+$. Consider inverse mapping φ which maps a unit disk U onto domain B such that $\varphi(0) = a$. Then

$$R(B, a) = \frac{1}{|f'(a)|} = |\varphi'(0)|$$

is called conformal radius of the simply connected domain $B \subset \overline{\mathbb{C}}$ relative to a point $a \in B$.

Conformal radius of the domain B with respect to an infinitely distant point is $r(B, \infty) = r(\varphi(B), 0)$, where $\varphi(z) = 1/z$.

I. Denega (✉)
Institute of Mathematics of the National Academy of Sciences of Ukraine, Kyiv, Ukraine

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_21

Definition 2 A function $g_B(z,a)$ which is continuous in $\overline{\mathbb{C}}$, harmonic in $B\backslash\{a\}$ apart from z, vanishes outside B, and in the neighborhood of a has the following asymptotic expansion

$$g_B(z,a) = -\ln|z-a| + \delta + o(1), \quad o(1) \to 0, \quad z \to a,$$

(if $a = \infty$, then $g_B(z,\infty) = \ln|z| + \delta + o(1)$, $o(1) \to 0$, $z \to \infty$) is called the (classical) Green function of the domain B with pole at $a \in B$.

Denote by e^{δ} the inner radius $r(B,a)$ of the domain B with respect to a point a (see, for example, [1–3]).

It is known [4], that the following inequality holds

$$|\varphi'(0)| \le r(B,a).$$

For compact $E \subset \mathbb{C}$ its (logarithmic) capacity is determined by the following equalities

$$\operatorname{cap} E := \frac{1}{r(\overline{\mathbb{C}}\backslash E, \infty)},$$

if the value $r(\overline{\mathbb{C}}\backslash E, \infty)$ is finite; $\operatorname{cap} E := 0$ otherwise [2, 3].

Definition 3 Let G be a domain in $\overline{\mathbb{C}}_z$. By a quadratic differential in G we mean the expression

$$Q(z)dz^2, \tag{1}$$

where $Q(z)$ is a meromorphic function in G [2, 3, 5–7].

A finite point a in G is called a zero or a pole of order n of the differential (1) if it is a zero or a pole, respectively, of the function $Q(z)$.

A circular domain G for $Q(z)dz^2$ contains a unique double pole a of $Q(z)dz^2$ and $G\setminus\{a\}$ is filled by trajectories of $Q(z)dz^2$, which are closed Jordan curves, each separating a from the boundary of G. For a suitable choice of a purely imaginary constant τ the function $w = exp\left\{\tau \int (Q(z))^{\frac{1}{2}} dz\right\}$, set equal to zero at a, maps G conformally onto a disc $|w| < R$.

In 1934 Lavrentyev in paper [8], in particular, solved the problem of maximum of the product of conformal radii of two non-overlapping simply connected domains. Namely, the following result is true.

Theorem 1 ([8]) *Let a_1 and a_2 be some fixed points of the complex plane $\mathbb{C}$, B_k, $a_k \in B_k$, $k \in \{1,2\}$ be any non-overlapping simply connected domains on $\mathbb{C}$. Then the following inequality holds*

$$R(B_1,a_1)R(B_2,a_2) \le |a_1 - a_2|^2. \tag{2}$$

Equality in (2) *occurs only in the case where the domains* B_1 *and* B_2 *are two half-planes and the imaginary axis is their common boundary and points* a_1, a_2 *are symmetrical about their common boundary.*

Later (see, for example, [9]), Lavrentyev's result was generalized to the case of meromorphic functions. Then, for any domains $B_1 \subset \overline{\mathbb{C}}$ and $B_2 \subset \overline{\mathbb{C}}$ the inequality (2) is valid and equality in (2) is attended if domains B_1 and B_2 have the following form

$$B_1 = \left\{ w \in \overline{\mathbb{C}} : \left| \frac{w - a_1}{w - a_2} \right| < \rho \right\}, \quad B_2 = \left\{ w \in \overline{\mathbb{C}} : \left| \frac{w - a_1}{w - a_2} \right| > \rho \right\}$$

or vice versa, $\rho \in \mathbb{R}^+$. An example of such configuration of the domains is the case where one domain is bounded by some circle and the other is unrestricted, that is, a complement to the first domain. It should be noted that in this case we have a whole continuum of extremals.

In the Kolbina papers [10, 11] Lavrentyev's result summarized by taking functions in fixed positive degrees: for any finite different points a_1 and a_2 maximum

$$J_0 = \frac{4^{\alpha+\beta}\alpha^\alpha\beta^\beta}{|\alpha - \beta|^{\alpha+\beta}} \left| \frac{\sqrt{\alpha} - \sqrt{\beta}}{\sqrt{\alpha} + \sqrt{\beta}} \right|^{2\sqrt{\alpha\beta}} |a_1 - a_2|^{\alpha+\beta}$$

of the value $J = R^\alpha(B_1, a_1)R^\beta(B_2, a_2)$, $\alpha, \beta \in \mathbb{R}^+$, with respect to all possible pairs of the domains B_1, B_2 such that $a_1 \in B_1 \subset \mathbb{C}$, $a_2 \in B_2 \subset \mathbb{C}$, is attained for poles and circular domains of the following quadratic differential

$$Q(z)dz^2 = -\frac{(a_2 - a_1)[z - a_1 - \alpha(a_2 - a_1)/(\alpha - \beta)]}{(z - a_1)^2(z - a_2)^2} dz^2.$$

Further, generalization of the Theorem 1 was manifested by increasing the number of domains, refusing to fix the poles of the corresponding quadratic differentials, moving to an extended complex plane, expanding the object of study – domains that do not intersect replace some class of open sets or partially intersecting domains (see, for example, [2, 3, 12–18]).

The method proposed in this paper originates in the work [19], which considered the problem of finding the maximum of the product of inner radii of three mutually disjoint domains under the additional condition of symmetry with respect to the unit circle of two of them and the degree $\gamma = 1$ the inner radius of the domain relative to the origin. The ideas suggested in [19] were generalized in the works [20–25].

2 Estimates of Products of Inner Radii

We have the following three results as well.

Theorem 2 *Let $n \in \mathbb{N}$. Then for any fixed system of different points $\{a_k\}_{k=1}^n$ and for any collection of mutually non-overlapping domains $\{B_k\}_{k=0}^n$, $a_k \in B_k \subset \overline{\mathbb{C}}$, $k = \overline{0,n}$, $a_0 = 0$, such that $r(B_1, a_1) = r(B_2, a_2) = \ldots = r(B_n, a_n)$, the following inequality holds*

$$r(B_0, 0)\, r(B_k, a_k) \leq \frac{1}{\sqrt{n}} \left(\prod_{k=1}^{n} |a_k| \right)^{\frac{2}{n}}. \tag{3}$$

Proof Consider the product

$$r^n(B_0, 0) \prod_{k=1}^{n} r(B_k, a_k)$$

for any fixed system of different points $\{a_k\}_{k=1}^n$ and for any collection of mutually non-overlapping domains $\{B_k\}_{k=0}^n$, $a_k \in B_k \subset \overline{\mathbb{C}}$, $k = \overline{0,n}$, $a_0 = 0$.

Let $d(E)$ be the transfinite diameter of a compact set $E \subset \mathbb{C}$. It is known [1, 2], that the logarithmic capacity $\mathrm{cap}E$ coincides with the transfinite diameter $d(E)$ of the set E

$$\mathrm{cap}E = d(E).$$

Since the inner radius of the domain containing an infinitely distant point is reciprocal to the transfinite diameter of the complement to this domain, then

$$r(B_0, 0) = r\left(B_0^+, \infty\right) = \frac{1}{d\left(\overline{\mathbb{C}} \setminus B_0^+\right)}. \tag{4}$$

Here, $B^+ = \left\{ z : \frac{1}{z} \in B \right\}$. By virtue of the well-known Polya theorem [1, 26], the inequality

$$\mu E \leq \pi d^2(E), \tag{5}$$

where μE denotes the Lebesgue measure of a compact set E, holds. Then relation (5) yields

$$d(E) \geq \left(\frac{1}{\pi} \mu E \right)^{\frac{1}{2}}.$$

Thus,

$$\frac{1}{d\left(\overline{\mathbb{C}} \setminus B_0^+\right)} \leq \frac{1}{\sqrt{\frac{1}{\pi}\mu\left(\overline{\mathbb{C}} \setminus B_0^+\right)}}.$$

Using monotony and additivity of the Lebesgue measure

$$\frac{1}{\sqrt{\frac{1}{\pi}\mu\left(\overline{\mathbb{C}}\setminus B_0^+\right)}}\le\frac{1}{\sqrt{\frac{1}{\pi}\mu\left(\bigcup\limits_{k=1}^{n}\overline{B}_k^+\right)}}=\frac{1}{\sqrt{\frac{1}{\pi}\sum\limits_{k=1}^{n}\mu\overline{B}_k^+}}.$$

Then from (4), we obtain

$$r\left(B_0,0\right)\le\left(\frac{1}{\pi}\sum_{k=1}^{n}\mu\overline{B}_k^+\right)^{-\frac{1}{2}}. \tag{6}$$

The area-minimization theorem [1] implies that

$$\mu(B)\ge\pi r^2\left(B,a\right).$$

It follows from inequality (6) that

$$r\left(B_0,0\right)\le\left[\sum_{k=1}^{n}r^2\left(B_k^+,a_k^+\right)\right]^{-\frac{1}{2}}.$$

Taking into account an invariance of the Green function in conformal and univalent mapping we get

$$g_{B_k}(z,a_k)=g_{B_k^+}(w^+,a_k^+),\quad w^+=\frac{1}{z}.$$

Using the asymptotic expansion

$$g_{B_k^+}(w^+,a_k^+)=g_{B_k^+}\left(\frac{1}{z},\frac{1}{a_k}\right)=\ln\frac{1}{|\frac{1}{z}-a_k^+|}+\ln r(B_k^+,a_k^+)+o(1)$$

we obtain

$$r\left(B_k^+,a_k^+\right)=\frac{r\left(B_k,a_k\right)}{|a_k|^2} \tag{7}$$

and the inequality

$$r\left(B_0,0\right)\le\frac{1}{\left[\sum\limits_{k=1}^{n}\frac{r^2(B_k,a_k)}{|a_k|^4}\right]^{\frac{1}{2}}}.$$

From whence, the following estimate holds

$$r^n(B_0,0)\prod_{k=1}^{n} r(B_k,a_k) \le \frac{\prod\limits_{k=1}^{n} r(B_k,a_k)}{\left[\sum\limits_{k=1}^{n}\frac{r^2(B_k,a_k)}{|a_k|^4}\right]^{\frac{n}{2}}}.$$

Then from the Cauchy inequality of arithmetic and geometric means, we have the assertion

$$\frac{1}{n}\sum_{k=1}^{n}\frac{r^2(B_k,a_k)}{|a_k|^4} \ge \left(\prod_{k=1}^{n}\frac{r^2(B_k,a_k)}{|a_k|^4}\right)^{\frac{1}{n}}.$$

It is clear that

$$\left(\sum_{k=1}^{n}\frac{r^2(B_k,a_k)}{|a_k|^4}\right)^{\frac{n}{2}} \ge n^{\frac{n}{2}}\prod_{k=1}^{n}\frac{r(B_k,a_k)}{|a_k|^2}.$$

Thus, it follows that

$$r^n(B_0,0)\prod_{k=1}^{n} r(B_k,a_k) \le n^{-\frac{n}{2}}\left(\prod_{k=1}^{n}|a_k|\right)^2. \tag{8}$$

Using inequality (8) and conditions of the Theorem 2

$$r^n(B_0,0)\prod_{k=1}^{n} r(B_k,a_k) = (r(B_0,0)\,r(B_k,a_k))^n \le n^{-\frac{n}{2}}\left(\prod_{k=1}^{n}|a_k|\right)^2.$$

Taking from the left and right parts of the last inequality the root of degree n, the inequality (2) holds. Theorem 2 is proved. □

Corollary 1 *Let $n \in \mathbb{N}$ and a fixed system of different points $\{a_k\}_{k=1}^{n}$ such that $|a_k| = \rho \in \mathbb{R}^+$, $k = \overline{1,n}$. Then for all conditions of the Theorem 2, the inequality*

$$r(B_0,0)\,r(B_k,a_k) \le \frac{1}{\sqrt{n}}\cdot\rho^2, \quad k = \overline{1,n}$$

is valid.

Remark 1 On the other hand, for all conditions of the Corollary 1 from Lavrentyev's result [8] (see also Theorem 1) the following inequality holds

$$r(B_0,0)\,r(B_k,a_k) \le \rho^2, \quad k = \overline{1,n}.$$

Theorem 3 *Let $n \in \mathbb{N}$. Then for any fixed system of different points $\{a_k\}_{k=1}^n$ and for any collection of mutually non-overlapping domains B_∞, $\{B_k\}_{k=1}^n$, $\infty \in B_\infty \subset \overline{\mathbb{C}}$, $a_k \in B_k \subset \overline{\mathbb{C}}$, $k = \overline{1,n}$, such that $r(B_1, a_1) = r(B_2, a_2) = \ldots = r(B_n, a_n)$, the following inequality holds*

$$r(B_\infty, \infty)\, r(B_k, a_k) \leq \frac{1}{\sqrt{n}}, \quad k = \overline{1,n}. \tag{9}$$

Proof Consider the product

$$r^n(B_\infty, \infty) \prod_{k=1}^{n} r(B_k, a_k)$$

for any fixed system of different points $\{a_k\}_{k=1}^n$ and for any collection of mutually non-overlapping domains B_∞, $\{B_k\}_{k=1}^n$, $\infty \in B_\infty \subset \overline{\mathbb{C}}$, $a_k \in B_k \subset \overline{\mathbb{C}}$, $k = \overline{1,n}$.

The proof of the Theorem 3 is based on the constructions given above in the proof of the Theorem 2. Using inequalities (4), (5) and the area-minimization theorem [1, 2], the relation holds

$$r(B_\infty, \infty) \leq \frac{1}{\left[\sum\limits_{k=1}^{n} r^2(B_k, a_k)\right]^{\frac{1}{2}}}.$$

Then from the Cauchy inequality of arithmetic and geometric means, we have

$$\left(\sum_{k=1}^{n} r^2(B_k, a_k)\right)^{\frac{1}{2}} \geq n^{\frac{1}{2}} \left[\prod_{k=1}^{n} r(B_k, a_k)\right]^{\frac{1}{n}}$$

and therefore

$$r^n(B_\infty, \infty) \leq n^{-\frac{n}{2}} \left[\prod_{k=1}^{n} r(B_k, a_k)\right]^{-1}.$$

From here

$$r^n(B_\infty, \infty) \prod_{k=1}^{n} r(B_k, a_k) \leq n^{-\frac{n}{2}}.$$

Using last inequality and conditions of the Theorem 3

$$r^n(B_\infty, \infty) \prod_{k=1}^{n} r(B_k, a_k) = (r(B_\infty, \infty)\, r(B_k, a_k))^n \leq n^{-\frac{n}{2}}.$$

Taking from the left and right parts of the last inequality the root of degree n, the inequality (9) holds. Theorem 3 is proved. □

Remark 2 For all conditions of the Theorem 3 from the Lavrentyev theorem, the inequality

$$r\left(B_{\infty},\infty\right)r\left(B_k,a_k\right)\leq 1,\quad k=\overline{1,n}$$

is true.

Theorem 4 *Let $n\in\mathbb{N}$. Then for any fixed system of different points $\{a_k\}_{k=1}^n$ and for any collection of mutually non-overlapping domains B_0, B_{∞}, $\{B_k\}_{k=1}^n$, $a_0=0\in B_0\subset\overline{\mathbb{C}}$, $\infty\in B_{\infty}\subset\overline{\mathbb{C}}$, $a_k\in B_k\subset\overline{\mathbb{C}}$, $k=\overline{1,n}$, such that $r\left(B_1,a_1\right)=r\left(B_2,a_2\right)=\ldots=r\left(B_n,a_n\right)$, the following inequality holds*

$$r\left(B_0,0\right)r\left(B_{\infty},\infty\right)r\left(B_k,a_k\right)\leq(n+1)^{-\frac{n+1}{2n}}\left(\prod_{k=1}^{n}|a_k|\right)^{\frac{1}{n}}. \tag{10}$$

Proof Consider the product

$$\left[r\left(B_0,0\right)r\left(B_{\infty},\infty\right)\right]^n\prod_{k=1}^{n}r\left(B_k,a_k\right)$$

for any fixed system of different points $\{a_k\}_{k=1}^n$ and for any collection of mutually non-overlapping domains B_0, B_{∞}, $\{B_k\}_{k=1}^n$, $a_0=0\in B_0\subset\overline{\mathbb{C}}$, $\infty\in B_{\infty}\subset\overline{\mathbb{C}}$, $a_k\in B_k\subset\overline{\mathbb{C}}$, $k=\overline{1,n}$.

Using the arguments in proving of the Theorem 2, inequalities (4), (5) and the area-minimization theorem [1, 2], it follows that

$$r\left(B_0,0\right)\leq\left(r^2\left(B_{\infty},\infty\right)+\sum_{k=1}^{n}r^2\left(B_k^{+},a_k^{+}\right)\right)^{-\frac{1}{2}}.$$

From here, using the equality (7), we have

$$r\left(B_0,0\right)\leq\left[r^2\left(B_{\infty},\infty\right)+\sum_{k=1}^{n}\frac{r^2\left(B_k,a_k\right)}{|a_k|^4}\right]^{-\frac{1}{2}}.$$

Similarly,

$$r\left(B_{\infty},\infty\right)\leq\left[r^2\left(B_0,0\right)+\sum_{k=1}^{n}r^2\left(B_k,a_k\right)\right]^{-\frac{1}{2}}.$$

Taking into account the Cauchy inequality of arithmetic and geometric means

$$r\left(B_0,0\right)\leq\frac{\left(\prod\limits_{k=1}^{n}|a_k|\right)^{\frac{2}{n+1}}}{(n+1)^{\frac{1}{2}}\left(r\left(B_\infty,\infty\right)\right)^{\frac{1}{n+1}}\left(\prod\limits_{k=1}^{n}r\left(B_k,a_k\right)\right)^{\frac{1}{n+1}}}$$

and, analogically,

$$r\left(B_\infty,\infty\right)\leq\frac{1}{(n+1)^{\frac{1}{2}}\left(r\left(B_0,0\right)\right)^{\frac{1}{n+1}}\left(\prod\limits_{k=1}^{n}r\left(B_k,a_k\right)\right)^{\frac{1}{n+1}}}.$$

Using simple transformations, we obtain the relation

$$r\left(B_0,0\right)r\left(B_\infty,\infty\right)\leq\frac{\left(\prod\limits_{k=1}^{n}|a_k|\right)^{\frac{2}{n+2}}}{(n+1)^{\frac{n+1}{n+2}}\left(\prod\limits_{k=1}^{n}r\left(B_k,a_k\right)\right)^{\frac{2}{n+2}}}$$

from which the inequality follows

$$\left[r\left(B_0,0\right)r\left(B_\infty,\infty\right)\right]^n\prod_{k=1}^{n}r\left(B_k,a_k\right)\leq$$

$$\leq(n+1)^{-\frac{n(n+1)}{n+2}}\left(\prod_{k=1}^{n}r\left(B_k,a_k\right)\right)^{1-\frac{2n}{n+2}}\left(\prod_{k=1}^{n}|a_k|\right)^{\frac{2n}{n+2}}.$$

According to the Theorem 1 (see also Theorem 2.3.1 [3]), we get

$$r\left(B_0,0\right)r\left(B_\infty,\infty\right)\leq 1,$$

then

$$\left[r\left(B_0,0\right)r\left(B_\infty,\infty\right)\right]^n\prod_{k=1}^{n}r\left(B_k,a_k\right)\leq$$

$$\leq\frac{(n+1)^{-\frac{n(n+1)}{n+2}}}{\left(\left[r\left(B_0,0\right)r\left(B_\infty,\infty\right)\right]^n\prod\limits_{k=1}^{n}r\left(B_k,a_k\right)\right)^{\frac{2n}{n+2}-1}}\prod_{k=1}^{n}|a_k|^{\frac{2n}{n+2}}.$$

From here there follows that

$$\left([r(B_0,0)\,r(B_\infty,\infty)]^n\prod_{k=1}^{n}r(B_k,a_k)\right)^{\frac{2n}{n+2}}\le(n+1)^{-\frac{n(n+1)}{n+2}}\prod_{k=1}^{n}|a_k|^{\frac{2n}{n+2}}.$$

Therefore,

$$[r(B_0,0)\,r(B_\infty,\infty)]^n\prod_{k=1}^{n}r(B_k,a_k)\le(n+1)^{-\frac{n+1}{2}}\prod_{k=1}^{n}|a_k|. \tag{11}$$

From conditions of the Theorem 4

$$[r(B_0,0)\,r(B_\infty,\infty)]^n\prod_{k=1}^{n}r(B_k,a_k)=(r(B_0,0)\,r(B_\infty,\infty)\,r(B_k,a_k))^n.$$

Using inequality (11), it follows that

$$(r(B_0,0)\,r(B_\infty,\infty)\,r(B_k,a_k))^n\le(n+1)^{-\frac{n+1}{2}}\prod_{k=1}^{n}|a_k|.$$

Taking from the left and right parts of the last inequality the root of degree n, the inequality (10) holds. Theorem 4 is proved. □

Corollary 2 *Let $n\in\mathbb{N}$ and a fixed system of different points $\{a_k\}_{k=1}^n$ such that $|a_k|=\rho\in\mathbb{R}^+$, $k=\overline{1,n}$. Then for all conditions of the Theorem 4, the inequality*

$$r(B_0,0)\,r(B_\infty,\infty)\,r(B_k,a_k)\le(n+1)^{-\frac{n+1}{2n}}\cdot\rho,\quad k=\overline{1,n}$$

is valid.

References

1. Goluzin, G.M.: Geometric Theory of Functions of a Complex Variable. American Mathematical Society, Providence (1969)
2. Dubinin, V.N.: Condenser Capacities and Symmetrization in Geometric Function Theory. Birkhäuser/Springer, Basel (2014)
3. Bakhtin, A.K., Bakhtina, G.P., Zelinskii, Y.B.: Topological-Algebraic Structures and Geometric Methods in Complex Analysis (in Russian). Zb. Prats of the Inst. of Math. of NASU (2008)
4. Hayman, W.K.: Multivalent Functions. Cambridge University Press, Cambridge (1994)
5. Jenkins, A.: Univalent Functions and Conformal Mappings. Springer, Berlin (1958)
6. Strebel, K.: Quadratic Differentials. Springer, Berlin (1984)
7. Duren, P.: Univalent Functions. Springer, Heidelberg (1983)

8. Lavrentyev, M.A.: On the theory of conformal mappings (in Russian). Tr. Sci. Inst An USSR **5**, 159–245 (1934)
9. Schiffer, M., Spencer, D.C.: Functionals of Finite Riemann Surfaces. Princeton University Press, Princeton (1954)
10. Kolbina, L.I.: Some extremal problems in conformal mapping (in Russian). Dokl. Akad. Nauk SSSR, Ser. Mat. **84**, 865–868 (1952)
11. Kolbina, L.I.: Conformal mapping of a unit circle onto nonoverlapping domains (in Russian). Vestn. Lenin. Univ. **5**, 37–43 (1955)
12. Duren, P.L., Schiffer M.: Conformal mappings onto non-overlapping regions. In: Complex Analysis, pp. 27–39. Birkhauser Verlag, Basel (1988)
13. Schaeffer, A.C., Spencer, D.C.: Coefficient Regions for Schlicht Functions. Amer. Math. Soc. Coll. Publ., New York (1950)
14. Schiffer, M.: A method of variation within the family of simple functions. Proc. Lond. Math. Soc. **44**, 432–449 (1938)
15. Tamrazov, P.M.: Extremal conformal mappings and poles of quadratic differentials. Izv. Akad. Nauk SSSR, Ser. Mat. **32**, 1033–1043 (1968)
16. Kuz'mina, G.V.: Methods of the geometric theory of functions. I, II. St. Petersburg Math. J. **9**, 455–507 (1998); **9**, 889–930 (1998)
17. Bakhtin, A.K.: Separating transformation and extremal problems on nonoverlapping simply connected domains. J. Math. Sci. **234**, 1–13 (2018)
18. Bakhtin, A.K.: Extremal decomposition of the complex plane with restrictions for free poles. J. Math. Sci. **231**, 1–15 (2018)
19. Kovalev, L.V.: On three disjoint domains (in Russian). Dalnevost Mat. Sb. **1**, 3–7 (2000)
20. Bakhtin, A.K., Denega, I.V.: Inequalities for the inner radii of nonoverlapping domains. Ukr. Math. J. **71**, 1138–1145 (2019)
21. Denega, I.: Estimates of the inner radii of non-overlapping domains. J. Math. Sci. **242**, 787–795 (2019)
22. Bakhtin, A.K., Denega, I.V.: Weakened problem on extremal decomposition of the complex plane. Mat. Stud. **51**, 35–40 (2019)
23. Bakhtin, A.K.: A problem of extreme decomposition of the complex plane with free poles. Ukr. Math. J. **71**, 1485–1509 (2020)
24. Bakhtin, A.K., Denega, I.V.: Extremal decomposition of the complex plane with free poles. J. Math. Sci. **246**, 1–17 (2020)
25. Bakhtin, A.K., Denega, I.V.: Extremal decomposition of the complex plane with free poles II. J. Math. Sci. **246**, 602–616 (2020)
26. Polya, G., Szego, G.: Isoperimetric Inequalities in Mathematical Physics. Princeton University Press, Princeton (1962)

Schwarz Lemma Type Estimates for Solutions to Nonlinear Beltrami Equation

Bogdan Klishchuk, Ruslan Salimov, and Mariia Stefanchuk

Abstract We continue to investigate the regular homeomorphic solutions to nonlinear Beltrami equation introduced in Golberg and Salimov (Complex Var Elliptic Equ 65(1):6–21, 2020). Schwarz Lemma type estimates are obtained involving the length-area method. The lower bounds for the inverses are also established.

1 Introduction

Let G be a domain in the complex plane $\mathbb{C}$ and $\mu : G \to \mathbb{C}$ be a measurable function with $|\mu(z)| < 1$ a.e. (almost everywhere) in G. Recall that the *Beltrami equation* has a form

$$f_{\overline{z}} = \mu(z) f_z \,, \tag{1}$$

where $f_{\overline{z}} = (f_x + i f_y)/2$, $f_z = (f_x - i f_y)/2$, $z = x + iy$, and f_x and f_y are the partial derivatives of f by x and y, respectively.

Various existence theorems for solutions of the Sobolev class $W^{1,1}_{\text{loc}}$ have been recently established applying the modulus approach for a quite wide class of linear and quasilinear degenerate Beltrami equations; see, e.g. [1–5].

Let $\sigma : G \to \mathbb{C}$ be a measurable function and $m \geqslant 0$. We consider the following equation written in the polar coordinates (r, θ) :

$$f_r = \sigma(re^{i\theta})\, |f_\theta|^m \, f_\theta, \tag{2}$$

B. Klishchuk (✉) · R. Salimov · M. Stefanchuk
National Academy of Sciences of Ukraine, Kiev, Ukraine

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_22

where f_r and f_θ are the partial derivatives of f by r and θ, respectively. Applying the relations between these derivatives and the formal derivatives $rf_r = zf_z + \overline{z}f_{\overline{z}}$, and $f_\theta = i(zf_z - \overline{z}f_{\overline{z}})$, see, e.g. [6, (21.25)], one can rewrite Eq. (2) in the form:

$$f_{\overline{z}} = \frac{z}{\overline{z}} \frac{\sigma(z)\,|z||zf_z - \overline{z}f_{\overline{z}}|^m + i}{\sigma(z)\,|z||zf_z - \overline{z}f_{\overline{z}}|^m - i} f_z \tag{3}$$

with the condition $\overline{z}\,(\sigma(z)\,|z||zf_z - \overline{z}f_{\overline{z}}|^m - i) \neq 0$ a.e. Under $m = 0$ Eq. (3) reduces to the standard linear Beltrami equation (1). Picking $m = 0$ and $\sigma = -i/|z|$ in (3), we arrive at the classical Cauchy-Riemann system. For $m > 0$ Eq. (3) provides a partial case of the general nonlinear system of equations (7.33) given in [6, Sect. 7.7]. Later on we assume that $m > 0$.

The nonlinear equation (3) provides a partial case of the nonlinear system of two real partial differential equations; see (1) in [8, 9], cf. [10]. Note that various nonlinear systems of partial differential equations studied in a quite large specter of aspects can be found in [6–24].

A mapping $f : G \to \mathbb{C}$ is called *regular at a point* $z_0 \in G$, if f has the total differential at this point and its Jacobian $J_f = |f_z|^2 - |f_{\overline{z}}|^2$ does not vanish, cf. [25, I. 1.6]. A homeomorphism f of Sobolev class $W^{1,1}_{\rm loc}$ is called *regular*, if $J_f > 0$ a.e. By a *regular solution* of equation (3) we call a regular homeomorphism $f : G \to \mathbb{C}$, which satisfies (3) a.e. in G.

Further we use the following notations

$$B_r = \{z \in \mathbb{C} : |z| < r\}, \quad \mathbb{B} = \{z \in \mathbb{C} : |z| < 1\}$$

and

$$\gamma_r = \{z \in \mathbb{C} : |z| = r\}, \quad \mathbb{A}(0, r_1, r_2) = \{z \in \mathbb{C} : r_1 < |z| < r_2\}.$$

The area of set $f(B_r)$ we denote by $S(r) = |f(B_r)|$.

The following statement provides a differential inequality for the function $S(r)$; see Lemma 1 in [22].

Proposition 1 *Let $m > 0$ and $f : \mathbb{B} \to \mathbb{C}$ be a regular homeomorphic solution to equality (3) of Sobolev class $W^{1,2}_{\rm loc}$ normalized by $f(0) = 0$. Then*

$$S'(r) \geqslant 2^{m+2}\pi^{\frac{m}{2}+1} \left(\int\limits_{\gamma_r} \frac{ds}{|z| \left(\operatorname{Im} \overline{\sigma(z)} \right)^{\frac{1}{m+1}}} \right)^{-(m+1)} S^{\frac{m+2}{2}}(r) \tag{4}$$

for almost all $r \in [0, 1)$.

2 Main Results

In this section we provide a series of theorems related to the asymptotic behavior of regular homeomorphic solutions to nonlinear equation (3).

The following result is an analogue of the well-known Ikoma-Schwartz lemma on estimating the lower bound [27, Theorem 2].

Theorem 1 *Let $f: \mathbb{B} \to \mathbb{C}$ be a regular homeomorphic solution to nonlinear equation (3) of Sobolev class $W^{1,2}_{\mathrm{loc}}$ satisfying $f(0) = 0$.*

(a) If for some $\varepsilon_0 \in (0, 1)$, $\alpha > 0$ and $C > 0$

$$\int\limits_{\varepsilon}^{\varepsilon_0} \frac{dt}{I_{m,\sigma}(t)} \sim C\,\varepsilon^{-\alpha} \quad as \quad \varepsilon \to 0\,, \tag{5}$$

where $I_{m,\sigma}(t) = \left(\int_{\gamma_t} \frac{ds}{|z|\left(\operatorname{Im}\overline{\sigma(z)}\right)^{1/(m+1)}}\right)^{m+1}$, then the following estimate

$$\liminf_{z\to 0} \frac{|f(z)|}{|z|^{\alpha/m}} \leqslant c_0\, C^{-1/m} < \infty \tag{6}$$

holds, where $c_0 = (2\pi)^{-(m+1)/m} m^{-1/m}$.

(b) If for some $\varepsilon_0 \in (0, 1)$ and $\alpha > 0$

$$\lim_{\varepsilon\to 0} \varepsilon^{\alpha} \int\limits_{\varepsilon}^{\varepsilon_0} \frac{dt}{I_{m,\sigma}(t)} = \infty\,, \tag{7}$$

then

$$\liminf_{z\to 0} \frac{|f(z)|}{|z|^{\alpha/m}} = 0\,. \tag{8}$$

Proof Proposition 1 implies

$$\left(\frac{S^{-\frac{m}{2}}(t)}{-\frac{m}{2}}\right)' = \frac{S'(t)}{S^{\frac{m+2}{2}}(t)} \geqslant 2^{m+2}\pi^{\frac{m}{2}+1} \left(\int\limits_{\gamma_t} \frac{ds}{|z|\left(\operatorname{Im}\overline{\sigma(z)}\right)^{\frac{1}{m+1}}}\right)^{-(m+1)}$$

for almost all $t \in (0, 1)$. Integrating the last inequality by $t \in (\varepsilon, \varepsilon_0)$, $0 < \varepsilon < \varepsilon_0 < 1$, we have

$$\int_{\varepsilon}^{\varepsilon_0} \left(\frac{S^{-\frac{m}{2}}(t)}{-\frac{m}{2}} \right)' dt \geqslant 2^{m+2} \pi^{\frac{m}{2}+1} \int_{\varepsilon}^{\varepsilon_0} \left(\int_{\gamma_t} \frac{ds}{|z| \left(\operatorname{Im} \overline{\sigma(z)} \right)^{\frac{1}{m+1}}} \right)^{-(m+1)} dt . \tag{9}$$

Note that $g_m(t) = \frac{S^{-\frac{m}{2}}(t)}{-\frac{m}{2}}$ is nondecreasing. Then

$$\int_{\varepsilon}^{\varepsilon_0} \left(\frac{S^{-\frac{m}{2}}(t)}{-\frac{m}{2}} \right)' dt = \int_{\varepsilon}^{\varepsilon_0} g_m'(t) dt \leqslant g_m(\varepsilon_0) - g_m(\varepsilon) = \frac{2}{m} \left(S^{-\frac{m}{2}}(\varepsilon) - S^{-\frac{m}{2}}(\varepsilon_0) \right); \tag{10}$$

see [28, Theorem IV 7.4]).

Combining estimates (9) and (10), one gets

$$S^{-\frac{m}{2}}(\varepsilon) \geqslant S^{-\frac{m}{2}}(\varepsilon) - S^{-\frac{m}{2}}(\varepsilon_0) \geqslant 2^{m+1} m \pi^{\frac{m}{2}+1} \int_{\varepsilon}^{\varepsilon_0} \frac{dt}{I_{m,\sigma}(t)},$$

where $I_{m,\sigma}(t) = \left(\int_{\gamma_t} \frac{ds}{|z| \left(\operatorname{Im} \overline{\sigma(z)} \right)^{1/(m+1)}} \right)^{m+1}$. This yields the bound

$$S(\varepsilon) \leqslant 2^{-\frac{2(m+1)}{m}} m^{-\frac{2}{m}} \pi^{-\frac{m+2}{m}} \left(\int_{\varepsilon}^{\varepsilon_0} \frac{dt}{I_{m,\sigma}(t)} \right)^{-\frac{2}{m}}. \tag{11}$$

Letting $l_f(\varepsilon) = \min_{|z|=\varepsilon} |f(z)|$, $\varepsilon \in (0, \varepsilon_0)$, and since f is a homeomorphism satisfying $f(0) = 0$,

$$\pi l_f^2(\varepsilon) \leqslant S(\varepsilon) \leqslant 2^{-\frac{2(m+1)}{m}} m^{-\frac{2}{m}} \pi^{-\frac{m+2}{m}} \left(\int_{\varepsilon}^{\varepsilon_0} \frac{dt}{I_{m,\sigma}(t)} \right)^{-\frac{2}{m}}.$$

Hence,

$$l_f(\varepsilon) \left(\int_{\varepsilon}^{\varepsilon_0} \frac{dt}{I_{m,\sigma}(t)} \right)^{\frac{1}{m}} \leqslant 2^{-\frac{m+1}{m}} m^{-\frac{1}{m}} \pi^{-\frac{m+1}{m}}.$$

And thus,

$$\liminf_{z\to 0} |f(z)| \left(\int\limits_{|z|}^{\varepsilon_0} \frac{dt}{I_{m,\sigma}(t)} \right)^{\frac{1}{m}} = \liminf_{\varepsilon\to 0} l_f(\varepsilon) \left(\int\limits_{\varepsilon}^{\varepsilon_0} \frac{dt}{I_{m,\sigma}(t)} \right)^{\frac{1}{m}} \leqslant c_0\,, \tag{12}$$

where $c_0 = 2^{-\frac{m+1}{m}}\, m^{-\frac{1}{m}}\, \pi^{-\frac{m+1}{m}} > 0$ is a constant depending only on m.

In view of the condition (5),

$$\liminf_{z\to 0} \frac{|f(z)|}{|z|^{\alpha/m}} = \liminf_{z\to 0} |f(z)| \left(\int\limits_{|z|}^{\varepsilon_0} \frac{dt}{I_{m,\sigma}(t)} \right)^{\frac{1}{m}} \cdot \lim_{z\to 0} |z|^{-\frac{\alpha}{m}} \left(\int\limits_{|z|}^{\varepsilon_0} \frac{dt}{I_{m,\sigma}(t)} \right)^{-\frac{1}{m}} \leqslant$$

$$\leqslant c_0\, C^{-1/m}\,.$$

It is obvious that from condition (7) and from the last estimate statement b) follows.

Letting in Theorem 1 $\alpha = m$, we derive the following statement.

Corollary 1 *If for $C > 0$*

$$\int\limits_{\varepsilon}^{\varepsilon_0} \frac{dt}{I_{m,\sigma}(t)} \sim C\,\varepsilon^{-m} \quad \text{as} \quad \varepsilon \to 0\,, \tag{13}$$

then

$$\liminf_{z\to 0} \frac{|f(z)|}{|z|} \leqslant c_0\, C^{-1/m} < \infty\,, \tag{14}$$

where $c_0 = (2\pi)^{-(m+1)/m} m^{-1/m}$.

As consequences of Theorem 1, we obtain the following statements.

Theorem 2 *Let $f\colon \mathbb{B} \to \mathbb{C}$ be a regular homeomorphic solution to nonlinear equation (3) of Sobolev class $W^{1,2}_{\rm loc}$ satisfying $f(0) = 0$ and $\operatorname{Im} \overline{\sigma(re^{i\theta})} \geqslant \lambda(r)$ for a.a. $r \in (0, \varepsilon_0)$, $\varepsilon_0 \in (0, 1)$, where $\lambda(r)\colon [0, 1) \to [0, \infty)$ is a measurable function.*

(a) If for $C > 0$ and $\alpha > 0$

$$\int\limits_{\varepsilon}^{\varepsilon_0} \lambda(t)\, dt \sim C\,\varepsilon^{-\alpha} \quad \text{as} \quad \varepsilon \to 0\,, \tag{15}$$

then

$$\liminf_{z\to 0} \frac{|f(z)|}{|z|^{\alpha/m}} \leqslant (C\,m)^{-1/m} < \infty\,. \tag{16}$$

(b) If for $C > 0$ and $\alpha > 0$

$$\lim_{\varepsilon\to 0} \varepsilon^{\alpha} \int\limits_{\varepsilon}^{\varepsilon_0} \lambda(t)\,dt = \infty\,, \tag{17}$$

then

$$\liminf_{z\to 0} \frac{|f(z)|}{|z|^{\alpha/m}} = 0. \tag{18}$$

Proof Indeed, the condition $\operatorname{Im}\overline{\sigma(re^{i\theta})} \geqslant \lambda(r)$ for a.a. $r \in (0, \varepsilon_0)$ yields

$$I_{m,\sigma}(t) = \left(\int\limits_{\gamma_t} \frac{ds}{|z| \left(\operatorname{Im}\overline{\sigma(z)}\right)^{\frac{1}{m+1}}} \right)^{m+1} \leqslant \frac{(2\pi)^{m+1}}{\lambda(t)}$$

for a.a. $t \in (0, \varepsilon_0)$. Hence the following bound

$$\left(\int\limits_{|z|}^{\varepsilon_0} \frac{dt}{I_{m,\sigma}(t)} \right)^{\frac{1}{m}} \geqslant (2\pi)^{-\frac{m+1}{m}} \left(\int\limits_{|z|}^{\varepsilon_0} \lambda(t)\,dt \right)^{\frac{1}{m}}$$

holds for all z such that $0 < |z| < \varepsilon_0$.

Combining the last estimate and estimate (12) from the proof of Theorem 1, we obtain

$$\liminf_{z\to 0} |f(z)| \left(\int\limits_{|z|}^{\varepsilon_0} \lambda(t)\,dt \right)^{\frac{1}{m}} \leqslant (2\pi)^{\frac{m+1}{m}} \liminf_{z\to 0} |f(z)| \left(\int\limits_{|z|}^{\varepsilon_0} \frac{dt}{I_{m,\sigma}(t)} \right)^{\frac{1}{m}} \leqslant$$

$$\leqslant (2\pi)^{\frac{m+1}{m}} c_0 = m^{-\frac{1}{m}},$$

where $c_0 = (2\pi)^{-\frac{m+1}{m}} m^{-\frac{1}{m}}$.

Finally, in view of the condition (15), we get

$$\liminf_{z\to 0}\frac{|f(z)|}{|z|^{\alpha/m}}=\liminf_{z\to 0}|f(z)|\left(\int\limits_{|z|}^{\varepsilon_0}\lambda(t)\,dt\right)^{\frac{1}{m}}\lim_{z\to 0}\left(|z|^{\alpha}\int\limits_{|z|}^{\varepsilon_0}\lambda(t)\,dt\right)^{-\frac{1}{m}}\leqslant$$

$$\leqslant (C\,m)^{-1/m}\,.$$

Letting in Theorem 2 $\alpha=m$, we derive the following statement.

Corollary 2 *If for* $C>0$

$$\int\limits_{\varepsilon}^{\varepsilon_0}\lambda(t)\,dt\sim C\,\varepsilon^{-m}\quad as\quad \varepsilon\to 0\,, \tag{19}$$

then

$$\liminf_{z\to 0}\frac{|f(z)|}{|z|}\leqslant (C\,m)^{-1/m}<\infty\,. \tag{20}$$

Consider the equation

$$f_r=-\frac{\alpha i}{mk^m r^{\alpha+1}}\,|f_\theta|^m\,f_\theta,\quad \alpha>0,\ k>0, \tag{21}$$

in the unit disk $\mathbb{B}$.

It is easy to check that $f=k\,r^{\frac{\alpha}{m}}e^{i\theta}$ is a regular homeomorphic solution to equation (21) of Sobolev class $W^{1,2}_{\rm loc}(\mathbb{B})$. Further, we have

$$\sigma=-\frac{\alpha i}{mk^m r^{\alpha+1}}\,,\quad \left(\operatorname{Im}\overline{\sigma(z)}\right)^{\frac{1}{m+1}}=\left(\frac{\alpha}{m}\right)^{\frac{1}{m+1}}\left(\frac{1}{k}\right)^{\frac{m}{m+1}}\frac{1}{|z|^{\frac{\alpha+1}{m+1}}}$$

and

$$I_{m,\sigma}(t)=\left(\int\limits_{\gamma_t}\frac{ds}{|z|\left(\operatorname{Im}\overline{\sigma(z)}\right)^{\frac{1}{m+1}}}\right)^{m+1}=\frac{(2\pi)^{m+1}mk^m}{\alpha}\cdot t^{\alpha+1}.$$

We obviously get $\int\limits_{\varepsilon}^{\varepsilon_0}\frac{dt}{I_{m,\sigma}(t)}\sim C\,\varepsilon^{-\alpha}$ as $\varepsilon\to 0$, where $C=\frac{1}{(2\pi)^{m+1}k^m m}$.

Thus, the conditions of Theorem 1(a) are satisfied. On the other hand, it is easy to see that $\lim\limits_{z\to 0}\frac{|f(z)|}{|z|^{\alpha/m}}=k$.

Consider the equation

$$f_r = -\frac{(m+\alpha)i}{mk^m}\,\frac{1}{r^{1+m+\alpha}}\,|f_\theta|^m\,f_\theta\,,\quad k>0\,,\quad \alpha>0\,, \tag{22}$$

in the unit disk $\mathbb{B}$.

It is easy to check that $f = k\,r^{1+\frac{\alpha}{m}}e^{i\theta}$ is a regular homeomorphic solution to equation (22) of Sobolev class $W^{1,2}_{\mathrm{loc}}(\mathbb{B})$. Further, we have

$$|z|\,\left(\mathrm{Im}\,\overline{\sigma(z)}\right)^{\frac{1}{m+1}} = \left(\frac{m+\alpha}{mk^m}\right)^{\frac{1}{m+1}}\,\frac{1}{|z|^{\frac{\alpha}{m+1}}}$$

and $I_{m,\sigma}(t) = (2\pi)^{m+1}\,\frac{mk^m}{m+\alpha}\,t^{\alpha+m+1}$. Thus, $\lim\limits_{\varepsilon\to 0}\varepsilon^\alpha\int\limits_\varepsilon^{\varepsilon_0}\frac{dt}{I_{m,\sigma}(t)} = \infty$. On the other hand, clearly $\lim\limits_{z\to 0}\frac{|f(z)|}{|z|^{\alpha/m}} = 0\,.$

Proposition 2 *Equation (2) with $\sigma = -\frac{i}{m\,r}$, $0<m<2$, has a homeomorphic solution $f:\mathbb{B}\to\mathbb{C}$ of Sobolev class $W^{1,2}_{\mathrm{loc}}(\mathbb{B})$ satisfying $f(0)=0$, such that $\lim\limits_{z\to 0}\frac{|f(z)|}{|z|^{\alpha/m}} = \infty$ for all $\alpha>0$ and, moreover, $\lim\limits_{\varepsilon\to 0}\varepsilon^\alpha\int\limits_\varepsilon^{\varepsilon_0}\frac{dt}{I_{m,\sigma}(t)} = 0$.*

Proof Consider the equation

$$f_r = -\frac{i}{mr}\cdot|f_\theta|^m\cdot f_\theta \tag{23}$$

in the unit disk $\mathbb{B}$.

It is easy to check that $f = \left(\ln\frac{e}{r}\right)^{-1/m}e^{i\theta}$ is a regular homeomorphic solution of equation (23). We show first that the mapping $f = \left(\ln\frac{e}{r}\right)^{-1/m}e^{i\theta}$ belongs to Sobolev class $W^{1,2}_{\mathrm{loc}}(\mathbb{B})$. Indeed, f is a diffeomorphism in $\mathbb{B}\setminus\{0\}$, and, therefore, $f\in W^{1,2}_{\mathrm{loc}}(\mathbb{B}\setminus\{0\})$. Let $\overline{B}_{r_0} = \{z\in\mathbb{C}:|z|\leqslant r_0\}$, $r_0\in(0,1)$. The partial derivatives of f by r and θ are

$$f_r = \frac{1}{m\,r}\left(\ln\frac{e}{r}\right)^{-1/m-1}\,e^{i\theta}\,,\quad f_\theta = \left(\ln\frac{e}{r}\right)^{-1/m}ie^{i\theta}\,,$$

and using the formula [6, p. 611], we have

$$\int\limits_{\overline{B}_{r_0}}\left(|f_z|^2+|f_{\overline{z}}|^2\right)dxdy = \frac{1}{2}\int\limits_{\overline{B}_{r_0}}\left(|f_r|^2+r^{-2}|f_\theta|^2\right)rdrd\theta =$$

$$= \frac{\pi}{m^2}\int\limits_0^{r_0}\left(\ln\frac{e}{r}\right)^{-2/m-2}\frac{dr}{r}+\pi\int\limits_0^{r_0}\left(\ln\frac{e}{r}\right)^{-2/m}\frac{dr}{r}\,.$$

Obviously, both integrals converge under $0 < m < 2$.

Further, we have

$$\left(\operatorname{Im}\overline{\sigma(z)}\right)^{\frac{1}{m+1}} = \left(\frac{1}{m}\right)^{\frac{1}{m+1}} \frac{1}{|z|^{\frac{1}{m+1}}}$$

and $I_{m,\sigma}(t) = (2\pi)^{m+1}\, m\, t$. Obviously, $\lim\limits_{\varepsilon\to 0} \varepsilon^{\alpha} \int\limits_{\varepsilon}^{\varepsilon_0} \frac{dt}{I_{m,\sigma}(t)} = 0$. On the other hand, clearly, that $\lim\limits_{z\to 0} \frac{|f(z)|}{|z|^{\alpha/m}} = \infty$ for all $\alpha > 0$.

3 Consequences for Inverse Solves

Here we formulate some results for the inverse mappings applying the statements of previous section. Obviously, the lower bounds will be derived instead of upper ones.

Theorem 3 *Let $f\colon \mathbb{B} \to \mathbb{C}$ be a regular homeomorphic solution to nonlinear equation (3) of Sobolev class $W^{1,2}_{\mathrm{loc}}$ satisfying $f(0) = 0$.*

(a) If for some $\varepsilon_0 \in (0, 1)$, $\alpha > 0$ and $C > 0$

$$\int\limits_{\varepsilon}^{\varepsilon_0} \frac{dt}{I_{m,\sigma}(t)} \sim C\, \varepsilon^{-\alpha} \quad as \quad \varepsilon \to 0\,,$$

where $I_{m,\sigma}(t) = \left(\int_{\gamma_t} \frac{ds}{|z|\left(\operatorname{Im}\overline{\sigma(z)}\right)^{1/(m+1)}}\right)^{m+1}$, then for f^{-1} the following estimate

$$\limsup_{w\to 0} \frac{|f^{-1}(w)|}{|w|^{m/\alpha}} \geqslant \left(C\, c_0^{-m}\right)^{\frac{1}{\alpha}}$$

holds, where $c_0 = (2\pi)^{-(m+1)/m} m^{-1/m}$.

(b) If for some $\varepsilon_0 \in (0, 1)$ and $\alpha > 0$ $\lim\limits_{\varepsilon\to 0} \varepsilon^{\alpha} \int\limits_{\varepsilon}^{\varepsilon_0} \frac{dt}{I_{m,\sigma}(t)} = \infty$, then

$$\limsup_{w\to 0} \frac{|f^{-1}(w)|}{|w|^{m/\alpha}} = \infty\,.$$

Corollary 3 *If for* $C > 0$

$$\int\limits_{\varepsilon}^{\varepsilon_0} \frac{dt}{I_{m,\sigma}(t)} \sim C\,\varepsilon^{-m} \quad as \quad \varepsilon \to 0\,,$$

then $\limsup\limits_{w\to 0} \frac{|f^{-1}(w)|}{|w|} \geqslant c_0^{-1} C^{1/m}$, *where* $c_0 = (2\pi)^{-(m+1)/m} m^{-1/m}$.

Theorem 4 *Let* $f\colon \mathbb{B} \to \mathbb{C}$ *be a regular homeomorphic solution of nonlinear equation (3) of Sobolev class* $W^{1,2}_{\mathrm{loc}}$ *satisfying* $f(0) = 0$ *and* $\operatorname{Im} \overline{\sigma(re^{i\theta})} \geqslant \lambda(r)$ *for a.a.* $r \in (0, \varepsilon_0)$, $\varepsilon_0 \in (0, 1)$, *where* $\lambda(r)\colon [0, 1) \to [0, \infty)$ *is a measurable function.*

(a) If for $C > 0$ *and* $\alpha > 0$

$$\int\limits_{\varepsilon}^{\varepsilon_0} \lambda(t)\,dt \sim C\,\varepsilon^{-\alpha} \quad as \quad \varepsilon \to 0\,,$$

then $\limsup\limits_{w\to 0} \frac{|f^{-1}(w)|}{|w|^{m/\alpha}} \geqslant (C\,m)^{1/\alpha}$.

(b) If for $C > 0$ *and* $\alpha > 0$ $\lim\limits_{\varepsilon\to 0} \varepsilon^{\alpha} \int\limits_{\varepsilon}^{\varepsilon_0} \lambda(t)\,dt = \infty$, *then* $\limsup\limits_{w\to 0} \frac{|f^{-1}(w)|}{|w|^{m/\alpha}} = \infty$.

Corollary 4 *If for* $C > 0$

$$\int\limits_{\varepsilon}^{\varepsilon_0} \lambda(t)\,dt \sim C\,\varepsilon^{-m} \quad as \quad \varepsilon \to 0\,,$$

then $\limsup\limits_{w\to 0} \frac{|f^{-1}(w)|}{|w|} \geqslant (C\,m)^{1/m}$.

References

1. Gutlyanskiĭ, V., Ryazanov, V., Srebro, U., Yakubov, E.: The Beltrami Equation. A Geometric Approach. Developments in Mathematics, vol. 26. Springer, New York (2012)
2. Martio, O., Ryazanov, V., Srebro, U., Yakubov, E.: Moduli in Modern Mapping Theory. Springer Monographs in Mathematics. Springer, New York (2009)
3. Gutlyanskiĭ, V., Ryazanov, V., Srebro, U., Yakubov, E.: On recent advances in the Beltrami equations. Ukr. Mat. Visn. **7**(4), 467–515 (2010). Reprinted in J. Math. Sci. (N.Y.) **175**(4), 413–449 (2011)
4. Srebro, U., Yakubov, E.: Beltrami equation. In: Handbook of Complex Analysis: Geometric Function Theory, vol. 2, pp. 555–597. Elsevier Sci. B. V., Amsterdam (2005)

5. Sevost'yanov, E.A.: On quasilinear Beltrami-type equations with degeneration (Russian). Mat. Zametki **90**(3), 445–453 (2011). Transl. Math. Notes **90**(3–4), 431–438 (2011)
6. Astala, K., Iwaniec, T., Martin, G.: Elliptic Partial Differential Equations and Quasiconformal Mappings in the Plane. Princeton Mathematical Series, vol. 48. Princeton University Press, Princeton (2009)
7. Guo, C.-Y., Kar, M.: Quantitative uniqueness estimates for p-Laplace type equations in the plane. Nonlinear Anal. **143**, 19–44 (2016)
8. Lavrent'ev, M.A., Šabat, B.V.: Geometrical properties of solutions of non-linear systems of partial differential equations (Russian). Dokl. Akad. Nauk SSSR (N.S.) **112**, 810–811 (1957)
9. Lavrent'ev, M.A.: A general problem of the theory of quasi-conformal representation of plane regions (Russian). Mat. Sbornik N.S. **21**(63), 285–320 (1947)
10. Lavrent'ev, M.A.: The variational method in boundary-value problems for systems of equations of elliptic type. Izdat. Akad. Nauk SSSR, Moscow (1962)
11. Šabat, B.V.: Geometric interpretation of the concept of ellipticity (Russian). Uspehi Mat. Nauk **12**(6(78)), 181–188 (1957)
12. Šabat, B.V.: On the notion of derivative system according to M. A. Lavrent'ev. Dokl. Akad. Nauk SSSR **136**, 1298–1301 (Russian). Translated as Soviet Math. Dokl. **2**, 202–205 (1961)
13. Kühnau, R.: Minimal surfaces and quasiconformal mappings in the mean. Trans. Inst. Math. Natl. Acad. Sci. Ukraine, **7**(2), 104–131 (2010)
14. Kruschkal, S.L., Kühnau, R.: Quasikonforme Abbildungenneue Methoden und Anwendungen (German). With English, French and Russian summaries. Teubner-Texte zur Mathematik [Teubner Texts in Mathematics], 54. BSB B. G. Teubner Verlagsgesellschaft, Leipzig (1983)
15. Adamowicz, T.: On p-harmonic mappings in the plane. Nonlinear Anal. **71**(1–2), 502–511 (2009)
16. Aronsson, G.: On certain p-harmonic functions in the plane. Manuscripta Math. **61**(1), 79–101 (1988)
17. Romanov, A.S.: Capacity relations in a planar quadrilateral (Russian). Sibirsk. Mat. Zh. **49**(4), 886–897 (2008). Translation in Sib. Math. J. **49**(4), 709–717 (2008)
18. Bojarski, B., Iwaniec, T.: p-Harmonic equation and quasiregular mappings. In: Partial Differential Equations (Warsaw, 1984), vol. 19, pp. 25–38. Banach Center Publ., PWN, Warsaw (1987)
19. Astala, K., Clop, A., Faraco, D., Jääskeläinen, J., Koski, A.: Nonlinear Beltrami operators, Schauder estimates and bounds for the Jacobian. Ann. Inst. H. Poincaré Anal. Non Linéaire **34**(6), 1543–1559 (2017)
20. Carozza, M., Giannetti, F., Passarelli di Napoli, A., Sbordone, C., Schiattarella, R.: Bi-Sobolev mappings and K_p-distortions in the plane. J. Math. Anal. Appl. **457**(2), 1232–1246 (2018)
21. Golberg, A., Salimov, R., Stefanchuk, M.: Asymptotic dilation of regular homeomorphisms. Complex Anal. Oper. Theory **13**(6), 2813–2827 (2019)
22. Salimov, R.R., Stefanchuk, M.V.: On the local properties of solutions of the nonlinear Beltrami equation. J. Math. Sci. **248**, 203–216 (2020)
23. Salimov, R.R., Stefanchuk, M.V.: Logarithmic asymptotics of the nonlinear Cauchy-Riemann-Beltrami equation. Ukr. Math. J. **73**, 463–478 (2021)
24. Golberg, A., Salimov, R.: Nonlinear Beltrami equation. Complex Var. Elliptic Equ. **65**(1), 6–21 (2020)
25. Lehto, O., Virtanen, K.I.: Quasiconformal mappings in the plane, 2nd edn. Translated from the German by K. W. Lucas. Die Grundlehren der mathematischen Wissenschaften, Band 126. Springer, New York (1973)
26. Bojarski, B., Gutlyanskiĭ, V., Martio, O., Ryazanov, V.: Infinitesimal geometry of quasiconformal and bi-Lipschitz mappings in the plane. In: EMS Tracts in Mathematics, vol. 19. European Mathematical Society (EMS), Zürich (2013)
27. Ikoma, K.: On the distortion and correspondence under quasiconformal mappings in space. Nagoya Math. J. **25**, 175–203 (1965)
28. Saks, S.: Theory of the Integral, 2nd revised edn. English translation by L.C. Young. With two additional notes by Stefan Banach Dover Publications, New York (1964)

On Conditions of Local Lineal Convexity Generalized to Commutative Algebras

Tetiana M. Osipchuk

Abstract The notion of lineally convex domains in the finite-dimensional complex space $\mathbb{C}^n$ and some of their properties are generalized to the finite-dimensional space $\mathcal{A}^n$, $n \geq 2$, that is the Cartesian product of n commutative and associative algebras $\mathcal{A}$. Namely, a domain in $\mathcal{A}^n$ is said to be ***(locally) $\mathcal{A}$-lineally convex*** if, for every boundary point of the domain, there exists a hyperplane in $\mathcal{A}^n$ passing through the point but not intersecting the domain (in some neighborhood of the point). It is proved that $\mathcal{A}_3$-lineal convexity of bounded domains with a smooth boundary in the space $\mathcal{A}_3^n$ follows from their local $\mathcal{A}_3$-lineal convexity for a three-dimensional algebra $\mathcal{A}_3$.

1 Introduction

The notion of lineal convexity that is studied in the theory of functions of many complex variables was coined in 1935 by Heinrich Behnke and Ernst F. Peschl [1], but it has been actively used only since the 60s due to the works of André Martineau [2, 3] and Lev A. Aizenberg [4, 5] who defined a lineally convex set in the finite-dimensional complex space $\mathbb{C}^n$, $n \geqslant 2$, independently in slightly different ways.

Consider a complex hyperplane

$$\Pi_{\mathbb{C}}(\boldsymbol{w}) := \left\{ (z_1, \ldots, z_n) \in \mathbb{C}^n : \sum_{j=1}^{n} \boldsymbol{c}_j \left(z_j - \boldsymbol{w}_j \right) = \mathbf{0}, (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n) \in \mathbb{C}^n \setminus \{\mathbf{0}\} \right\}$$

passing through a point $\boldsymbol{w} = (\boldsymbol{w}_1, \boldsymbol{w}_2 \ldots, \boldsymbol{w}_n) \in \mathbb{C}^n$.

Definition 1 (A. Martineau [2]) A set $E \subset \mathbb{C}^n$ is said to be **lineally convex in the sense of Martineau** if its complement is a union of complex hyperplanes.

T. M. Osipchuk (✉)
Institute of Mathematics of the National Academy of Sciences of Ukraine, Kyiv, Ukraine

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_23

The lineal convexity of a set $E \subset \mathbb{C}^n$ in the sense of Martineau is equivalent to the condition that, for any point $\boldsymbol{w} \in \mathbb{C}^n \setminus E$, there is a complex hyperplane $\Pi_{\mathbb{C}}(\boldsymbol{w})$ not intersecting E.

Definition 2 (L. Aizenberg [4]) A domain $D \subset \mathbb{C}^n$ is said to be **lineally convex** if, for every boundary point $\boldsymbol{w} \in \partial D$, there exists a complex hyperplane $\Pi_{\mathbb{C}}(\boldsymbol{w})$ not intersecting D.

A domain lineally convex in the sense of Martineau is obviously lineally convex by Aizenberg. In [6] it is proved that there exist domains lineally convex by Aizenberg and not lineally convex in the sense of Martineau. The notion of lineal convexity in the sense of the Aizenberg definition is also known as ***weak lineal convexity*** [7–9].

Definition 3 ([1, 10, 11]) A domain $D \subset \mathbb{C}^n$ is said to be **locally lineally convex** if, for every boundary point $\boldsymbol{w} \in \partial D$, there exists a complex hyperplane $\Pi_{\mathbb{C}}(\boldsymbol{w})$ passing through $\boldsymbol{w}$ but not intersecting D in some neighborhood of the point $\boldsymbol{w}$.

There is also another definition of local lineal convexity:

Definition 4 ([12]) An open set $D \subset \mathbb{C}^n$ is said to be **locally lineally convex in the sense of Kiselman** if, for every point $\boldsymbol{w} \in \mathbb{C}^n$, there exists a neighborhood U of $\boldsymbol{w}$ such that $D \cap U$ is lineally convex.

Local lineal convexity in the sense of Kiselman implies local lineal convexity for all open sets. But there exists a bounded domain in $\mathbb{C}^2$ with Lipschitz boundary which is locally lineally convex but not locally lineally convex in the sense of Kiselman (see Example 4.4 in [12]).

H. Behnke and E. Peschl in [1] proved that the global lineal convexity follows from the local one for bounded domains with a smooth boundary in $\mathbb{C}^2$. For the case of $\mathbb{C}^n$ this result was obtained in 1971 by Alexander P. Yuzhakov and Viachelsav P. Krivokolesko [10]. The statement fails for domains with a non-smooth boundary. An example of a locally lineally convex domain that is not lineally convex is constructed in [10]. Moreover, Yuri B. Zelinskii showed that the condition of the boundedness of the domain is essential by constructing an example of an unbounded locally lineally convex domain with a smooth boundary that is not lineally convex [13]. Later an example of an unbounded Hartogs domain having these properties was also constructed by Christer O. Kiselman [12].

In the work [1], the separate necessary and sufficient analytical conditions of local lineal convexity of domains with a smooth boundary in $\mathbb{C}^2$ were also obtained. In 1971 B. S. Zinoviev got a generalization of Behnke-Peschl conditions for the case $\mathbb{C}^n$, $n \geqslant 2$, in terms of nonnegativity and positivity of the differential of the second order of a real function defining a domain with a boundary of the class C^2, respectively. Moreover, the sign of the differential is determined on the boundary of the domain and on the vectors of a complex hyperplane tangent to the domain [11]. In 1998 Christer O. Kiselman managed to obtain the criterion of lineal convexity of a bounded domain in the space $\mathbb{C}^n$ with a boundary of the class C^2 in terms of nonnegativity of the differential of the second order of the function defining

the domain [9]. In 2008 Lars Hörmander improved Kiselman's result by loosening conditions imposed on the boundary of the domain [14].

In 1980s, the theory of lineally convex sets begins to be generalized to the spaces of hypercomplex numbers by Henzel A. Mkrtchyan and Yuri B. Zelinskii [15, 16]. In [15] it is proved that the global hypercomplex convexity follows from the local one for bounded domains with a smooth boundary in the multi-dimensional quaternion space. Conditions similar to those of Zinoviev were obtained for the algebra of real quaternions [17], the algebra of real generalized quaternions [18], and Clifford algebras [19].

Consider a commutative and associative algebra $\mathcal{A}$ over the field of real numbers $\mathbb{R}$ with identity $\boldsymbol{e}$. Let $\dim\mathcal{A} = m$ and elements $\{\boldsymbol{e}_k\}_{k=1}^m$ be a basis of $\mathcal{A}$. Consider the vector space $\mathcal{A}^n$, $n \geqslant 2$, which is the Cartesian product of n algebras $\mathcal{A}$. Let $z = (z_1, z_2, \ldots, z_n) \in \mathcal{A}^n$, where

$$z_j := x_1^j \boldsymbol{e}_1 + x_2^j \boldsymbol{e}_2 + \cdots + x_m^j \boldsymbol{e}_m \in \mathcal{A}, \quad x_q^j \in \mathbb{R}, \ \ j = \overline{1, n}.$$

And let a neighbourhood $U(\boldsymbol{w})$ of a point $\boldsymbol{w} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_n) \in \mathcal{A}^n$ be an open ball with center at $\boldsymbol{w}$. Consider a hyperplane

$$\Pi_{\mathcal{A}}(\boldsymbol{w}) := \left\{ z \in \mathcal{A}^n : \sum_{j=1}^{n} \boldsymbol{c}_j \left(z_j - \boldsymbol{w}_j\right) = \boldsymbol{0}, (\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_n) \in \mathcal{A}^n \setminus \{\boldsymbol{0}\} \right\} \tag{1}$$

which is called ***analytic***. An analytic hyperplane $\Pi_{\mathcal{A}}(\boldsymbol{w})$ is called ***(locally) supporting*** for a domain $\Omega \subset \mathcal{A}^n$ at a point $\boldsymbol{w} \in \partial\Omega$ if it does not intersect Ω (in some neighborhood of the point $\boldsymbol{w}$). A domain $\Omega \subset \mathcal{A}^n$ is said to be ***(locally) $\mathcal{A}$-lineally convex*** if it has an analytic, (locally) supporting hyperplane $\Pi_{\mathcal{A}}(\boldsymbol{w})$ at every point $\boldsymbol{w} \in \partial\Omega$.

Let $\gamma_{lk}^p \in \mathbb{R}$ be structure constants of $\mathcal{A}$ defined as follows:

$$\boldsymbol{e}_l \boldsymbol{e}_k = \sum_{p=1}^{m} \gamma_{lk}^p \boldsymbol{e}_p, \quad l, k = \overline{1, m}. \tag{2}$$

Moreover, let the basis satisfy the following conditions:

(1) there exist the inverse elements $\boldsymbol{e}_k^{-1} = \dfrac{1}{\boldsymbol{e}_k}$, $k = \overline{1, m}$;

(2) there exists $p = \tilde{p}$ such that the matrix $\Gamma^{\tilde{p}} = (\gamma_{lk}^{\tilde{p}})$ is non-degenerate.

Consider the following matrices

$$\boldsymbol{Z}_j = \begin{pmatrix} z_j^{\mathbf{1}} \\ z_j^{\mathbf{2}} \\ \dots \\ z_j^{\boldsymbol{m}} \end{pmatrix}, \ \boldsymbol{E} = \begin{pmatrix} \boldsymbol{e}_1 & 0 & \dots & 0 \\ 0 & \boldsymbol{e}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{e}_m \end{pmatrix}, \ X^j = \begin{pmatrix} x_1^j \\ x_2^j \\ \dots \\ x_m^j \end{pmatrix}, \ j = \overline{1,n},$$

and a non-degenerate $m \times m$ matrix

$$\Gamma = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1m} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{m1} & \gamma_{m2} & \cdots & \gamma_{mm} \end{pmatrix}, \ \text{where } \gamma_{lq} \in \mathbb{R}. \tag{3}$$

And let

$$\boldsymbol{Z}_j = \Gamma \boldsymbol{E} X^j, \tag{4}$$

i. e.,

$$z_j^{\boldsymbol{l}} := \gamma_{l1} x_1^j \boldsymbol{e}_1 + \gamma_{l2} x_2^j \boldsymbol{e}_2 + \dots + \gamma_{lm} x_m^j \boldsymbol{e}_m, \ l = \overline{1,m}, \ j = \overline{1,n}.$$

We obtain from (4):

$$X^j = \boldsymbol{E}^{-1} \Gamma^{-1} \boldsymbol{Z}_j,$$

where $\Gamma^{-1} = (\eta_{lp})$, $\eta_{lp} \in \mathbb{R}$, $l, p = \overline{1,m}$. That is to say,

$$x_l^j = \boldsymbol{e}_l^{-1} \sum_{p=1}^{m} \eta_{lp} z_j^{\boldsymbol{p}}, \ j = \overline{1,n}, \ l = \overline{1,m}. \tag{5}$$

Let $\rho(\boldsymbol{z}) = \rho(z) = \rho(x_1^1, x_2^1, \dots, x_m^n) : \mathbb{R}^{mn} \to \mathbb{R}$, $\boldsymbol{z} \in \mathcal{A}^n$, $z \in \mathbb{R}^{mn}$. Substituting x_l^j, $j = \overline{1,n}$, $l = \overline{1,m}$, for their values (5) in the expression of the function $\rho(z)$, we get

$$\rho(z) = \rho(x_1^1(z_1^{\mathbf{1}}, z_1^{\mathbf{2}} \dots, z_1^{\boldsymbol{m}}), x_2^1(z_1^{\mathbf{1}}, z_1^{\mathbf{2}} \dots, z_1^{\boldsymbol{m}}), \dots, x_m^n(z_n^{\mathbf{1}}, z_n^{\mathbf{2}} \dots, z_n^{\boldsymbol{m}})). \tag{6}$$

Now consider a domain

$$\Omega = \{\boldsymbol{z} \in \mathcal{A}^n : \rho(\boldsymbol{z}) = \rho(z^{\mathbf{1}}, z^{\mathbf{2}}, \dots, z^{\boldsymbol{m}}) < 0\}, \tag{7}$$

where $z^l = (z_1^l, z_1^l, z_2^l \dots z_n^l)$, $l = \overline{1,m}$, with the boundary $\partial\Omega = \{z \in \mathcal{A}^n : \rho(z) = 0\}$, where the function $\rho : \mathcal{A}^n \to \mathbb{R}$ is k times continuously differentiable in a neighborhood of $\partial\Omega$ with respect to its real variables and such that grad$\rho \neq 0$ everywhere on $\partial\Omega$. If $k = 1$, then we say that Ω has a ***smooth boundary***. If $k = 2$, then such a domain is called ***regular***.

In the work [20] the conditions similar to those of Zinoviev were obtained for regular, locally $\mathcal{A}$-lineally convex domains and are presented in Sect. 2.

In Sect. 3 the three-dimensional commutative algebra $\mathcal{A}_3$ with the basis $\{\boldsymbol{e}_1 = 1,\ \boldsymbol{e}_2 = \rho,\ \boldsymbol{e}_3 = \rho^2\}$ having the following multiplication table:

$\cdot$	1	$\boldsymbol{e}_2$	$\boldsymbol{e}_3$
1	1	$\boldsymbol{e}_2$	$\boldsymbol{e}_3$
$\boldsymbol{e}_2$	$\boldsymbol{e}_2$	$\boldsymbol{e}_3$	-1
$\boldsymbol{e}_3$	$\boldsymbol{e}_3$	-1	$-\boldsymbol{e}_2$

is considered. It is proved that the $\mathcal{A}_3$-lineal convexity of a bounded domain $\Omega \subset \mathcal{A}_3^n$ with a smooth boundary follows from its local $\mathcal{A}_3$-lineal convexity.

2 Analytical Conditions of Local $\mathcal{A}$-Lineal Convexity

We say that an analytic hyperplane $\Pi_{\mathcal{A}}(\boldsymbol{w})$ ***lies in a real hyperplane***

$$\Pi_{\mathbb{R}^{mn}}(w) := \left\{ \left(s_1^1, s_2^1, \dots, s_m^n\right) \in \mathbb{R}^{mn} : \sum_{j=1}^{n}\sum_{l=1}^{m} a_l^j s_l^j = 0, \right.$$
$$\left. (a_1^1, a_2^1, \dots, a_m^n) \in \mathbb{R}^{mn} \setminus \{0\} \right\}, \tag{8}$$

if any vector $\boldsymbol{s} = (\boldsymbol{s}_1, \boldsymbol{s}_2, \dots, \boldsymbol{s}_n) \in \mathcal{A}^n$, $\boldsymbol{s}_j = \sum\limits_{l=1}^{m} s_l^j \boldsymbol{e}_l = \sum\limits_{l=1}^{m} (x_l^j - w_l^j)\boldsymbol{e}_l = \boldsymbol{z}_j - \boldsymbol{w}_j$, $j = \overline{1,n}$, satisfying the equation of the hyperplane (1) satisfies the equation of the hyperplane (8).

Lemma 1 *Let a real hyperplane $\Pi_{\mathbb{R}^{mn}}(w)$ (8) be given and let $\Pi_{\mathcal{A}}^{\tilde{p}}(\boldsymbol{w})$ be the analytic hyperplane (1) such that*

$$\boldsymbol{c}_j = \sum_{k,l=1}^{m} \eta_{kl}^{\tilde{p}} a_l^j \boldsymbol{e}_k, \quad j = \overline{1,n}, \tag{9}$$

where $\eta_{kl}^{\tilde{p}}$ are the elements of the matrix inverse to the matrix $\Gamma^{\tilde{p}}$ satisfying condition 2). Then $\Pi_{\mathcal{A}}^{\tilde{p}}(\boldsymbol{w})$ lies in $\Pi_{\mathbb{R}^{mn}}(w)$.

Proof Substitute the constants $\boldsymbol{c}_j$ in (1) for their values (9) and, after multiplying by $\boldsymbol{s}_j = \sum\limits_{p=1}^{m} s_p^j \boldsymbol{e}_p$, group together the terms with each basis element $\boldsymbol{e}_k$, $k = \overline{1,m}$, separately. Set the grouped expressions to zero. We obtain that the equation in (1) is equivalent to the system of m real equations defining real hyperplanes in the mn-dimensional real space. Moreover, the equation obtained after grouping terms with the unit $\boldsymbol{e}_{\tilde{p}}$ defines the real hyperplane $\Pi_{\mathbb{R}^{mn}}(w)$. The lemma is proved. □

By Lemma 1, for any $\tilde{p}$ satisfying condition 2), the analytic hyperplane

$$T_{\mathcal{A}}^{\tilde{p}}(\boldsymbol{w}) := \left\{ \boldsymbol{s} = (\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n) \in \mathcal{A}^n : \sum_{j=1}^{n} \sum_{k,l=1}^{m} \eta_{kl}^{\tilde{p}} \frac{\partial \rho(w)}{\partial x_l^j} \boldsymbol{e}_k \boldsymbol{s}_j = \boldsymbol{0} \right\}$$

lies in the real hyperplane tangent to a domain $\Omega \subset \mathcal{A}^n$ (7) with a smooth boundary at $\boldsymbol{w} \in \partial\Omega$.

If the function ρ (6) is twice continuously differentiable with respect to its real variables x_l^j, $j = \overline{1,n}$, $l = \overline{1,m}$, at a point $\boldsymbol{w} \in \mathcal{A}^n$, then, formally differentiating ρ as a composite function with respect to the variables $\boldsymbol{z}_j^{\boldsymbol{l}}$, $j = \overline{1,n}$, $l = \overline{1,m}$, we obtain the following formulas for the formal partial derivatives:

$$\frac{\partial \rho(\boldsymbol{w})}{\partial \boldsymbol{z}_j^{\boldsymbol{p}}} := \sum_{l=1}^{m} \eta_{lp} \frac{\partial \rho(w)}{\partial x_l^j} \boldsymbol{e}_l^{-1}, \quad j = \overline{1,n}, \quad p = \overline{1,m},$$

$$\frac{\partial^2 \rho(\boldsymbol{w})}{\partial \boldsymbol{z}_j^{\boldsymbol{p}} \partial \boldsymbol{z}_i^{\boldsymbol{q}}} := \sum_{l,k=1}^{m} \eta_{lp} \eta_{kq} \frac{\partial^2 \rho(w)}{\partial x_l^j \partial x_k^i} \boldsymbol{e}_l^{-1} \boldsymbol{e}_k^{-1}, \quad j,i = \overline{1,n}, \quad p,q = \overline{1,m}.$$

Theorem 1 *If a regular domain $\Omega \subset \mathcal{A}^n$ is locally $\mathcal{A}$-lineally convex and $T_{\mathcal{A}}^{\tilde{p}}(\boldsymbol{w})$ is locally supporting for Ω at any point $\boldsymbol{w} \in \partial\Omega$, then, for $\boldsymbol{w}$ and any vector $\boldsymbol{s} \in T_{\mathcal{A}}^{\tilde{p}}(\boldsymbol{w})$, $\|\boldsymbol{s}\| = 1$, the following inequality is true*

$$\sum_{i,j=1}^{n} \sum_{k,l=1}^{m} \frac{\partial^2 \rho(\boldsymbol{w})}{\partial \boldsymbol{z}_i^{\boldsymbol{l}} \partial \boldsymbol{z}_j^{\boldsymbol{k}}} \boldsymbol{s}_j^{\boldsymbol{l}} \boldsymbol{s}_i^{\boldsymbol{k}} \geqslant 0. \tag{10}$$

If, for any point $\boldsymbol{w} \in \partial\Omega$ and any vector $\boldsymbol{s} \in T_{\mathcal{A}}^{\tilde{p}}(\boldsymbol{w})$, $\|\boldsymbol{s}\| = 1$,

$$\sum_{i,j=1}^{n} \sum_{k,l=1}^{m} \frac{\partial^2 \rho(\boldsymbol{w})}{\partial \boldsymbol{z}_i^{\boldsymbol{l}} \partial \boldsymbol{z}_j^{\boldsymbol{k}}} \boldsymbol{s}_j^{\boldsymbol{l}} \boldsymbol{s}_i^{\boldsymbol{k}} > 0, \tag{11}$$

then the regular domain $\Omega \subset \mathcal{A}^n$ is locally $\mathcal{A}$-lineally convex.

Proof *Sufficiency.* Write the Taylor series formally for the function $\rho(z) = \rho(z^{\mathbf{1}}, z^{\mathbf{2}}, \ldots, z^{\boldsymbol{m}})$, $z^{\boldsymbol{l}} = (z_1^{\boldsymbol{l}}, z_1^{\boldsymbol{l}}, z_2^{\boldsymbol{l}} \ldots z_n^{\boldsymbol{l}})$, $l = \overline{1, m}$, with respect to the variables $z_j^{\boldsymbol{l}}$ in the neighborhood $U(\boldsymbol{w})$ of any point $\boldsymbol{w} \in \partial\Omega$. Notice that $\rho(\boldsymbol{w}) = 0$ at any boundary point $\boldsymbol{w}$. Since $\boldsymbol{s} \in T_{\mathcal{A}}^{\tilde{p}}(\boldsymbol{w})$, the second summand in the Taylor decomposition also vanishes. Then

$$\rho(z) = \frac{1}{2}\left(\sum_{i,j=1}^{n}\sum_{k,l=1}^{m}\frac{\partial^2\rho(\boldsymbol{w})}{\partial z_i^{\boldsymbol{l}}\partial z_j^{\boldsymbol{k}}}\frac{(z_j^{\boldsymbol{l}} - \boldsymbol{w}_j^{\boldsymbol{l}})(z_i^{\boldsymbol{k}} - \boldsymbol{w}_i^{\boldsymbol{k}})}{\|z - \boldsymbol{w}\|^2}\right)\|z - \boldsymbol{w}\|^2 + o(\|z - \boldsymbol{w}\|^2),\ z \to \boldsymbol{w}, \tag{12}$$

for any point $z \in U(\boldsymbol{w}) \cap T_{\mathcal{A}}^{\tilde{p}}(\boldsymbol{w})$.

Thus, $\rho(z) \geqslant 0$ for any point $z \in U(\boldsymbol{w}) \cap T_{\mathcal{A}}^{\tilde{p}}(\boldsymbol{w})$ and any point $\boldsymbol{w} \in \partial\Omega$ by (11) and (12), which means local $\mathcal{A}$-lineal convexity of the domain Ω.

Necessity. Suppose a regular domain Ω is locally $\mathcal{A}$-lineally convex and, for a point $\widetilde{\boldsymbol{w}} = (\widetilde{\boldsymbol{w}}_1, \widetilde{\boldsymbol{w}}_2, \ldots, \widetilde{\boldsymbol{w}}_n) \in \partial\Omega$ and for a vector $\boldsymbol{t} = (\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n) \in T_{\mathcal{A}}^{\tilde{p}}(\widetilde{\boldsymbol{w}})$, the following inequality is true

$$\sum_{i,j=1}^{n}\sum_{k,l=1}^{m}\frac{\partial^2\rho(\widetilde{\boldsymbol{w}})}{\partial z_i^{\boldsymbol{l}}\partial z_j^{\boldsymbol{k}}}\boldsymbol{t}_j^{\boldsymbol{l}}\boldsymbol{t}_i^{\boldsymbol{k}} < 0. \tag{13}$$

On the other hand, for the points $z \in U(\widetilde{\boldsymbol{w}}) \cap T_{\mathcal{A}}^{\tilde{p}}(\widetilde{\boldsymbol{w}})$, the expansion (12) is valid. Thus, for the points $\widetilde{z} = (\widetilde{z}_1, \widetilde{z}_2, \ldots, \widetilde{z}_n) \in U(\widetilde{\boldsymbol{w}}) \cap T_{\mathcal{A}}^{\tilde{p}}(\widetilde{\boldsymbol{w}})$ which correspond to the tangent vector $\boldsymbol{t}$, where correspondence is defined by the relation $\boldsymbol{t}_i = (\widetilde{z}_i - \widetilde{\boldsymbol{w}}_i)/\|\widetilde{z} - \widetilde{\boldsymbol{w}}\|$, $i = \overline{1, n}$, the inequality $\rho(\widetilde{z}) < 0$ is true by (13), which contradicts the fact that the hyperplane $T_{\mathcal{A}}^{\tilde{p}}(\widetilde{\boldsymbol{w}})$ is locally supporting for Ω at $\widetilde{\boldsymbol{w}}$. □

3 Properties of Locally $\mathcal{A}_3$-Lineally Convex Domains with a Smooth Boundary

Lemma 2 *Let $\Omega = \{x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n : \rho(x) < 0\}$ be a bounded domain in $\mathbb{R}^n$ with a smooth boundary. Suppose that $x_0 \in \partial\Omega$ and L is an r-dimensional plane, $1 < r < n$, that passes through the point x_0 and is not tangent to Ω. Then there exists a neighbourhood $U(x_0)$ such that $\partial\Omega \cap U(x_0) \cap L$ is the graph of some smooth function.*

The statement of the lemma follows directly from the implicit-function theorem.

Using Lemma 1, it can be proved that, for any real hyperplane $\Pi_{\mathbb{R}^{3n}}(w)$, $w \in \mathbb{R}^{3n}$, there is the unique analytic hyperplane $\Pi_{\mathcal{A}_3}(\boldsymbol{w})$ lying in $\Pi_{\mathbb{R}^{3n}}(w)$.

For any $\boldsymbol{a}, \boldsymbol{b} \in \mathcal{A}_3^n$, the set

$$L := \left\{z = (z_1, z_2, \ldots, z_n) \in \mathcal{A}_3^n : z = \boldsymbol{t}\,(\boldsymbol{b} - \boldsymbol{a}) + \boldsymbol{a}, \quad \boldsymbol{t} \in \mathcal{A}_3\right\}$$

is called the ***analytic line*** passing through the points $\boldsymbol{a}, \boldsymbol{b}$.

Lemma 3 *An analytic line L that is contained in a real hyperplane tangent to a locally $\mathcal{A}_3$-lineally convex domain $\Omega \subset \mathcal{A}_3^n$ with a smooth boundary at a point $\boldsymbol{w} \in \partial\Omega$, and passing through the point $\boldsymbol{w}$ does not intersect Ω in some neighborhood of $\boldsymbol{w}$.*

Proof Since Ω is locally $\mathcal{A}_3$-lineally convex and has a smooth boundary, the unique analytic hyperplane $T_{\mathcal{A}_3}(\boldsymbol{w})$ locally supporting for Ω at $\boldsymbol{w}$ lies in the real hyperplane tangent to Ω at $\boldsymbol{w}$. Since the analytic hyperplane and analytic line are of the real dimensions $3n-3$ and 3, respectively, and they both are contained in the same $3n-1$ dimensional real hyperplane, they have the common 2-dimensional real plane. For all that, the analytic hyperplane and analytic line having a common point intersect only at this point or the line is completely contained in the hyperplane. Thus, in our case, $L \subset T_{\mathcal{A}_3}(\boldsymbol{w})$. Since $T_{\mathcal{A}_3}(\boldsymbol{w}) \cap \Omega = \varnothing$ in some neighborhood of the point $\boldsymbol{w}$, therefore $L \cap \Omega = \varnothing$ in the same neighborhood of $\boldsymbol{w}$. □

Lemma 4 *An arbitrary analytic line L intersects a bounded locally $\mathcal{A}_3$-lineally convex domain with a smooth boundary $\Omega \subset \mathcal{A}_3^n$ in at most one connected component.*

Proof Suppose $z^0, z^1 \in L \cap \Omega$ be two arbitrary points of the intersection $L \cap \Omega$. Note that upper indices which are not in bold are the part of the notation of the points of the space $\mathcal{A}_3^n$ and do not mean conjugation, as is customary above for upper indices in bold. Prove that z^0, z^1 belong to the same connected component of $L \cap \Omega$. Since Ω is connected, there exists a curve $\gamma : [0, 1] \to \Omega$, such that $\gamma(0) = z^0$, $\gamma(1) = z^1$. Let $\gamma(s) = z^s$, $s \in [0, 1]$. Without loss of generality, suppose $z^0 \equiv \mathbf{0}$. Consider the analytic lines

$$L_s = \left\{z \in \mathcal{A}_3^n : z = \boldsymbol{t}z^s, \quad \boldsymbol{t} \in \mathcal{A}_3\right\}, \quad s \in (0, 1),$$

passing through the points z^0, z^s. Consider the following sets:

- $\Sigma_1 := \{s$: points $\gamma(0), \gamma(s)$ are in the same component of $L_s \cap \Omega\}$;
- $\Sigma_2 := \{s$: points $\gamma(0), \gamma(s)$ are in the different components of $L_s \cap \overline{\Omega}\}$;
- $\Sigma_3 := \{s$: points $\gamma(0), \gamma(s)$ are in the same component of $L_s \cap \overline{\Omega}$ and in the different components of $L_s \cap \Omega\}$.

It is easy to see that $\Sigma_1 \cup \Sigma_2 \cup \Sigma_3 = (0, 1)$ and $\Sigma_1 \cap \Sigma_2 = \Sigma_1 \cap \Sigma_3 = \Sigma_2 \cap \Sigma_3 = \varnothing$.

1. First, show that $\Sigma_3 = \varnothing$. Without loss of generality, suppose that $L_s \cap \overline{\Omega}$, $s \in \Sigma_3$, consists of one connected component. Assume that there exists a point $\boldsymbol{x}^1$ such that $\boldsymbol{x}^1 \in L_s \cap \overline{\Omega}$, $\boldsymbol{x}^1 \notin \overline{L_s \cap \Omega}$. Then Lemma 2 is not fulfilled for the line L_s at the point $\boldsymbol{x}^1$. Thus, L_s is tangent to Ω at $\boldsymbol{x}^1$. Consider the set C of all points

$\boldsymbol{x} \in L_s \cap \overline{\Omega}$, such that L_s is tangent to Ω at the points $\boldsymbol{x}$. Since $\rho \in C^1$, the set C is closed, nonempty by assumption, and such that $C \cup \overline{L_s \cap \Omega} = L_s \cap \overline{\Omega}$. Since the set $L_s \cap \overline{\Omega}$ is connected, we have $C \cap \overline{L_s \cap \Omega} \neq \varnothing$. Therefore, there exists a point $\boldsymbol{x}^2 \in \overline{L_s \cap \Omega}$ at which L_s is tangent to Ω and an arbitrary neighborhood of the point $\boldsymbol{x}^2$ contains points of Ω, which contradicts Lemma 3. Thus, the assumption is wrong, and the point $\boldsymbol{x}^1$ with indicated properties does not exist. By virtue of the connectedness of $L_s \cap \overline{\Omega}$, there exists a point $\boldsymbol{x}^3$ that is a point of the boundary common for two different components of $L_s \cap \Omega$. Hence, according to Lemma 2, L_s is tangent to Ω at $\boldsymbol{x}^3$, which contradicts Lemma 3. Thus, the set Σ_3 is empty.

2. Show that the set Σ_2 is open. Namely, prove that, for any $s_0 \in \Sigma_2$, there exists $\delta > 0$ such that for any $s \in (s_0 - \delta, s_0 + \delta)$ the intersection $L_s \cap \overline{\Omega}$ is disconnected. The intersection $L_s \cap \Omega$ is isometric to the open set $\Omega_s = \{\boldsymbol{q} : f(\boldsymbol{q}, s) = \rho(p(s)\boldsymbol{q}) < 0\} \subset \mathcal{A}_3^1$, where $p(s) = \frac{\gamma(s)}{|\gamma(s)|}$, $\boldsymbol{q} \in \mathcal{A}_3$, $0 < s < 1$, since $\boldsymbol{z} = p(s)\boldsymbol{q}$ maps $\mathcal{A}_3^1$ homeomorphically onto $L_s \subset \mathcal{A}_3^n$ and $|p(s)\boldsymbol{q} - p(s)\boldsymbol{q}'| = |p(s)|\,|\boldsymbol{q} - \boldsymbol{q}'| = |\boldsymbol{q} - \boldsymbol{q}'|$. The points $\gamma(s), \gamma(0) \in L_s \cap \Omega$ are associated with the points $q(s) = |\gamma(s)|, 0 \in \Omega_s$. Choose an arbitrary $s_0 \in (0, 1)$ such that the intersection $L_{s_0} \cap \overline{\Omega}$ is disconnected. Consider components D_0 and $D_1 = (L_{s_0} \cap \overline{\Omega}) \setminus D_0$ of this intersection, and D_0', $D_1' \subset \overline{\Omega_{s_0}}$, respectively. The domain Ω is bounded. Hence, there exists a ball V_r such that $\overline{\Omega_{s_0}} \subset V_r$. Then $\mathcal{A}_3^1$ contains open sets V, U such that $D_0' \subset V$, $D_1' \subset U$, $V \cap U = \varnothing$, $U \cup V \subset V_r$. We have $f(\boldsymbol{q}, s_0) > 0$ on the compact $\overline{V}_r \setminus (U \cup V)$, since $\overline{\Omega_{s_0}} = \{\boldsymbol{q} : f(\boldsymbol{q}, s_0) \leq 0\} \subset U \cup V$. The function $f(\boldsymbol{q}, s)$ is continuous for $\boldsymbol{q} \in \mathcal{A}_3^1$, $0 < s < 1$. Consequently, there exists $\delta > 0$ such that, for all s such that $|s - s_0| < \delta$ and $\boldsymbol{q} \in \overline{V}_r \setminus (U \cup V)$ one has $f(\boldsymbol{q}, s) > 0$, i.e., $\overline{\Omega_s} \subset U \cup V$. Moreover, $0 \in D_0' \subset V$ and $q(s) \in D_1' \subset U$ for sufficiently small δ. Thus, $\gamma(0)$ and $\gamma(s)$ belong to different components of $L_s \cap \overline{\Omega}$ for $s_0 - \delta < s \leq s_0 + \delta$.

3. Show that the set Σ_1 is also open. Taking into account that $\Sigma_1 \neq \varnothing$, choose $s_0 \in (0, 1)$ so that the intersection $L_{s_0} \cap \Omega$ is connected. Prove that there exists $\varepsilon > 0$ such that, for any $s \in (s_0 - \varepsilon, s_0 + \varepsilon)$, the intersection $L_s \cap \Omega$ is also connected.

   Since the set $L_{s_0} \cap \Omega$ is connected and open in L_{s_0}, there exists a curve $\tau(l) \subset L_{s_0} \cap \Omega$, $l \in [0, 1]$, $\tau(0) = \gamma(0)$, $\tau(1) = \gamma(s_0)$. The distance r between the compact sets τ and $\partial\Omega$ is greater than zero. Consider the balls $B(\tau(l), r)$ of radius r and centered at the points $\tau(l)$, and their union $B = \bigcup\limits_{l \in [0,1]} B(\tau(l), r) \subset \Omega$. Choose $\varepsilon > 0$ so that, for any $s \in (s_0 - \varepsilon, s_0 + \varepsilon)$, the distance from $\gamma(s_0)$ to $\gamma(s)$ is less than r. Then L_s intersects every ball $B(\tau(l), r)$. Construct a continuous curve $\tau_s \subset L_s \cap \Omega$ connecting the points $\tau_s(0) = \gamma(0)$ and $\tau_s(1) = \gamma(s)$ as the union of the set of centers of the disks $C(l) := L_s \cap B(\tau(l), r)$, $l \in [0, 1]$, and the segment $[C(1), \gamma(s)]$ that lies inside the ball $L_s \cap B(\tau(1), r)$. It is obvious that the curve τ_s is completely contained in $L_s \cap \Omega$. Thus, the points $\gamma(0)$ and $\gamma(s)$ lie in the same component of the intersection $L_s \cap \Omega$.

Since $\Sigma_1 \cup \Sigma_2 \cup \Sigma_3 = (0, 1)$, $\Sigma_1 \cap \Sigma_2 = \Sigma_1 \cap \Sigma_3 = \Sigma_2 \cap \Sigma_3 = \varnothing$, $\Sigma_3 = \varnothing$, and both sets Σ_1, Σ_2 are open, we have either $\Sigma_1 = \varnothing$ or $\Sigma_2 = \varnothing$. Since the domain Ω

is connected, the set Σ_1 cannot be empty. Thus, $\Sigma_2 = \varnothing$, moreover, $\Sigma_1 = (0, 1)$. Then $L \cap \Omega$ is connected, otherwise, we will have the case 1 for $L \cap \Omega$, which contradicts Lemmas 2, 3.

The lemma is proved. □

Lemma 5 *Suppose that an analytic line L intersects a bounded locally $\mathcal{A}_3$-lineally convex domain $\Omega \subset \mathcal{A}_3^n$ with a smooth boundary. Then $\overline{L \cap \Omega} = L \cap \overline{\Omega}$.*

Proof By Lemma 4, $L \cap \Omega$ is connected. We have $\overline{L \cap \Omega} \subset L \cap \overline{\Omega}$. Let $L \cap \overline{\Omega} \setminus \overline{L \cap \Omega} \neq \varnothing$. The case, where $L \cap \overline{\Omega} \setminus \overline{L \cap \Omega}$ is not closed, is not possible (see case 1, Lemma 4). Thus, $L \cap \overline{\Omega} \setminus \overline{L \cap \Omega}$ is closed, i.e., $L \cap \overline{\Omega}$ is disconnected. Consider points z^0, z^1 such that $z^0 \in L \cap \Omega$, $z^1 \in L \cap \overline{\Omega} \setminus \overline{L \cap \Omega} \subset \partial\Omega$. Since $\partial\Omega$ is smooth, there exists a curve $\gamma : [0, 1] \to \overline{\Omega}$ such that $\gamma(1) = z^1$, $\gamma(s) = z^s \in \Omega$, $0 \leq s < 1$. Consider the analytic lines L_s passing through the points z^0, z^s, $0 \leq s \leq 1$. Since z^0, z^1 belong to different components of $L \cap \overline{\Omega}$, then z^0, z^s belong to different components of $L_s \cap \overline{\Omega}$, therefore, to different components of $L_s \cap \Omega$, for s close enough to 1 (see proof of the case 2, Lemma 4). This contradicts Lemma 4. Thus, $\overline{L \cap \Omega} = L \cap \overline{\Omega}$. □

Theorem 2 *If a bounded domain $\Omega \subset \mathcal{A}_3^n$ with a smooth boundary is locally $\mathcal{A}_3$-lineally convex, then it is $\mathcal{A}_3$-lineally convex.*

Proof Suppose Ω is not $\mathcal{A}_3$-lineally convex. Then there is a point $\boldsymbol{w} \in \partial\Omega$ and the analytic hyperplane $\Pi_{\mathcal{A}_3}(\boldsymbol{w})$ passing through $\boldsymbol{w}$, not intersecting Ω in some neighborhood $U(\boldsymbol{w})$ of $\boldsymbol{w}$, and such that $\Pi_{\mathcal{A}_3}(\boldsymbol{w}) \cap \Omega \neq \varnothing$. Take a point $z^0 \in \Pi_{\mathcal{A}_3}(\boldsymbol{w}) \cap \Omega$ and draw the analytic line $L \subset \Pi_{\mathcal{A}_3}(\boldsymbol{w})$ through the points $\boldsymbol{w}$, z^0. Then $z^0 \in L \cap \Omega$ and $\boldsymbol{w} \in L \cap \overline{\Omega}$. For all that, $\boldsymbol{w} \notin \overline{L \cap \Omega}$, since $U(\boldsymbol{w}) \cap \Omega \cap L \subset U(\boldsymbol{w}) \cap \Omega \cap \Pi_{\mathcal{A}_3}(\boldsymbol{w}) = \varnothing$. This contradicts Lemma 5. □

Remark 1 By Theorem 2, the analytical conditions (10), (11) of local $\mathcal{A}_3$-lineal convexity of a bounded regular domain in the space $\mathcal{A}_3^n$ are also the analytical conditions of $\mathcal{A}_3$-lineal convexity of the domain.

References

1. Behnke, H., Peschl, E.: Zur Theorie der Funktionen mehrerer komplexer Veränderlichen Konvexität in bezug auf analytische Ebenen im kleinen und großen. Math. Ann. **111**(2), 158–177 (1935; in German). https://doi.org/10.1007/BF01472211
2. Martineau, A.: Sur la topologie des espaces de fonctions holomorphes Math. Ann. **163**(1), 62–88 (1966; in French). https://doi.org/10.1007/BF02052485
3. Martineau, A.: Sur les équations aux dérivées partielles à coefficients constants avec second membre, dans le champ complexe. C. R. Acad. Sci. Paris Sér **163**(1), 62–88 (1967; in French)
4. Aizenberg, L.A.: Linear convexity in $\mathbb{C}^n$ and distribution of the singularities of holomorphic functions. Bull. Pol. Acad. Sci. Ser. Math. Astr. Phys. Sci. **15**(7), 487–495 (1967; in Russian)
5. Aizenberg, L.A.: Decomposition of holomorphic functions of several complex variables into partial fractions. Sib. Math. J. **8**, 859–872 (1967). https://doi.org/10.1007/BF01040660

6. Aizenberg, L.A., Yuzhakov, A.P., Makarova, Y.L.: On linear convexity in $\mathbb{C}^n$. Sibirsk. Mat. Zh, **9**(4), 731–746 (1968; in Russian). http://mi.mathnet.ru/smj5559
7. Hörmander, L.: Notions of Convexity. Progress in Mathematics, vol. 127. Birkhäuser, Boston (1994)
8. Kiselman, Ch.O.: Lineally convex Hartogs domains. Acta Math. Vietnamica **21**, 69–94 (1996)
9. Kiselman, Ch.O.: A differential inequality characterizing weak lineal convexity. Math. Ann. **311**(1), 1–10 (1998)
10. Yuzhakov, A.P., Krivokolesko, V.P.: Some properties of linearly convex domains with smooth boundaries in $\mathbb{C}^n$. Sib. Math. J. **12**, 323–327 (1971). https://doi.org/10.1007/BF00969055
11. Zinoviev, B.S.: Analytic conditions and some questions of approximation of linear convex domains with smooth boundaries in the space $\mathbb{C}^n$. Izv. vuzov, Math. **6**, 61–69 (1971; in Russian). http://mi.mathnet.ru/ivm3882
12. Kiselman, Ch.O.: Weak lineal convexity. In: Constructive Approximation of Functions, vol. 107, pp. 159–174. Banach Center Publications, Polish Academy of Sciences (2016). https://doi.org/10.4064/bc107-0-11
13. Zelinskii, Yu.B.: On locally linearly convex domains. Ukr Mat. J. **54**(2), 280–284 (2002; in Russian)
14. Hörmander, L.: Weak linear convexity and a related notion of concavity. Math. Scand. **102**, 73–100 (2008). https://doi.org/10.7146/math.scand.a-15052
15. Mkrtchyan, H.A.: Hypercomplex-convex domains with smooth boundary. Doc. AN USSR Ser. A **3**, 15–17 (1986; in Russian)
16. Zelinskii, Yu.B., Mkrtchyan, H.A.: On extremal points and hypercomplex-convex domains. Doc. AN USSR **311**(6), 1299–1302 (1990; in Russian)
17. Osipchuk, T.M.: Analytic conditions of local linear convexity in $\mathbb{H}^n$. Zb. Prats Inst. Math. NASU **3**(3), 244–254 (2006; in Ukrainian)
18. Osipchuk, T.M.: Analytical conditions of local linear convexity in the space $\mathbb{H}^n_{\alpha,\beta}$. Zb. Prats Inst. Math. NASU **10**(4–5), 301–305 (2013)
19. Osipchuk, T.M., Zelinskii, Yu.B., Tkachuk, M.V.: Analytical conditions of locally general convexity in $C^n_{p,q}$. Zb. Prats Inst. Math. NASU **7**(2), 393–401 (2010; in Ukrainian)
20. Osipchuk, T.M.: On local linear convexity generalized to commutative algebras. ArXiv preprint. arXiv:2009.13306 [math.CV] (2020). https://doi.org/10.48550/arXiv.2009.13306

On a Quadrature Formula for the Direct Value of the Double Layer Potential

Igor O. Reznichenko, Pavel A. Krutitskii, and Valentina V. Kolybasova

Abstract A quadrature formula for the direct value of the double layer potential with continuous density given on a closed or open surface is derived. The double layer potential for the Helmholtz equations are considered, the potential for the Laplace equation is a particular case. The proposed quadrature formula gives significantly higher accuracy than standard quadrature formula, which is confirmed by numerical tests. The derived quadrature formula can be used for numerical solving boundary value problems for the Laplace and Helmholtz equations by the method of potentials and boundary integral equations.

1 Introduction

The double layer potential is used in the numerical solution of boundary value problems for the Laplace and Helmholtz equations by the method of integral equations in [1, 2]. With the help of potentials, boundary value problems are reduced to integral equations. For the numerical solution of integral equations, it is necessary to have quadrature formulas that calculate with good accuracy the direct values of potentials on the surface, where the potential density is given. Engineering calculations use standard quadrature formulas for potentials [3], but their accuracy leaves much to be desired the best. An improved quadrature formula for the direct value of the potential of a simple layer is proposed in [4], and for the direct value of the normal derivative of the potential of a simple layerin [5]. In this paper, an improved quadrature formula is derived for the directvalue of the double layer potential. The improved formula gives significantly higher accuracy than the standard one, which is confirmed by numerical tests.

I. O. Reznichenko (✉) · V. V. Kolybasova
Lomonosov Moscow State University, Moscow, Russia
e-mail: liorb@mail.ru; kolybasova@physics.msu.ru

P. A. Krutitskii
Keldysh Institute of Applied Mathematics, Moscow, Russia
e-mail: biem@mail.ru

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_24

In the two-dimensional case, an improved quadrature formula for the simple-layer potential with density given on open curves and having power-law singularities at the ends of the curves is constructed in [6, 7]. This formula can be used to find numerical solutions to boundary value problems for the Laplace and Helmholtz equations outside of sections and open curves on the plane. Such problems were studied in [8–14].

2 Problem Statement

Here we define the properties of the given surface, the definition of the double layer potential for the Helmholtz equation and introduce the main objective of the paper.

We use the Cartesian coordinate system $x = (x_1, x_2, x_3) \in R^3$ in three-dimensional space. Let Γ be either a simple closed C^2-smooth surface or a simple bounded open oriented C^2-smooth surface containing the limit points of itself [15, Sec. 14.1]. If the surface Γ is closed, then it must bound a spatially simply connected interior domain [16, p. 201]. Assume that Γ is parameterized in such a way that a rectangle is mapped onto it,

$$y = (y_1, y_2, y_3) \in \Gamma, \quad y_1 = y_1(u, v), \ y_2 = y_2(u, v), \ y_3 = y_3(u, v);$$

$$u \in [0, A], \quad v \in [0, B]; \quad y_j(u, v) \in C^2([0, A] \times [0, B]), \quad j = 1, 2, 3. \tag{1}$$

The sphere, the surface of an ellipsoid, smooth surfaces of figures of revolution, the torus surface, and many other more complicated surfaces can be parameterized in this way. Let us introduce N points u_n with step h on the interval $[0; A]$ and M points v_m with step H on the interval $[0; B]$ by the formulas

$$A = Nh, \quad B = MH, \quad u_n = (n + 1/2)h, \quad n = 0, \ldots, N - 1;$$

$$v_m = (m + 1/2)H, \quad m = 0, \ldots, M - 1.$$

We divide the rectangle $[0, A] \times [0, B]$ mapped onto the surface Γ into $N \times M$ small rectangles of size $h \times H$; then the points $(u_n; v_m)$ are the midpoints of these rectangles.

It is well known [15, Sec. 14.1] that the components of a (not necessarily unit) normal vector $\eta(y) = (\eta_1(y), \eta_2(y), \eta_3(y))$ at a point $y = (y_1, y_2, y_3) \in \Gamma$ of the surface can be expressed via second-order determinants by the formulas

$$\eta_1 = \begin{vmatrix} (y_2)_u & (y_3)_u \\ (y_2)_v & (y_3)_v \end{vmatrix}, \quad \eta_2 = \begin{vmatrix} (y_3)_u & (y_1)_u \\ (y_3)_v & (y_1)_v \end{vmatrix}, \quad \eta_3 = \begin{vmatrix} (y_1)_u & (y_2)_u \\ (y_1)_v & (y_2)_v \end{vmatrix}. \tag{2}$$

Set $|\eta(y)| = \sqrt{(\eta_1(y))^2 + (\eta_2(y))^2 + (\eta_3(y))^2}$. It is well known [15, Secs. 14.1 and 14.2] that

$$\int_\Gamma F(y)ds_y = \int_0^A du \int_0^B dv F(y(u,v))|\eta(y(u,v))|.$$

We require that the inequality

$$|\eta(y(u,v))| > 0, \quad \forall\, (u,v) \in ((0,A) \times (0,B)). \tag{3}$$

be satisfied. It follows from condition (3) that $|\eta(y(u,v))| \in C^1((0,A) \times (0,B))$.

By $\mathbf{n}_y$ we denote the unit normal at a point $y \in \Gamma$; i.e. $\mathbf{n}_y = \eta(y)/|\eta(y)|$. The derivative along the normal ny has the form

$$\frac{\partial}{\partial \mathbf{n}_y} = |\eta(y)|^{-1}(\eta(y), \nabla_y).$$

Set $|x - y(u,v)| = \sqrt{(x_1 - y_1(u,v))^2 + (x_2 - y_2(u,v))^2 + (x_3 - y_3(u,v))^2}$ and note that

$$\frac{\partial}{\partial \mathbf{n}_y}|x-y| = \frac{1}{|\eta(y)|}\sum_{j=1}^{3}\eta_j(y)\frac{y_j - x_j}{|x-y|}.$$

The double layer potential is used to solve boundary value problems for the Helmholtz equation by the method of integral equations. Let $\mu(y) \in C^0(\Gamma)$. Consider the direct value of the double layer potential at the point $x = y(u_{\hat{n}}, v_{\hat{m}}) \in \Gamma$

$$\mathcal{W}_k[\mu](x) = \frac{1}{4\pi}\int_\Gamma \mu(y)\frac{\partial}{\partial \mathbf{n}_y}\frac{e^{ik|x-y|}}{|x-y|}ds_y =$$

$$= \frac{1}{4\pi}\int_\Gamma \mu(y)\frac{1}{|\eta(y)|}\frac{\exp(ik|x-y|)(ik|x-y|-1)}{|x-y|^2}\sum_{j=1}^{3}\frac{\eta_j(y)(y_j - x_j)}{|x-y|}ds_y =$$

$$= \frac{1}{4\pi}\int_0^A du \int_0^B dv\, \mu(y(u,v))\exp(ik|x-y(u,v)|)(ik|x-y(u,v)|-1)\times$$

$$\times \sum_{j=1}^{3}\frac{\eta_j(y(u,v))(y_j(u,v) - x_j)}{|x-y(u,v)|^3} =$$

$$= \frac{1}{4\pi} \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} \int_{u_n-h/2}^{u_n+h/2} du \int_{v_m-H/2}^{v_m+H/2} dv\, \mu(y(u,v)) \times$$

$$\times \exp(ik|x-y(u,v)|)(ik|x-y(u,v)|-1) \sum_{j=1}^{3} \frac{\eta_j(y(u,v))(y_j(u,v)-x_j)}{|x-y(u,v)|^3}, \quad (4)$$

where $k \geq 0$. It is well known that [17, Sec. 27.5] the direct value of the double layer potential under our assumptions is continuous function on the surface Γ. Set $\mu_{nm} = \mu(y(u_n, v_m))$, then

$$\mu(y(u,v)) = \mu_{nm} + o(1), \quad (5)$$

for $u \in [u_n - h/2, u_n + h/2]$ and $v \in [v_m - H/2, v_m + H/2]$.

The same as in [4] it can be shown that with $u \in [u_n - h/2, u_n + h/2]$ and $v \in [v_m - H/2, v_m + H/2]$

$$|x - y(u,v)| = |x - y(u_n, v_m)| + O(h+H),$$

$$\exp(ik|x - y(u,v)|) = \exp(ik|x - y(u_n, v_m)|) + O(h+H).$$

Constants in estimates of functions denoted as $O(h+H)$, do not depend on n, m and on the location of x in the nodes of Γ. Therefore,

$$\mathcal{W}_k[\mu](x)|_{x=y(u_{\hat{n}}, v_{\hat{m}})\in\Gamma} \approx$$

$$\approx \frac{1}{4\pi} \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} \mu_{nm} \exp(ik|x - y(u_n, v_m)|)(ik|x - y(u_n, v_m)| - 1) \times$$

$$\times \int_{u_n-h/2}^{u_n+h/2} du \int_{v_m-H/2}^{v_m+H/2} dv \sum_{j=1}^{3} \frac{\eta_j(y(u,v))(y_j(u,v) - x_j)}{|x - y(u,v)|^3}. \quad (6)$$

Thus, to obtain a quadrature formula for the direct value of the double layer potential at a point $x = y(u_{\hat{n}}, v_{\hat{m}}) \in \Gamma$, one must calculate the integral

$$\int_{u_n-h/2}^{u_n+h/2} du \int_{v_m-H/2}^{v_m+H/2} dv \sum_{j=1}^{3} \frac{\eta_j(y(u,v))(y_j(u,v) - x_j)}{|x - y(u,v)|^3}, \quad (7)$$

which we refer to as the canonical integral.

3 Calculation of the Canonical Integral When the Point x Lies in the Domain of Integration

In this case, the integration is carried out on a rectangle with the center at the point $(u_{\hat{n}}, v_{\hat{m}})$, which corresponds to $y(u_{\hat{n}}, v_{\hat{m}}) = x$ on the surface Γ. Using the Taylor formula around the point $(u_{\hat{n}}, v_{\hat{m}})$, we obtain

$$|y(u,v)-x|^2 = |y(u,v)-y(u_{\hat{n}},v_{\hat{m}})|^2 \approx \sum_{j=1}^{3}((y_j)'_u(u-u_{\hat{n}})+(y_j)'_v(v-v_{\hat{m}}))^2 =$$

$$= \sum_{j=1}^{3}(((y_j)'_u)^2(u-u_{\hat{n}})^2+((y_j)'_v)^2(v-v_{\hat{m}})^2+2(y_j)'_u(y_j)'_v(u-u_{\hat{n}})(v-v_{\hat{m}})) =$$

$$= \alpha^2(u-u_{\hat{n}})^2+\beta^2(v-v_{\hat{m}})^2+2\delta(u-u_{\hat{n}})(v-v_{\hat{m}}),$$

$$\alpha^2 = \sum_{j=1}^{3}((y_j)'_u)^2, \quad \beta^2 = \sum_{j=1}^{3}((y_j)'_v)^2, \quad \delta = \sum_{j=1}^{3}(y_j)'_u(y_j)'_v,$$

where $(y_j)'_u$ and $(y_j)'_v$ are calculated at the point $(u_{\hat{n}}, v_{\hat{m}})$. Note that $\alpha^2\beta^2 - \delta^2 = |\eta(x)|^2$, according to [15, Sec. 14.1], therefore $\alpha^2 > 0$ and $\beta^2 > 0$ by virtue of the condition (3). Applying the Taylor expansion centered at the point $(u_{\hat{n}}, v_{\hat{m}})$ with remainder in the Peano form [11, Sec. 10.5.3], we find that

$$y_j - x_j = (y_j)'_u(u-u_{\hat{n}})+(y_j)'_v(v-v_{\hat{m}})+\frac{1}{2}(y_j)''_{uu}(u-u_{\hat{n}})^2+$$

$$+\frac{1}{2}(y_j)''_{vv}(v-v_{\hat{m}})^2+(y_j)''_{uv}(u-u_{\hat{n}})(v-v_{\hat{m}})+o\left((u-u_{\hat{n}})^2+(v-v_{\hat{m}})^2\right),$$

$$\eta_j(y(u,v)) = \eta_j(y(u_{\hat{n}},v_{\hat{m}}))+(\eta_j)'_u(u-u_{\hat{n}})+(\eta_j)'_v(v-v_{\hat{m}})+$$

$$+o\left(\sqrt{(u-u_{\hat{n}})^2+(v-v_{\hat{m}})^2}\right).$$

The derivatives with respect to u and v are taken at the point $(u_{\hat{n}}, v_{\hat{m}})$.

It is easy to check that [15, Sec. 14.1 and 14.2]

$$\sum_{j=1}^{3}\eta_j(y(u_{\hat{n}},v_{\hat{m}}))(y_j)'_u = \sum_{j=1}^{3}\eta_j(y(u_{\hat{n}},v_{\hat{m}}))(y_j)'_v = 0,$$

therefore

$$\sum_{j=1}^{3}\eta_j(y(u,v))(y_j-x_j)\approx\xi_1(u-u_{\hat{n}})^2+\xi_2(v-v_{\hat{m}})^2+\xi_3(u-u_{\hat{n}})(v-v_{\hat{m}}),$$

$$\xi_1=\sum_{j=1}^{3}\left(\frac{1}{2}\eta_j(y(u_{\hat{n}},v_{\hat{m}}))(y_j)''_{uu}+(\eta_j)'_u(y_j)'_u\right),$$

$$\xi_2=\sum_{j=1}^{3}\left(\frac{1}{2}\eta_j(y(u_{\hat{n}},v_{\hat{m}}))(y_j)''_{vv}+(\eta_j)'_v(y_j)'_v\right),$$

$$\xi_3=\sum_{j=1}^{3}\left(\eta_j(y(u_{\hat{n}},v_{\hat{m}}))(y_j)''_{uv}+(\eta_j)'_u(y_j)'_v+(\eta_j)'_v(y_j)'_u\right).$$

The derivatives with respect to u and v are taken at the point $(u_{\hat{n}}, v_{\hat{m}})$.

It follows from the above relations that in the case under consideration, the canonical integral (7) is approximately equal to the following integral, which we denote by $\mathcal{J}_{\hat{n}\hat{m}}$

$$\int_{u_{\hat{n}}-h/2}^{u_{\hat{n}}+h/2}du\int_{v_{\hat{m}}-H/2}^{v_{\hat{m}}+H/2}dv$$
$$\times\frac{\xi_1(u-u_{\hat{n}})^2+\xi_2(v-v_{\hat{m}})^2+\xi_3(u-u_{\hat{n}})(v-v_{\hat{m}})}{\left(\alpha^2(u-u_{\hat{n}})^2+\beta^2(v-v_{\hat{m}})^2+2\delta(u-u_{\hat{n}})(v-v_{\hat{m}})\right)^{3/2}}=$$
$$=\int_{-h/2}^{h/2}dU\int_{-H/2}^{H/2}dV\frac{\xi_1U^2+\xi_2V^2+\xi_3UV}{\left(\alpha^2U^2+\beta^2V^2+2\delta UV\right)^{3/2}}=\mathcal{J}_{\hat{n}\hat{m}},$$

where $U=u-u_{\hat{n}},\ V=v-v_{\hat{m}}$. By calculating the integral $\mathcal{J}_{\hat{n}\hat{m}}$ and moving on to the new integration variable z, we find this integral explicitly

$$\mathcal{J}_{\hat{n}\hat{m}}=\frac{h}{\beta^3}\left(-\frac{\xi_2 z}{\sqrt{z^2+(\alpha/\beta)^2-(\delta/\beta^2)^2}}+\xi_2\ln\left|z+\sqrt{z^2+(\alpha/\beta)^2-(\delta/\beta^2)^2}\right|-\right.$$
$$-\frac{\xi_3-2\xi_2\delta/\beta^2}{\sqrt{z^2+(\alpha/\beta)^2-(\delta/\beta^2)^2}}+$$
$$\left.+z\frac{\xi_2(\delta/\beta^2)^2+\xi_1-\xi_3\delta/\beta^2}{((\alpha/\beta)^2-(\delta/\beta^2)^2)\sqrt{z^2+(\alpha/\beta)^2-(\delta/\beta^2)^2}}\right)\Bigg|_{-H/h+\delta/\beta^2}^{H/h+\delta/\beta^2}-$$

$$-\frac{H}{\alpha^3}\left(-\frac{\xi_1 z}{\sqrt{z^2-(\delta/\alpha^2)^2+(\beta/\alpha)^2}}+\xi_1\ln\left|z+\sqrt{z^2-(\delta/\alpha^2)^2+(\beta/\alpha)^2)^2}\right|-\right.$$

$$-\frac{\xi_3-2\xi_1\delta/\alpha^2}{\sqrt{z^2-(\delta/\alpha^2)^2+(\beta/\alpha)^2}}+$$

$$\left.+z\,\frac{\xi_1(\delta/\alpha^2)^2+\xi_2-\xi_3\delta/\alpha^2}{(-(\delta/\alpha^2)^2+(\beta/\alpha)^2)\sqrt{z^2-(\delta/\alpha^2)^2+(\beta/\alpha)^2}}\right)\Bigg|_{h/H+\delta/\alpha^2}^{-h/H+\delta/\alpha^2}.$$

4 Calculation of the Canonical Integral When the Point *x* Does Not Lie in the Domain of Integration

Let x be a point not belonging to the small piece of Γ where the point $y = y(u, v)$ ranges for $(u - u_n) \in [-h/2, h/2]$ and $(v - v_m) \in [-H/2, H/2]$. We expand the function $y_j(u, v)$ by the Taylor formula around the point (u_n, v_m) and for $j = 1, 2, 3$ we obtain

$$y_j(u, v) = y_j(u_n, v_m) + D_j + O(H^2 + h^2),$$

where

$$D_j = (y_j)'_u(u - u_n) + (y_j)'_v(v - v_m).$$

Here and later the derivatives with respect to u and v are taken at the point(u_n, v_m). Set

$$r^2 = |x - y(u_n, v_m)|^2 = \sum_{j=1}^{3} r_j^2 \neq 0, \quad r_j = y_j(u_n, v_m) - x_j, \quad j = 1, 2, 3,$$

then

$$y_j(u, v) - x_j = r_j + D_j + O(H^2 + h^2), \quad j = 1, 2, 3.$$

Therefore,

$$|x - y(u,v)|^2 = \sum_{j=1}^{3}(x_j - y_j(u,v))^2 \approx \sum_{j=1}^{3}(r_j^2 + 2r_j D_j + D_j^2) =$$

$$= 2P(u-u_n) + 2Q(v-v_m) + \alpha^2(u-u_n)^2 + \beta^2(v-v_m)^2 + 2\delta(u-u_n)(v-v_m) +$$

$$+ r^2 = \beta^2(V + \delta U/\beta^2 + Q/\beta^2)^2 - (\delta U + Q)^2/\beta^2 + \alpha^2 U^2 + 2PU + r^2,$$

where $U = u - u_n,\ V = v - v_m,$

$$P = \sum_{j=1}^{3} r_j (y_j)'_u, \quad Q = \sum_{j=1}^{3} r_j (y_j)'_v, \quad \alpha^2 = \sum_{j=1}^{3} ((y_j)'_u)^2,$$

$$\beta^2 = \sum_{j=1}^{3} ((y_j)'_v)^2, \quad \delta = \sum_{j=1}^{3} (y_j)'_u (y_j)'_v.$$

It can be shown that [15, Sec. 14.1]

$$\alpha^2\beta^2 - \delta^2 = |\eta(y(u_n, v_m))|^2. \tag{8}$$

Since $|\eta(y(u_n, v_m))| > 0$ for all possible $n,\ m$ by condition (3), we have

$$\alpha^2\beta^2 - \delta^2 > 0. \tag{9}$$

It follows that $\alpha^2 > 0$ and $\beta^2 > 0$.

Applying the Taylor expansion centered at the point (u_n, v_m) with remainder in the Peano form [15, Sec. 10.5.3], we find that

$$\eta_j(y(u,v)) = \eta_j(y(u_n, v_m)) + (\eta_j)'_u(u-u_n) + (\eta_j)'_v(v-v_m) +$$

$$+ o\left(\sqrt{(u-u_n)^2 + (v-v_m)^2}\right).$$

The derivatives with respect to u and v are taken at the point (u_n, v_m). To compute the expression

$$\sum_{j=1}^{3} \eta_j(y(u,v))(y_j(u,v) - x_j)$$

with regard to the formulas

$$\sum_{j=1}^{3} \eta_j(y(u_n, v_m))(y_j)'_u = \sum_{j=1}^{3} \eta_j(y(u_n, v_m))(y_j)'_v = 0,$$

meaning the orthogonality of the normal vector to tangent vectors to the surface [15, Sec. 14, § 1.2]), we use the Taylor expansion around the point (u_n, v_m) with remainder in the Peano form,

$$y_j(u, v) - x_j = r_j + (y_j)'_u(u - u_n) + (y_j)'_v(v - v_m) + \frac{1}{2}(y_j)''_{uu}(u - u_n)^2 +$$

$$+ \frac{1}{2}(y_j)''_{vv}(v - v_m)^2 + (y_j)''_{uv}(u - u_n)(v - v_m) + o\left((u - u_n)^2 + (v - v_m)^2\right),$$

then

$$\sum_{j=1}^{3} \eta_j(y(u, v))(y_j(u, v) - x_j) \approx R + \xi_4 U + \xi_5 V + \xi_1 U^2 + \xi_2 V^2 + \xi_3 UV,$$

where $U = u - u_n$, $V = v - v_m$,

$$\xi_1 = \sum_{j=1}^{3} \left(\frac{1}{2} \eta_j(y(u_n, v_m))(y_j)''_{uu} + (\eta_j)'_u (y_j)'_u \right),$$

$$\xi_2 = \sum_{j=1}^{3} \left(\frac{1}{2} \eta_j(y(u_n, v_m))(y_j)''_{vv} + (\eta_j)'_v (y_j)'_v \right),$$

$$\xi_3 = \sum_{j=1}^{3} \left(\eta_j(y(u_n, v_m))(y_j)''_{uv} + (\eta_j)'_u (y_j)'_v + (\eta_j)'_v (y_j)'_u \right),$$

$$\xi_4 = \sum_{j=1}^{3} (\eta_j)'_u r_j, \quad \xi_5 = \sum_{j=1}^{3} (\eta_j)'_v r_j, \quad R = \sum_{j=1}^{3} \eta_j(y(u_n, v_m)) r_j.$$

All the derivatives with respect to u and v are taken at the point (u_n, v_m).

It follows from the above relations that the canonical integral (7) is approximately equal to the following integral, which we denote by $K_{nm}(x)$

$$\int_{u_n-h/2}^{u_n+h/2} du \int_{v_m-H/2}^{v_m+H/2} dv \frac{1}{|x-y(u,v)|^3} \sum_{j=1}^{3} \eta_j(y(u,v))(y_j(u,v)-x_j) \approx$$

$$\approx \int_{-h/2}^{h/2} dU \int_{-H/2}^{H/2} dV \times$$

$$\times \frac{R+\xi_4 U+\xi_5 V+\xi_1 U^2+\xi_2 V^2+\xi_3 UV}{\beta^3((V+\delta U/\beta^2+Q/\beta^2)^2-(\delta U+Q)^2/\beta^4+(\alpha^2U^2+2PU+r^2)/\beta^2)^{3/2}} =$$

$$= K_{nm}(x). \tag{10}$$

The integral $K_{nm}(x)$ is calculated explicitly in [18].

5 The Main Result

Let us state the main result of the present paper.

Theorem 1 *Let Γ be either a simple C^2-smooth closed surface bounding a spatially simply connected interior domain or a simple C^2-smooth bounded open oriented surface containing the limit points of itself. Let Γ admit the parametrization (1) with the property (3), and let $\mu(y) \in C^0(\Gamma)$. Then for the direct value of the double layer potential (4) for $x = y(u_{\hat{n}}, v_{\hat{m}}) \in \Gamma$ and $k \geq 0$ we have the quadrature formula*

$$\mathcal{W}_k[\mu](x)|_{x=y(u_{\hat{n}},v_{\hat{m}})\in\Gamma} \approx -\frac{1}{4\pi}\mu_{\hat{n}\hat{m}}\mathcal{J}_{\hat{n}\hat{m}} +$$

$$+\frac{1}{4\pi} \sum_{\substack{n=0,\, m=0 \\ (n,m)\neq(\hat{n},\hat{m})}}^{n=N-1,\, m=M-1} \mu_{nm} \exp(ik|x-y(u_n,v_m)|)(ik|x-y(u_n,v_m)|-1)K_{nm}(x). \tag{11}$$

where the integral $\mathcal{J}_{\hat{n}\hat{m}}$ is explicitly calculated in Sect. 3, and the integral $K_{nm}(x)$ from (10) is calculated explicitly in [18].

If $k = 0$, then the potential of the double layer for the Helmholtz equation passes into the potential of the double layer for the Laplace equation, respectively, the quadrature formula (11) at $k = 0$ takes the form of a quadrature formula for the direct value of the harmonic potential of the double layer on the surface of Γ.

6 Numerical Tests

Quadrature formula (11) is an alternative to the standard quadrature formula for the direct value of a double-layer potential on surface Γ, commonly used in engineering calculations [3, Chapter 2]

$$\mathcal{W}_k[\mu](x) \approx \frac{1}{4\pi} \sum_{\substack{n=0,\, m=0 \\ (n,m)\neq(\hat{n},\hat{m})}}^{n=N-1,\, m=M-1} \mu_{nm} \exp(ik|x-y(u_n, v_m)|)(ik|x-y(u_n, v_m)|-1)\times$$

$$\times \frac{hH}{|x - y(u_n, v_m)|^3} \sum_{j=1}^{3} \eta_j(y(u_n, v_m))(y_j(u_n, v_m) - x_j). \tag{12}$$

Testing improved (11) and standard (12) quadrature formulas was carried out in the case when the surface Γ is a sphere of unit radius. In tests, the exact direct value of the double layer potential at the nodal points was compared with the approximate values calculated by quadrature formulas—by the improved formula (11) according to the Theorem and by the standard formula (12). At each nodal point, the absolute error was calculated for both formulas. Calculations were carried for different values of M and N. Values of steps are given by formulas $h = 2\pi/N$, $H = \pi/M$. If $N/2 = M = 25$, then $h = H \approx 0.13$; if $N/2 = M = 50$, then $h = H \approx 0.063$; if $N/2 = M = 100$, then $h = H \approx 0.031$. The table for each test shows the maximum absolute calculation error for all nodal points of the sphere. The first line of the table contains the values of N, M, in the subsequent lines—maximum errors for the standard and improved quadrature formulas in each test.

Test 1 for quadrature formulas in the case of the Laplace equation. In this test, the potential density $\mu(y(u, v)) = 1$ was used, then the harmonic potential of the double layer and its direct value on the unit sphere have the form:

$$\mathcal{W}_0[\mu](x) = \begin{cases} 1 \text{ if } |x| < 1 \\ 0 \text{ if } |x| > 1 \end{cases}, \qquad \mathcal{W}_0[\mu](x)|_{|x|=1} = \frac{1}{2}.$$

Test 2 for quadrature formulas in the case of the Laplace equation. In this test, the potential density $\mu(y(u, v)) = \cos u \sin v$, was used, then the harmonic potential of the double layer and its direct value on the unit sphere have the form:

$$\mathcal{W}_0[\mu](x) = \begin{cases} \dfrac{2|x|\cos\varphi\sin\vartheta}{3} & \text{if } |x| < 1, \\ -\dfrac{\cos\varphi\sin\vartheta}{3|x|^2} & \text{if } |x| > 1. \end{cases}, \qquad \mathcal{W}_0[\mu](x)|_{|x|=1} = \frac{\cos\varphi\sin\vartheta}{6},$$

where φ and ϑ are azimuth and zenith angles in spherical coordinates with origin at the center of the sphere.

Test 3 for quadrature formulas in the case of the Laplace equation. In this test, the potential density $\mu(y(u, v)) = (3\cos^2 v - 1)/2$, was used, then the harmonic potential of the double layer and its direct value on the unit sphere have the form:

$$\mathcal{W}_0[\mu](x) = \begin{cases} \dfrac{3|x|^2(3\cos^2\vartheta - 1)}{10} & \text{if } |x| < 1 \\ -\dfrac{3\cos^2\vartheta - 1}{5|x|^3} & \text{if } |x| > 1 \end{cases}, \qquad \mathcal{W}_0[\mu](x)|_{|x|=1} = \frac{3\cos^2\vartheta - 1}{20},$$

Test 4 for quadrature formulas in the case of the Helmholtz equation. In this test, the potential density $\mu(y(u, v)) = \mu(y(u, v)) = k$, was used, then the harmonic potential of the double layer and its direct value on the unit sphere have the form:

$$\mathcal{W}_k[\mu](x) = \begin{cases} (1 - ik)\exp(ik)\dfrac{\sin(k|x|)}{|x|} & \text{if } |x| < 1, \\ (\sin k - k\cos k)\dfrac{\exp(ik|x|)}{|x|} & \text{if } |x| > 1, \end{cases}$$

$$\mathcal{W}_k[\mu](x)|_{|x|=1} = \frac{1}{2}((2 - ik)\sin k - \cos k)\exp(ik),$$

where $k = 1$ (Table 1).

The results of the test calculations show that the improved quadrature formula (11) has the first order of convergence, and the standard formula [3] converges more slowly. The error of calculations according to the improved quadrature formula proposed in the Theorem 1 is less than the error of calculations according to

Table 1 Maximum absolute error of quadrature formulas in tests 1–4

Number of the test	Quadrature formula	$M = N/2 = 25$	$M = N/2 = 50$	$M = N/2 = 100$
1	Standard	0.019	0.0097	0.0062
1	Improved	0.012	0.0063	0.0032
2	Standard	0.019	0.0097	0.0049
2	Improved	0.00050	0.00014	3.8E-5
3	Standard	0.011	0.0089	0.0062
3	Improved	0.011	0.0060	0.0031
4	Standard	0.019	0.0097	0.0062
4	Improved	0.012	0.0063	0.0032

the standard quadrature formula. Thus, the improved quadrature formula provides higher accuracy of calculations of the direct value of the potential of the double layer.

The improved quadrature formula can find application in the numerical solution of boundary integral equations arising in the process of solving boundary value problems for the Laplace and Helmholtz equations by the method of potentials.

References

1. Belotserkovsky, S.M., Lifanov, I.K.: Method of Discrete Vortices. CRC Press, Boca Raton (1993)
2. Lifanov, I.K.: Singular Integral Equations and Discrete Vortices. VSP, Zeist (1996)
3. Brebbia, C.A., Telles, J.S.F., Wrobel, L.C.: Boundary Elements Technique. Springer, Berlin (1984)
4. Krutitskii, P.A., Fedotova, A.D., Kolybasova, V.V.: Quadrature formula for the simple layer potential. Differ. Equ. **55**(9), 1226–1241 (2019)
5. Krutitskii, P.A., Reznichenko, I.O., Kolybasova, V.V.: Quadrature formula for the direct value of the normal derivative of the single layer potential. Differ. Equ. **56**(9), 1237–1255 (2020)
6. Krutitskii, P.A., Kwak, D.Y., Hyon, Y.K.: Numerical treatment of a skew-derivative problem for the Laplace equation in the exterior of an open arc. J. Eng. Math. **59**, 25–60 (2007)
7. Krutitskii, P.A., Kolybasova, V.V.: Numerical method for the solution of integral equations in a problem with directional derivative for the Laplace equation outside open curves. Differ. Equ. **52**(9), 1219–1233 (2016)
8. Krutitskii, P.A.: The Dirichlet problem for dissipative Helmholtz equation in a plane domain bounded by closed and open curves. Hiroshima Math. J. **28**(1), 149–168 (1998)
9. Krutitskii, P.A.: The Neumann problem for the 2-D Helmholtz equation in a domain bounded by closed and open curves. Int. J. Math. Math. Sci. **21**(2), 209–216 (1998)
10. Krutitskii, P.A.: The skew derivative problem in the exterior of open curves in a plane. Zeitschrift fur Analysis und ihre Anwendungen **16**(3), 739–747 (1997)
11. Krutitskii, P.A.: The 2-dimensional Dirichlet problem in an external domain with cuts. Zeitschrift fur Analysis und ihre Anwendungen **17**(2), 361–378 (1998)
12. Krutitskii, P.A.: The Neumann problem in a 2-D exterior domain with cuts and singularities at the tips. J. Differ. Equ. **176**(1), 269–289 (2001)
13. Krutitskii, P.A.: The 2-D Neumann problem in a domain with cuts. Rendiconti di Matematica e delle sue Applicazioni, Serie VII **19**(1), 65–88 (1999)
14. Krutitskii, P.A.: The mixed harmonic problem in an exterior cracked domain with Dirichlet condition on cracks. Comp. Math. with Appl. **50**, 769–782 (2005)
15. Butuzov, V.F., Krutitskaya, N.Ch., Medvedev, G.N., Shishkin, A.A.: Mathematical Analysis in Questions and Problems. Fizmatlit, Moscow (2000; in Russian)
16. Ilyin, V.A., Poznyak, E.G.: Fundamentals of Mathematical Analysis. Part 2. Mir Publishers, Moscow (1982)
17. Vladimirov, V.S.: Equations of Mathematical Physics. Mir Publishers, Moscow (1984)
18. Krutitskii, P.A., Reznichenko, I.O.: Quadrature formula for the harmonic double layer potential. Differ. Equ. **57**(7), 901–920 (2021)

Menchov–Trokhimchuk Theorem Generalized for Monogenic Functions in a Three-Dimensional Algebra

Maxim V. Tkachuk and Sergiy A. Plaksa

Abstract The aim of this work is to prove an analog of Menchov–Trokhimchuk theorem on weakening conditions of monogeneity for functions given in a concrete three-dimensional commutative algebra over the field of complex numbers. The property of monogeneity of a function is understood as a combination of its continuity with the existence of its Gâteaux derivative.

1 Introduction

In the algebra of complex numbers $\mathbb{C}$, a function $F: \mathbb{C} \longrightarrow \mathbb{C}$ is called monogenic at the point $\xi_0 \in \mathbb{C}$ if there exists the following finite limit:

$$\lim_{\xi \to \xi_0} \frac{F(\xi) - F(\xi_0)}{\xi - \xi_0}, \tag{1}$$

which is called the derivative of function F at the point ξ_0. A function, which is monogenic at all points of a domain $D \subset \mathbb{C}$, is called holomorphic in this domain (see [1]).

An idea to weaken conditions of holomorphicity of complex-valued functions is developed in papers of H. Bohr [2], H. Rademacher [3], D. Menchov [4–6], V. Fedorov [7], G. Tolstov [8], Yu. Trokhimchuk [9, 10], G. Sindalovski [11], D. Teliakovski [12], E. Dolgenko [13], M. Brodovich [14].

Let us introduce one of Menchov's conditions denoted as K''': it is said that a *function $F(\xi)$ satisfies the condition K''' at a point ξ_0 if the limit (1) exists for ξ belonging to the union of two noncollinear rays with origin at the point ξ_0.*

D. Menchov [4–6] showed that the fulfillment of condition K''' at every point of domain D, except an at most countable set of points, is sufficient for the mapping F to be conformal in the case where $F: D \to \mathbb{C}$ is continuous univalent function.

M. V. Tkachuk (✉) · S. A. Plaksa
Institute of Mathematics, National Academy of Sciences of Ukraine, Kiev, Ukraine

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_25

Yu. Trokhimchuk [9] removed the condition of univalence of the function F and proved the following theorem:

Menchov–Trokhimchuk Theorem *If a function $F : D \to \mathbb{C}$ is continuous in a domain D and the condition K''' is satisfied at any its point, except an at most countable set of points, then the function F is holomorphic in the domain D.*

A. Bondar [15] proved an analogue of this theorem for functions given in multidimensional complex space $\mathbb{C}^n$. More precisely, he proved that the continuity of function and the existence of its Fréche derivative along $2n$ specially chosen directions are sufficient for the holomorphy of such a function. For functions given in $\mathbb{C}^n$, A. Bondar [15] and V. Siryk [16] proved analogs of another Menchov–Trokhimchuk theorem using a certain condition of preservation of the angles. O. Gretskii [17] generalized the mentioned Bondar's results to the case of mappings of Banach spaces.

Our aim is to weaken the monogeneity conditions for functions given in commutative algebras over the complex field. The property of monogeneity of a function is understood as a combination of its continuity with the existence of its Gâteaux derivative.

In the paper [18], we proved an analog of Menchov–Trokhimchuk Theorem for functions given in a special real three-dimensional subspace of a three-dimensional commutative algebra $\mathbb{A}_3$ over the complex field. In the present paper, we give a complete proof of a similar statement for a function given in a real subspace of dimension k, $2 \le k \le 6$, of the algebra $\mathbb{A}_3$, that was announced in the paper [18], where a sketch of its proof was only presented.

2 Monogenic Functions in a Three-Dimensional Commutative Algebra with Two-Dimensional Radical

Consider a three-dimensional commutative algebra $\mathbb{A}_3$ with unit 1 over the complex field $\mathbb{C}$ and with a basis $\{1, \rho, \rho^2\}$ for which $\rho^3 = 0$. We define the Euclidean norm by the equality

$$\|a + b\rho + c\rho^2\| := \sqrt{|a|^2 + |b|^2 + |c|^2}\,, \qquad a, b, c \in \mathbb{C}\,.$$

The algebra $\mathbb{A}_3$ has a unique maximal ideal $\mathcal{I} := \{\lambda_1\rho + \lambda_2\rho^2 : \lambda_1, \lambda_2 \in \mathbb{C}\}$ which is also a radical of the algebra.

Consider a linear functional $f : \mathbb{A}_3 \to \mathbb{C}$ defined by the equality

$$f(a + b\rho + c\rho^2) = a\,. \tag{2}$$

Since the kernel of f is the maximal ideal $\mathcal{I}$, one can conclude that f is a continuous multiplicative functional (see [19, p. 135]).

Fix a real n-dimensional subspace $E_n := \{\zeta = x_1e_1 + x_2e_2 + \ldots + x_ne_n : x_1, x_2, \ldots, x_n \in \mathbb{R}\} \subset \mathbb{A}_3$, where $2 \leq n \leq 6$ and the vectors $e_1, e_2, \ldots, e_n$ are linearly independent over the field of real numbers $\mathbb{R}$ but, generally speaking, they do not form a basis of the algebra $\mathbb{A}_3$. Impose only one restriction on the choice of the subspace E_n: the image of E_n under the mapping f must be the whole complex plane $\mathbb{C}$ (see [20, 21]).

As particular cases of such subspaces, we can mention subspaces constructed on the harmonic bases $\{e_1, e_2, e_3\}$ of the algebra $\mathbb{A}_3$ that satisfy the condition $e_1^2+e_2^2+e_3^2 = 0$. These cases are important from the viewpoint of applications (see [22, 23]). The existence of harmonic bases is an essential prerequisite for the representation of solutions of the three-dimensional Laplace equation in the form of components of the expansions of differentiable functions with respect to the basis (see [22, 24, 25]).

It is well known that there are different types of differentiability of mappings in linear normalized spaces. First of all, the concepts of strong Fréchet differentiability and weak Gâteaux differentiability are used for the mentioned mappings (see, e.g. [19]). Let us note that the corresponding Fréchet and Gâteaux derivatives are defined as linear operators.

Formerly, for functions given in a domain of a finite-dimensional commutative associative algebra, G. Scheffers [26] considered a derivative, which is understood as a function given in the same domain. Generalizing such an approach to the case of mappings given in a domain of an arbitrary commutative associative Banach algebra, E. Lorch [27] introduced a strong derivative, which is also understood as a function given in the same domain.

A function $\Phi\colon \Omega \longrightarrow \mathbb{A}_3$ is called *differentiable in the sense of Lorch* in a domain $\Omega \subset E_3$ if for every $\zeta \in \Omega$ there exists an element $\Phi'_L(\zeta) \in \mathbb{A}_3$ such that for any $\varepsilon > 0$ there exists $\delta > 0$ such that for all $h \in E_3$ with $\|h\| < \delta$ the following inequality is fulfilled:

$$\|\Phi(\zeta + h) - \Phi(\zeta) - h\Phi'_L(\zeta)\| \leq \|h\|\varepsilon. \tag{3}$$

The Lorch derivative $\Phi'_L(\zeta)$ is a function of the variable ζ, i.e. $\Phi'_L : \Omega \to \mathbb{A}_3$. In this case, the mapping $B_\zeta : E_3 \to \mathbb{A}_3$, defined by the equality $B_\zeta h = h\Phi'_L(\zeta)$, is a bounded linear operator. Therefore, a function Φ, which is differentiable in the sense of Lorch in a domain Ω, has the Fréchet derivative B_ζ at every point $\zeta \in \Omega$. The converse statement is not true, see an example in [19, p. 116].

Using the Gâteaux differential, I. Mel'nichenko [25] suggested to consider the Gâteaux derivative as a function $\Phi'_G\colon \Omega \longrightarrow \mathbb{A}_3$ too.

If, for a function $\Phi\colon \Omega \longrightarrow \mathbb{A}_3$ given in a domain $\Omega \subset E_3$ and for every $\zeta \in \Omega$, there exists an element $\Phi'_G(\zeta) \in \mathbb{A}_3$ such that

$$\lim_{\delta \to 0+0} (\Phi(\zeta + \delta h) - \Phi(\zeta))\, \delta^{-1} = h\Phi'_G(\zeta) \quad \forall h \in E_3\,, \tag{4}$$

then we say that the function $\Phi'_G\colon \Omega \longrightarrow \mathbb{A}_3$ is the *Gâteaux derivative* of the function Φ.

It is clear that the existence of stronger Lorch derivative $\Phi'_L(\zeta)$ implies the existence of weaker Gâteaux derivative $\Phi'_G(\zeta)$ and the equality $\Phi'_L(\zeta) = \Phi'_G(\zeta)$. However, the existence of Fréchet derivative does not imply the existence of Gâteaux derivative $\Phi'_G(\zeta)$ that is illustrated by the mentioned example in [19, p. 116].

We say that a function $\Phi\colon \Omega \longrightarrow \mathbb{A}_3$ is *monogenic* in a domain $\Omega \subset E_3$ if Φ is continuous and has the Gâteaux derivative at every point of the domain Ω [23, 28, 29].

Despite the fact that the existence of the Gâteaux derivative does not imply the existence of the Lorch derivative, the monogenic functions $\Phi\colon \Omega \longrightarrow \mathbb{A}_3$ in a domain $\Omega \subset E_3$ are differentiable in the sense of Lorch in this domain. It follows from the representation of monogenic functions $\Phi(\zeta)$, $\zeta \in \Omega$, via holomorphic functions of the complex variable $f(\zeta)$ that is established in [23].

In the paper [30] one of the monogeneity conditions is weakened in the case $n = 3$, videlicet, it is proved that if the Gâteaux derivative of the function $\Phi : \Omega \longrightarrow \mathbb{A}_3$ exists at all points of a domain $\Omega \subset E_3$, then the continuity of the function Φ can be replaced by its local boundedness in the domain Ω.

3 Analog of the Menchov–Trokhimchuk Theorem for Monogenic Functions in Domains of a Fixed Subspace E_n of the Algebra $\mathbb{A}_3$

Let us introduce some notations and the terminology.

First of all, note that the radical $\mathcal{I}$ considered as a linear space over the field $\mathbb{R}$ has the dimension 4, that we shall call the real dimension. The intersection of the radical $\mathcal{I}$ with the linear space E_n is a set of noninvertable elements belonging to E_n. This set is a plane L_{E_n} of the real dimension $(n-2)$ due to the assumption that the image of E_n under the mapping f is the whole complex plane. In particular, L_{E_3} is a straight line and $L_{E_2} = \{0\}$.

Preimage of an arbitrary point $\xi \in \mathbb{C}$ under the mapping f is a plane $L^{\zeta}_{E_n} := \{\zeta + \eta : \eta \in L_{E_n}\}$, where ζ is an element from E_n such that $\xi = f(\zeta)$. It is obvious that the planes $L^{\zeta}_{E_n}$ and L_{E_n} are parallel.

Consider the following hypercomplex analog of the Menchov condition K''' in the algebra $\mathbb{A}_3$ for functions $\Phi : \Omega \to \mathbb{A}_3$ given in a domain $\Omega \subset E_n$:

Definition 1 We say that a function $\Phi : \Omega \to \mathbb{A}_3$ satisfies the condition $K'''_{\mathbb{A}_3, E_n}$ at a point $\zeta \in \Omega \subset E_n$ if there exists an element $\Phi_*(\zeta) \in \mathbb{A}_3$ such that the equality

$$\lim_{\delta \to 0+0} \left(\Phi(\zeta + \delta h) - \Phi(\zeta)\right) \delta^{-1} = h\Phi_*(\zeta) \tag{5}$$

is fulfilled for n vectors $h_1, h_2, h_3, \ldots, h_n$, which form a basis in E_n and, moreover, $h_3, \ldots, h_n$ form a basis in the plane L_{E_n}.

Note that, in the case where the function $\Phi : \Omega \to \mathbb{A}_3$ satisfies the condition $K'''_{\mathbb{A}_3,E_n}$ at different points of domain $\Omega \subset E_n$, the set of vectors $h_1, h_2, \ldots, h_n$ can be different at different points of this domain.

Lemma 1 *Let a domain $\Omega \subset E_n$ have connected intersections with the planes $L^{\zeta}_{E_n}$ for all $\zeta \in \Omega$ and a function $\Phi : \Omega \to \mathbb{A}_3$ of the form $\Phi(\zeta) = \rho^2\Phi_2(\zeta)$, where $\Phi_2(\zeta) \in \mathbb{C}$, be continuous in Ω and satisfy the condition $K'''_{\mathbb{A}_3,E_n}$ at all points $\zeta \in \Omega$, except an at most countable set of points. Then $\Phi_2(\zeta) = F_2(f(\zeta))$, where $F_2 : D \to \mathbb{C}$ is a holomorphic function in the domain D, which is the image of domain Ω under the mapping f.*

Proof Let $\zeta \in \Omega$ be an arbitrary point, where the function Φ satisfies the condition $K'''_{\mathbb{A}_3,E_n}$. We rewrite equality (5) for the function $\Phi(\zeta) = \rho^2\Phi_2(\zeta)$:

$$\lim_{\delta\to 0+0} \rho^2 \left(\Phi_2(\zeta + \delta h) - \Phi_2(\zeta)\right) \delta^{-1} = h\Phi_*(\zeta), \tag{6}$$

and note that it is fulfilled for $h \in \{h_1, h_2, \ldots, h_n\}$.

Substituting $h = h_1$ in equality (6) and taking into account the fact that h_1 is an invertible element of the algebra $\mathbb{A}_3$, we obtain

$$\Phi_*(\zeta) = \rho^2 h_1^{-1} \lim_{\delta\to 0+0} \left(\Phi_2(\zeta + \delta h_1) - \Phi_2(\zeta)\right) \delta^{-1} =: \rho^2\, \Psi(\zeta). \tag{7}$$

After the substitution of expression (7) for Φ_* in equality (6), we get:

$$\lim_{\delta\to 0+0} \rho^2 \left(\Phi_2(\zeta + \delta h) - \Phi_2(\zeta)\right) \delta^{-1} = h\rho^2\Psi(\zeta). \tag{8}$$

Now, after the substitution of values $h = h_3, \ldots, h_n$ into (8), we obtain zero in the right-hand part of equality (8) because $\{h_3, \ldots, h_n\} \subset \mathcal{I}$. Thus, the restriction of the function Φ_2 to the intersection of Ω with the plane $L^{\zeta}_{E_n}$ has the directional derivatives along the vectors $h = h_3, \ldots, h_n$ that are equal to zero at all points, except for a countable set. Moreover, the intersection of Ω with the plane $L^{\zeta}_{E_n}$ is a connected set. Then, by Trokhimchuk's Theorem 9 in the monograph [10, p. 103], the function Φ_2 is constant in the intersection of the domain Ω with the plane $L^{\zeta}_{E_n}$. This implies that the function Φ_2 can be represented in the form $\Phi_2(\zeta) = F_2(f(\zeta))$, where $F_2 : D \to \mathbb{C}$ is a function continuous in the domain D.

Let us prove that the function F_2 is holomorphic in the domain D.

First, we note that the equality

$$\rho^2\, h\Psi(\zeta) = \rho^2\, f(h) f(\Psi(\zeta))$$

follows from definition (2) of the functional f. We denote $\xi := f(\zeta)$ and rewrite equality (8) in the form

$$\rho^2 \lim_{\delta\to 0+0} \left(F_2(\xi + \delta f(h)) - F_2(\xi)\right) \delta^{-1} = \rho^2\, f(h) f(\Psi(\zeta)). \tag{9}$$

Since the multipliers next to ρ^2 on both sides of equality (9) take complex values, in view of the uniqueness of representation of an element of the algebra in the form of the linear combination of basis elements, we can conclude that the equality

$$\lim_{\delta\to 0+0}\left(F_2(\xi+\delta f(h))-F_2(\xi)\right)\delta^{-1}=f(h)f(\Psi(\zeta)),$$

is true for $h\in\{h_1,h_2\}$.

This yields the equalities

$$f(\Psi(\zeta))=\lim_{\delta\to 0+0}\left(F_2(\xi+\delta t_1)-F_2(\xi)\right)(\delta t_1)^{-1}=$$

$$=\lim_{\delta\to 0+0}\left(F_2(\xi+\delta t_2)-F_2(\xi)\right)(\delta t_2)^{-1},$$

where $t_1:=f(h_1)$, $t_2:=f(h_2)$.

Thus, at any point $\xi\in D$, except at most countably many points, the function F_2 has equal derivatives along two noncollinear rays with origin at the point ξ. This means that the continuous function F_2 satisfies the Menchov condition K''' at the point ξ. Therefore, by virtue of the Menchov–Trokhimchuk Theorem, the function F_2 is holomorphic in the domain D. □

Under the condition $a\neq 0$, every element $a+b\rho+c\rho^2$, $a,b,c\in\mathbb{C}$, has the inverse element, and its decomposition with respect to the basis $\{1,\rho,\rho^2\}$ has the following form:

$$(a+b\rho+c\rho^2)^{-1}=\frac{1}{a}-\frac{b}{a^2}\,\rho+\left(\frac{b^2}{a^3}-\frac{c}{a^2}\right)\rho^2.$$

Then

$$(t-a-b\rho-c\rho^2)^{-1}=\frac{1}{t-a}+\frac{b}{(t-a)^2}\,\rho+\left(\frac{c}{(t-a)^2}+\frac{b^2}{(t-a)^3}\right)\rho^2. \qquad (10)$$

Using this decomposition, we can easily write the decomposition with respect to the basis $\{1,\rho,\rho^2\}$ of the principal extension of a holomorphic function $F:D\to\mathbb{C}$ into the domain $\Pi:=\{\zeta\in E_n: f(\zeta)\in D\}$:

$$\frac{1}{2\pi i}\int\limits_{\gamma}F(t)(t-\zeta)^{-1}\,dt=F(f(\zeta))+(b_1x_1+b_2x_2+\ldots+b_nx_n)F'(f(\zeta))\,\rho+$$

$$+\left((c_1x_1+c_2x_2+\ldots+c_nx_n)F'(f(\zeta))+\frac{(b_1x_1+b_2x_2+\ldots+b_nx_n)^2}{2}F''(f(\zeta))\right)\rho^2$$

$$\forall\,\zeta=x_1e_1+x_2e_2+\ldots+x_ne_n\in\Pi, \qquad (11)$$

where i is the imaginary unit, the closed Jordan rectifiable curve γ lies in the domain D and encloses the point $f(\zeta) = a_1x_1 + a_2x_2 + \ldots + a_nx_n$, and the complex constants a_k, b_k, c_k, $k = 1, 2, \ldots, n$, are the coefficients of decompositions of the elements $e_1, e_2, \ldots, e_n$ with respect to the basis $\{1, \rho, \rho^2\}$:

$$\begin{aligned} e_1 &= a_1 + b_1\rho + c_1\rho^2, \\ e_2 &= a_2 + b_2\rho + c_2\rho^2, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ e_n &= a_n + b_n\rho + c_n\rho^2. \end{aligned}$$

Decomposition (11) generalizes a similar decomposition obtained in Theorem 1.7 from [22] under the additional assumption that $e_1 = 1$.

Lemma 2 *Let a domain $\Omega \subset E_n$ have connected intersections with the planes $L^{\zeta}_{E_n}$ for all $\zeta \in \Omega$ and a function $\Phi : \Omega \to \mathbb{A}_3$ be continuous in Ω and satisfy the condition $K'''_{\mathbb{A}_3, E_n}$ at all points $\zeta \in \Omega$, except an at most countable set of points. Then for all $\zeta \in \Omega$, the function Φ can be represented in the form*

$$\Phi(\zeta) = \frac{1}{2\pi i} \int\limits_{\gamma} \Big(F_0(\xi) + F_1(\xi)\rho + F_2(\xi)\,\rho^2\Big)(\xi - \zeta)^{-1}\, d\xi\,, \tag{12}$$

where F_0, F_1, F_2 are functions holomorphic in the domain D, which is the image of the domain Ω under the mapping f.

Proof Consider the decomposition of $\Phi(\zeta)$ with respect to the basis $\{1, \rho, \rho^2\}$:

$$\Phi(\zeta) = \Phi_0(\zeta) + \Phi_1(\zeta)\rho + \Phi_2(\zeta)\rho^2.$$

The function $\rho^2\Phi(\zeta) = \rho^2\Phi_0(\zeta)$ is continuous in Ω and satisfies the condition $K'''_{\mathbb{A}_3, E_n}$ at all points $\zeta \in \Omega$, except an at most countable set of points. Then, by Lemma 1, $\Phi_0(\zeta) = F_0(f(\zeta))$, where F_0 is a holomorphic function in the domain D, which is the image of the domain Ω under the mapping f.

It follows from equality (11) that the first components of decompositions with respect to the basis $\{1, \rho, \rho^2\}$ of the functions $\Phi(\zeta)$ and $\frac{1}{2\pi i}\int_{\gamma} F_0(\xi)(\xi - \zeta)^{-1}\, d\xi$ coincide in the domain Ω. This yields the equality

$$\Phi(\zeta) - \frac{1}{2\pi i} \int\limits_{\gamma} F_0(\xi)(\xi - \zeta)^{-1}\, d\xi = \Phi_{11}(\zeta)\,\rho + \Phi_{12}(\zeta)\,\rho^2 \quad \forall \zeta \in \Omega, \tag{13}$$

where Φ_{11}, Φ_{12} are complex-valued functions continuous in Ω.

Then the function $\rho(\Phi_{11}(\zeta)\rho + \Phi_{12}(\zeta)\rho^2) = \rho^2\Phi_{11}(\zeta)$ is continuous in Ω and satisfies the condition $K'''_{\mathbb{A}_3, E_3}$ at all points $\zeta \in \Omega$, except an at most countable set of points. By Lemma 1, $\Phi_{11}(\zeta) = F_1(f(\zeta))$, where F_1 is a holomorphic function in the domain D.

Further, as in the proof of equality (13), we obtain

$$\Phi_{11}(\zeta)\,\rho + \Phi_{12}(\zeta)\,\rho^2 - \rho\,\frac{1}{2\pi i}\int\limits_{\gamma} F_1(\xi)(\xi-\zeta)^{-1}\,d\xi = \Phi_{22}(\zeta)\,\rho^2 \quad \forall \zeta \in \Omega, \tag{14}$$

where Φ_{22} is a complex-valued function continuous in Ω.

As a consequence of equalities (13), (14), we get

$$\Phi(\zeta) - \frac{1}{2\pi i}\int\limits_{\gamma} F_0(\xi)(\xi-\zeta)^{-1}\,d\xi - \\ - \rho\,\frac{1}{2\pi i}\int\limits_{\gamma} F_1(\xi)(\xi-\zeta)^{-1}\,d\xi = \Phi_{22}(\zeta)\rho^2 \quad \forall \zeta \in \Omega. \tag{15}$$

Now, by Lemma 1, we have the equality $\Phi_{22}(\zeta) = F_2(f(\zeta))$, where F_2 is a holomorphic function in the domain D. This yields the equality

$$\rho^2\,\Phi_{22}(\zeta) = \rho^2\,F_2(f(\zeta)) = \rho^2\,\frac{1}{2\pi i}\int\limits_{\gamma} F_2(\xi)(\xi-\zeta)^{-1}\,d\xi \quad \forall \zeta \in \Omega\,. \tag{16}$$

Finally, we obtain representation (12) as a consequence of equalities (15) and (16). □

The main result is the following statement:

Theorem 1 *Let a domain $\Omega \subset E_n$ have connected intersections with the planes $L^{\zeta}_{E_n}$ for all $\zeta \in \Omega$ and a function $\Phi : \Omega \to \mathbb{A}_3$ be continuous in Ω and satisfy the condition $K'''_{\mathbb{A}_3,E_n}$ at all points $\zeta \in \Omega$, except an at most countable set of points. Then:*

(1) the function Φ is monogenic in the domain Ω;
(2) the function Φ can be extended to a function monogenic in the domain $\Pi := \{\zeta \in E_n : f(\zeta) \in D\}$. This extension is unique and is represented by equality (12) for all $\zeta \in \Pi$;
(3) the monogenic extension (12) of the function Φ is differentiable in the sense of Lorch in the domain Π.

All assertions of Theorem 1 are obvious consequences of representation (12).

References

1. Goursat, E.: Cours d'Analyse Mathématique, vol. 2. Gauthier-Villars, Paris (1910)
2. Bohr, H.: Über streckentreue und konforme Abbildung. Math. Z. **1**, 403–420 (1918)
3. Rademacher, H.: Über streckentreue und winkeltreue Abbildung. Math. Z., **4**, 131–138 (1919)
4. Menchov, D.: Sur les différentielles totales des fonctions univalentes. Math. Ann. **105**, 75–85 (1931)
5. Menchov, D.: Sur les fonctions monogénes. Bull. Soc. Math. France **59**, 141–182 (1931)
6. Menchov, D.: Les conditions de monogénéité. Act. Sci. Ind. No. 329 (1936)
7. Fedorov, V.S.: On monogenic functions. Mat. Sb. **42**(4), 485–500 (1935)
8. Tolstov, G.P.: On the curvilinear and iterated integral. Tr. Mat. Inst. Akad. Nauk SSSR **35**, 3–101 (1950)
9. Trokhimchuk, Yu.Yu.: Continuous Mappings and the Conditions of Monogeneity. Fizmatgiz, Moscow (1963; in Russian)
10. Trokhimchuk, Yu.Yu.: Differentiation, Inner Mappings, and Criteria of Analyticity. Institute of Mathematics, National Academy of Sciences of Ukraine, Kiev (2007; in Russian)
11. Sindalovskii, G.Kh.: On the Cauchy–Riemann conditions in the class of functions with summable modulus and some boundary properties of analytic functions. Mat. Sb. **128(170)**(3(11)), 364–382 (1985)
12. Telyakovskii, D.S.: Generalization of the Menchov theorem on functions that satisfy the condition K''. Mat. Zametki **76**(4), 578–591 (2004)
13. Dolzhenko, E.P.: Works by D. E. Menchov in the theory of analytic functions and the contemporary state of the theory of monogeneity. Usp. Mat. Nauk **47**(5), 67–96 (1992)
14. Brodovich, M.T.: On the mapping of a space domain that preserves angles and extensions along a system of rays. Sib. Mat. Zh. **38**(2), 260–262 (1997)
15. Bondar, A.V.: Local Geometric Characteristics of Holomorphic Mappings. Naukova Dumka, Kiev (1992; in Russian)
16. Sirik, V.I.: Some criteria for continuous mappings to be holomorphic. Ukr. Math. J. **37**(6), 621–626 (1985)
17. Gretskii, O.S.: On the C-differentiability of mappings of Banach spaces. Ukr. Math. J. **46**(10), 1472–1479 (1994)
18. Tkachuk, M.V., Plaksa, S.A.: An analog of the Menchov–Trokhimchuk theorem for monogenic functions in a three-dimensional commutative algebra. Ukr. Math. J. **73**(8), 1299–1308 (2022)
19. Hille, E., Phillips, R.S.: Functional Analysis and Semi-Groups. American Mathematical Society, Providence (1957)
20. Plaksa S.A., Pukhtaievych, R.P.: Monogenic functions in a finite-dimensional semi-simple commutative algebra. An. Stiint. Univ. "Ovidius" Constanta, Ser. Mat. **22**(1), 221–235 (2014)
21. Shpakivskyi, V.: Constructive description of monogenic functions in a finite-dimensional commutative associative algebra. Adv. Pure Appl. Math. **7**(1), 63–75 (2016)
22. Mel'nichenko, I.P., Plaksa, S.A.: Commutative Algebras and Spatial Potential Fields. Institute of Mathematics, National Academy of Sciences of Ukraine, Kiev (2008; in Russian)
23. Plaksa, S.A., Shpakovskii, V.S.: Constructive description of monogenic functions in a harmonic algebra of the third rank. Ukr. Math. J. **62**(8), 1251–1266 (2011)
24. Ketchum, P.W.: Analytic functions of hypercomplex variables. Trans. Am. Math. Soc. **30**, 641–667 (1928)
25. Mel'nichenko, I.P.: The representation of harmonic mappings by monogenic functions. Ukr. Math. J. **27**(5), 499–505 (1975)
26. Scheffers, G.: Verallgemeinerung der Grundlagen der gewöhnlich complexen Funktionen, I. Ber. Verh. Sachs. Akad. Wiss. Leipzig Math. Phys. Kl. **45**, 828–848 (1893)
27. Lorch, E.R.: The theory of analytic function in normed Abelian vector rings. Trans. Am. Math. Soc. **54**, 414–425 (1943)
28. Plaksa, S.A.: Commutative algebras associated with classic equations of mathematical physics. Adv. Appl. Anal. Trends Math., 177–223 (2012)

29. Plaksa, S.A.: Monogenic functions in commutative algebras associated with classical equations of mathematical physics. J. Math. Sci. **242**(3), 432–456 (2019)
30. Plaksa, S.A.: On differentiable and monogenic functions in a harmonic algebra. In: Proc. of the Institute of Mathematics, pp. 210–221. National Academy of Sciences of Ukraine, vol. 14, No. 1 (2017)

Part VI
Constructive Methods in the Theory of Composite and Porous Media

Monodromy of Pfaffian Equations for Group-Valued Functions on Riemann Surfaces

Grigory Giorgadze

Abstract We discuss several generalizations of Riemann-Hilbert monodromy problem formulated in terms of representations of compact Lie groups and motivated by recent applications in mechanics and modern mathematical physics. Generalizations of Riemann-Hilbert monodromy problem are developed in the framework of principal bundles of compact Lie groups and meromorphic connections on Riemann surfaces.

1 Introduction

It is known that for any vector bundle there exists connection with regular singularities at the given points (Plemelj's theorem) [2, 3]. This result can be generalized for holomorphic principal G-bundles [8]. To describe this generalization we consider a system of differential equations of the form $Df = \alpha f$, where α is a $\mathbf{g}$-valued 1-form defined on Riemann surface X, and $f : X \to G$ is a G-valued unknown function. Namely, let us define the operator

$$D : \Lambda^0(X, \mathbf{g}) \to \Lambda^1(X, \mathbf{g}) \tag{1}$$

by the formula

$$D_x(f)(u) = dr^{-1}_{f(x)}(df)_x(u),$$

where $r_g : G \to G$ be right shift on the group G, $C(X, G)$ be the group of all smooth functions $f : X \to G$, $\Lambda^p(X, G)$, $p = 0, 1, 2$, be the space of all $\mathbf{g}$-valued p-forms on X and $\mathbf{g}$ be the Lie algebra of G.

G. Giorgadze (✉)
Tbilisi State University, Tbilisi, Georgia
e-mail: gia.giorgadze@tsu.ge

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_26

An expression of the form

$$Df = \omega, \tag{2}$$

where ω is a $\mathbf{g}_{\mathbf{C}}$-valued 1-form on X and $f : X \to G_{\mathbf{C}}$ is an unknown smooth function, is called a *G-system of differential equations* [8, 9, 12], where $\mathbf{g}_{\mathbf{C}}$ and $G_{\mathbf{C}}$ denotes the complexification of $\mathbf{g}$ and G.

For a G-system, it is possible to formulate *Riemann-Hilbert monodromy problem* (RHMP) as follows:

(**RHMP**) *prove that, for a given discrete set $S = \{s_1, \dots, s_m\} \subset X$ and for a given homomorphism $\rho : \pi_1(X \setminus S, z_0) \to G_{\mathbf{C}}$, there exists a G-system of the type (2) with a 1-form ω which is holomorphic in $X \setminus S$ and monodromy of which coincides with ρ.*

The generalizations of Riemann-Hilbert monodromy problem, discussed below, are partially motivated by its many applications in geometry and mathematical physics, see, e.g. [1, 10–13].

It is known that solution of RHMP depends on group G [14]. In particular,
1) if $G = U(n)$, then

$$Df = df \cdot f^{-1}$$

and ω is a matrix of 1-forms on X, so that one obtains a usual system of the form

$$df = \omega f.$$

2) If $n = 1$, then $G_{\mathbf{C}} = \mathbf{C}^*$ and $Df = d \log f$, the logarithmic derivative of the function f.

Let

$$* : \Lambda^1(X; \mathbf{g}) \to \Lambda^1(X; \mathbf{g})$$

be the Hodge operator, then the complexification of de Rham complex $\Lambda^p_{\mathbf{C}}(X; \mathbf{g})$, $p = 0, 1, 2$, decomposes into the direct sum

$$\Lambda^1_{\mathbf{C}}(X; \mathbf{g}) = \Lambda^{1,0}(X; \mathbf{g}) \oplus \Lambda^{0,1}(X; \mathbf{g})$$

by the requirement that $* = -i$ on $\Lambda^{1,0}(X; \mathbf{g})$ and $* = i$ on $\Lambda^{0,1}(X; \mathbf{g})$. The operator D decomposes into the direct sum $D = D' \oplus D''$, where

$$D' : \Lambda^0(X; \mathbf{g}) \to \Lambda^{1,0}(X; \mathbf{g}), \quad D'' : \Lambda^0(X; \mathbf{g}) \to \Lambda^{0,1}(X; \mathbf{g}),$$

are determined by the formulæ

$$D'_x(f)(u) = d'_{r^{-1}_{f(x)}}(d'f)_x(u), \quad D''_x(f)(u) = d''r^{-1}_{f(x)}(d''f)_x(u).$$

A $G_{\mathbf{C}}$-valued function $f : X \to G_{\mathbf{C}}$ is called *holomorphic* (resp. antiholomorphic) if $D''f = 0$ (resp. $D'f = 0$).

The operator D has the following properties:

(1) it is a crossed homomorphism, i. e.

$$D(f \cdot g) = (Df)_x + (\mathtt{ad} f(x)) \circ (Dg)_x$$

for any $f, g \in C(X, G)$. Note that the operator D'' is also a crossed homomorphism,

(2) the kernel $\mathtt{ker} D$ consists of constant functions.

We will say that system (2) is *integrable* if, for any $x_0 \in X$ and $g_0 \in G$, there exists a solution f of this system in a neighborhood of x_0 satisfying $f(x_0) = g_0$.

A point x_0 is called an *isolated singular point* of a map $f : U \to G_{\mathbf{C}}$ if there is a punctured neighborhood U_{x_0} such that the map f is analytic in U_{x_0}.

We also will say that a $G_{\mathbf{C}}$-valued function $f \in \Omega(U_\epsilon(x_0))$ has a *polynomial growth* at the point x_0 if for each sector

$$S = \{z | \theta_0 \leq \arg z \leq \theta_1, 0 \leq |z| < \epsilon\},$$

where z denotes a local coordinate system on X, for sufficiently small ϵ, there exist an integer $k > 0$ and a constant $c > 0$ such that the inequality

$$d(f(z), \mathbf{1}) < c|z|^{-k}$$

holds. Here $d(\_, \mathbf{1})$ denotes the distance to the unit of group $G_{\mathbf{C}}$.

The properties (1), (2) of the operator D imply that if f_0 is some solution of system (2), then $f = f_0 h$ is also a solution for any $h \in \mathtt{ker} D$.

2 *G*-Systems on the Riemann Sphere

Consider a G-system (2) on the Riemann sphere $\mathbf{CP}^1$. Let f_0 be a solution of (2) in a neighborhood $U \subset \mathbf{CP}^1$ of the point z_0 and suppose f_0 has the polynomial growth at the points from the set $S = \{s_1, \ldots, s_m\}$. After continuation of f_0 along a path $\gamma_i \in \pi_1(\mathbf{CP}^1 \setminus S, z_0)$ starting and ending in z_0 and circling once around a singular point s_i, the solution f_0 transforms into another solution f_1: i.e. $\gamma_i^* f_0 = g_i f_1$ for some $g_i \in G$. Thus f_0 determines a representation

$$\rho : \pi_1(\mathbf{CP}^1 \setminus S, z_0) \to G_{\mathbf{C}}. \tag{3}$$

The subgroup $\mathtt{Im}\rho \subset G_{\mathbf{C}}$ is called the *monodromy group* of G-system (2) and the representation (3) induces a principal $G_{\mathbf{C}}$-bundle $P'_\rho \to \mathbf{CP}^1 \setminus S$. The form ω being a holomorphic connection for this bundle [4–6].

Consider an extension of the bundle $P'_\rho \to \mathbf{CP}^1 \setminus S$ a holomorphic principal bundle $P_\rho \to \mathbf{CP}^1$ [7, 18].

To extend the bundle $P'_\rho \to \mathbf{CP}^1 \setminus S$ to some point $s_i \in S$, consider a simple covering $\{U_j\}$ of $X_m = CP^1 \setminus S$, such that every intersection $U_{\alpha_1} \cap U_{\alpha_2} \cap \ldots \cap U_{\alpha_k}$ is simply connected. For each U_α, we choose a point $z_\alpha \in U_\alpha$ and join z_0 and s_α by a simple path γ_α starting at z_0 and ending at s_α. For a point $z \in U_\alpha \cap U_\beta$, we choose a path $\tau_\alpha \subset U_\alpha$ which starts at s_α and ends at z. Consider

$$g_{\alpha\beta}(z) = \rho\left(\gamma_\alpha \tau_\alpha(z)\, \tau_\beta^{-1}(z)\, \gamma_\beta^{-1}\right). \tag{4}$$

It is clear that $g_{\alpha\beta}(z) = g_{\beta\alpha}^{-1}(z)$ on $U_\alpha \cap U_\beta$ and $g_{\alpha\beta} g_{\beta\gamma}(z) = g_{\alpha\gamma}(z)$ on $U_\alpha \cap U_\beta \cap U_\gamma$.

The cocycle $\left\{g_{\alpha\beta}(z)\right\}$ is constant [2]. Hence from (4) we obtain a flat principal bundle, which is denoted by P'_ρ.

Let $\{t_\alpha(z)\}$ be a trivialization of our bundle:

$$t_\alpha : p^{-1}(U_\alpha) \to G_{\mathbf{C}}.$$

Consider the $\mathbf{g}$ valued 1-form

$$\omega_\alpha = -t_\alpha^{-1} dt_\alpha.$$

The cocycle $\left\{g_{\alpha\beta}(z)\right\}$ is constant on the intersection $U_\alpha \cap U_\beta$ and $g_{\alpha\beta}(z) t_\beta(z) = t_\alpha(z)$, so the identity $\omega_\alpha = \omega_\beta$ holds on $U_\alpha \cap U_\beta$. Indeed, replacing t_β by $t_\beta^{-1} g_{\alpha\beta}$ in the expression $\omega_\beta = -t_\beta^{-1} dt_\beta$, we obtain

$$\omega_\beta = -t_\alpha^{-1} g_{\alpha\beta}(z)\, dt_\alpha g_{\alpha\beta}^{-1}(z) = -t_\alpha^{-1} dt_\alpha.$$

The 1-form $\omega = \{\omega_\alpha\}$ is holomorphic on X_m and therefore it defines a connection 1-form of the bundle $P'_\rho \to X_m$. The corresponding connection is denoted by ∇'. We will extend the pair $\left(P'_\rho, \nabla'\right)$ to X.

As the required construction is of local character, we shall extend $P'_\rho \to X_m$ to the bundle $P''_\rho \to X_m \cup \{s_i\}$, where $s_i \in S$.

Let a neighborhood V_i of the point s_i intersect each of the open sets $U_{\alpha_1}, U_{\alpha_2}, \ldots, U_{\alpha_k}$ having s_i in its closure.

As we noted when constructing the bundle from transition functions (4), only one of them is different from identity. Denote by g_{1k} nonconstant cocycle, then $g_{1k} = M_i$, where M_i is the monodromy which corresponds to the singular point s_i and is obtained from representation (3). Mark a branch of the multi-valued function

$$(\tilde{z} - s_i)^{A_i}$$

containing the point $\tilde{s}_i \in \tilde{U}_i$ (where $2\pi i \exp(A_i) = M_i$). Thus the marked branch defines a function

$$g_{01} := \exp(A_j \ln(z - s_j). \tag{5}$$

Denote by g_{02} the extension of g_{01} along the path which goes around s_i counterclockwise, and similarly for other points. Hence on $U_i \cap U_{\alpha_k} \cap U_{\alpha_1}$ we shall have:

$$g_{0k}(z) = g_{01}(z)M_i = g_{01}(z)g_{0k}(z).$$

The function $g_{0k} : V_i \to G_{\mathbf{C}}$ is defined at the point s_i and takes there the value coinciding with the monodromy.

In a neighborhood of s_i one will have

$$\omega_i = dg_{0k}g_{0k}^{-1}.$$

If we use the above construction of extension for all points from S we obtain a holomorphic principal bundle $P_\rho \to \mathbf{CP}^1$ on the Riemann sphere $\mathbf{CP}^1$.

The holomorphic sections of P_ρ are solutions of the equation

$$\nabla f = 0 \iff Df = \omega, \tag{6}$$

where ω is the meromorhic 1-form of connection ∇. It means that $P_\rho \to \mathbf{CP}^1$ is induced by the system of the form (2) and the Atiyah class $a(P_\rho)$ is nontrivial [4]. $P_\rho \to \mathbf{CP}^1$ does not admit holomorphic connections and hence the system (2) must necessary have singular points.

Here and in the sequel under singular points will be meant critical singular points, i. e. ramification points of the solution.

The construction of extension of the bundle at the singular points of equation described above has local character. From this follows that this construction may be applied for any Riemann surface of higher genus.

The Birkhoff stratum Ω_κ consists of the loops from $L_pG_{\mathbf{C}}$ with fixed partial indices $K = (k_1, \ldots, k_r)$. Existence of a one-to-one correspondence between the Birkhoff strata Ω_K and holomorphic equivalence classes of principal bundles on $\mathbf{CP}^1$ is a generalization of Birkhof-Grothendieck theorem for holomorphic vector bundles on Riemann sphere [6, 16].

Theorem 1 *[4] Each loop $f \in \Omega G$ determines a pair (P, ξ), where P is a holomorphic principal $G_{\mathbf{C}}$-bundle on $\mathbf{CP}^1$ and ξ is a smooth section of the bundle $P|_{\bar{X}_\infty}$ holomorphic in X_∞, and if (P', ξ') and (P, ξ) are holomorphically equivalent bundles, then f' and f lie in the same Birkhoff stratum.*

The Theorem 1 implies that to each principal bundle with a fixed trivialization corresponds a tuple of integers $(k_1, \ldots, k_r)$ which completely determine holomor-

phic type of the principal bundle and hence if the holomorphic principal G-bundle is induced by the system of the form (2) without singular points, then this bundle is trivial [4, 18, 19].

Theorem 2 *If* $\mathrm{Im}\rho$ *is connected then the Riemann-Hilbert monodromy problem is solvable for any m points* $s_1, \ldots, s_m$.

Indeed, let $\gamma_1, \ldots, \gamma_m$ be generators of $\pi_1(X_m, z_0)$. Suppose $\rho_1 = \rho(\gamma_1), .., \rho_m = \rho(\gamma_m)$. If $\mathrm{Im}\rho$ is a connected subgroup then there exists a continuous path $\rho_j(t)$, such that $\rho_i(0) = 1$ and $\rho_j(1) = \rho_j$. From this follows that there exists homomorphism $\chi_t : \pi_1(X_m, z_0) \to G$, such that $\chi_t(\gamma_j) = \rho_j(t)$.

Theorem 2 is proved.

Here we use the following general result from homological algebra: if G_1 is some connected group, then the homomorphism $h : \pi_1(X_m, z_0) \to G_1$ is the monodromy homomorphism of a G-system if and only if it is possible to connect h to 1 by continuous path in group of cochains $Z^1(\pi_1(M), G)$ [14].

Remark In general, ω may have a pole at infinity whose order is more that 1. For example, see [3].

Theorem 3 *[8, 12]. Suppose* $\rho(\gamma_j) \in \mathbf{T}$ *for some* j. *Then* $\rho : \pi_1(X \setminus S) \to G$ *is the monodromy of a regular G-system.*

Theorem 3 is proved using the Plemelj's scheme [18]. The properties $\rho(\gamma_j) \in \mathbf{T}$ guarantee that the gauge transformation reduces the regular system to a system with singularities of first order at all singular points [2].

3 G-System on the Riemann Surfaces of Genus $g \geq 2$

As above suppose that G is a connected compact Lie group and $G_{\mathbf{C}}$ is its complexification; $\mathbf{g}$ and $\mathbf{g}_{\mathbf{C}}$ are the Lie algebras of the group G and $G_{\mathbf{C}}$, respectively; Z is the centrum of the group $G_{\mathbf{C}}$, and Z_0 is the connected component of the unit; X is a compact connected Riemann surface of genus $g \geq 2$. If $\tilde{X} \to X$ is a universal covering and $\rho : \pi_1(X) \to G_{\mathbf{C}}$ is a representation, then the corresponding principal bundle will be denoted P_ρ.

Let $x_0 \in X$ be a fixed point and $p : \tilde{X} \to X \setminus \{x_0\}$ be a universal covering, then the triple $(\tilde{X}, p, X \setminus \{x_0\})$ is a principal bundle whose structure group Γ is a free group on $2g$ generators, and if γ is a loop circling around x_0 then $\gamma = \prod_{i=1}^{g}[a_i, b_i]$, where a_i, b_i are generators of $\Gamma \cong \pi_1(X \setminus \{x_0\})$ and $[\_, \_]$ denotes the commutator.

Let $P'_\rho \to X \setminus \{x_0\}$ be the principal bundle corresponding to the representation $\rho : \pi_1(X \setminus \{x_0\}) \to G_{\mathbf{C}}$. Since by Theorem 1 each loop $f : S^1_X \to G$ determines a holomorphic principal $G_{\mathbf{C}}$-bundle, using f one can extend the bundle $P'_\rho \to X \setminus \{x_0\}$ to X in the following way: let U_{x_0} be a neighborhood of x_0 homeomorphic to a unit disc and consider the trivial bundles $U_{x_0} \times G_{\mathbf{C}} \to U_{x_0}$ and $P'_\rho \to X \setminus \{x_0\}$. Let us glue these bundles over the intersection $(X \setminus \{x_0\}) \cap U_{x_0} = U_{x_0} \setminus \{x_0\}$ using the loop f. We thus obtain an extended bundle $P_\rho \to X$.

Consider the homomorphism of fundamental groups

$$f_* : \pi_1(S^1_X) \to \pi_1(G_{\mathbf{C}})$$

induced by f and suppose that γ is a generator of $\pi_1(S^1_X)$ mapped to $+1$ under the isomorphism $\pi_1(S^1_X) \cong \mathbf{Z}$. If $f' : S^1_X \to G_{\mathbf{C}}$ is homotopic to f, then $f'_* = f_*$ and therefore f and f' corresponds to topologically equivalent $G_{\mathbf{C}}$-bundles on X. Conversely, for any element $c \in \pi_1(G_{\mathbf{C}})$ there exists $f_* : \pi_1(S^1_X) \to \pi_1(G_{\mathbf{C}})$ with $f_*(\gamma) = c$ [17].

Let $P \to X$ be a principal bundle and f the corresponding loop. The element

$$\chi(P) := f_*(\gamma) \in \pi_1(G_{\mathbf{C}})$$

of the fundamental group is the *characteristic class* of the bundle P [17].

It is known that the map

$$\chi : H^1(X; C^\infty(G_{\mathbf{C}})) \to \pi_1(G_{\mathbf{C}})$$

determined by the formula $\chi(P) = c$ for each $P \in H^1(X; C^\infty(G_{\mathbf{C}}))$ is surjective. Here $C^\infty(G_{\mathbf{C}})$ denotes the sheaf of germs of continuous maps $X \to G_{\mathbf{C}}$.

Let $p : \tilde{G}_{\mathbf{C}} \to G_{\mathbf{C}}$ be the universal cover of the group, with fibre $\pi_1(G_{\mathbf{C}}) \cong \mathtt{ker}\, p$. The exact sequence of groups

$$1 \to \pi_1(G_{\mathbf{C}}) \to \tilde{G}_{\mathbf{C}} \to G_{\mathbf{C}} \to 1 \tag{7}$$

induces the exact sequence of sheaves

$$1 \to \pi_1(G_{\mathbf{C}}) \to C^\infty(\tilde{G}_{\mathbf{C}}) \to C^\infty(G_{\mathbf{C}}) \to 1. \tag{8}$$

Since $\pi_1(G_{\mathbf{C}})$ is contained in the center of the group $\tilde{G}$, the sequences (7) and (8) yield the following commutative diagram

$$\begin{array}{ccc} \delta : H^1(\pi_1(X); G_{\mathbf{C}}) & \to & H^2(\pi_1(X); \pi_1(G_{\mathbf{C}})) \\ \downarrow \mu & & \downarrow \nu \\ \delta : H^1(X; G_{\mathbf{C}}) & \to & H^2(X; \pi_1(G_{\mathbf{C}})) \\ \downarrow i^* & & \downarrow \mathrm{id} \\ \delta : H^1(X; C^\infty(G_{\mathbf{C}})) & \to & H^2(X; \pi_1(G_{\mathbf{C}})) \end{array} \tag{9}$$

where μ, ν are isomorphims and i^* is induced by the embedding $i : G_{\mathbf{C}} \hookrightarrow C^\infty(G_{\mathbf{C}})$. The coboundary operator

$$\delta : H^1(X; C^\infty(G_{\mathbf{C}})) \to H^2(X; \pi_1(G_{\mathbf{C}})) \cong \pi_1(G_{\mathbf{C}})$$

from the last row of the diagram (9) equals χ (see [17]).

Let $\rho : \pi_1(X \setminus x_0) \to G_{\mathbf{C}}$ be a representation such that $\rho(S^1_X) = c \in Z_0$. If $\tilde{Z}_0$ is the Lie algebra of group Z_0, then $\exp : \tilde{Z}_0 \to Z_0$ is a universal covering. Let us choose an element $\alpha \in \tilde{Z}_0$ such that $\exp \alpha = c$. Extend the bundle $P'_\rho \to X \setminus x_0$ to X using the loop $f : S^1_X \to G$, with

$$f(z) = \exp(\alpha \ln(z - x_0))$$

on S^1_X. Denote the obtained principal bundle by $P_{\rho,\alpha} \to X$.

The space $H \subset G$ is called *irreducible* if

$$\{Y \in \mathbf{g} \mid \forall h \in H \ \mathrm{ad}h(Y) = Y\} = \text{ center } \mathbf{g}.$$

The representation $\rho : \Gamma \to G_{\mathbf{C}}$ is called *unitary* if $\rho(\Gamma) \subset G$, and $\rho : \Gamma \to G$ is called *irreducible*, if $\rho(\Gamma)$ is irreducible. The following theorem gives a useful criterion of holomorphic equivalence of G-bundles.

Theorem 4 *[17] Let ρ and ρ' be unitary representations of the group $\Gamma \cong \pi_1(X \setminus \{x_0\})$ in G. The bundles $P_{\rho,\beta}$ and $P_{\rho',\beta'}$ are holomorphically equivalent if and only if ρ and ρ' are equivalent in a maximal compact subgroup of $G_{\mathbf{C}}$ and $\beta = \beta'$.*

Let M be any connected smooth manifold (compact or not) and let $\rho : \pi_1(M) \to G_{\mathbf{C}}$ be any homomorphism. Then from Theorem 4 follows, that

1) *if $\pi_1(M)$ is a free group and $G_{\mathbf{C}}$ is connected, then ρ is a monodromy homomorphism for a G-system (2)* (see [12]).
2) *If $\pi_1(M)$ is a free abelian group and G is a connected compact Lie group with torsion free cohomology, and if $\mathrm{Im}\rho \subset G$, then ρ is a monodromy homomorphism for some G-system of the type (2)* (see [12]).

Acknowledgments The research is partially supported by the GNSF project titled "Problem of factorization and invariants of holomorphic bundles on Riemann surfaces", under grant agreement number 22_354.

References

1. Abhijeet, A.: Boundary Physics and Bulk-Boundary Correspondence in Topological Phases of Matter. Springer, Berlin (2019). https://doi.org/10.1007/978-3-030-31960-1
2. Anosov, D.V., Bolibruch, A.A.: The Riemann-Hilbert problem. Aspects of Mathematics. Vieweg, Braunschweig, Wiesbaden (1994)
3. Arnold, V.I., Il'yashenko, Y.S.: Ordinary differential equations. Dynamical Systems I. Encyclopaedia of Mathematical Sciences, vol. 1. Springer, Berlin (1988)
4. Atiyah, M.F.: Vector bundles over an elliptic curve. Proc. Lond. Math. Soc. **3**(7), 414–452 (1957). https://doi.org/10.1112/plms/s3-7.1.414
5. Atiyah, M.F.: Instantons in two and four dimensions. Commun. Math. Phys. **93**(4), 437–451 (1984). https://doi.org/10.1007/BF01212288

6. Atiyah, M.F., Bott, R.: the yang-mills equation on the Riemann surface. Philos. Trans. R. Soc. Lond. A **308**, 523–615 (1982). https://doi.org/10.1098/rsta.1983.0017
7. Deligne, P.: Equations Différentiales á Points Singuliers Réguliers. Lecture Notes in Mathematics, vol. 163. Springer, Berlin (1970)
8. Giorgadze, G.: *G*-systems and holomorphic principal bundles on Riemann surfaces. J. Dyn. Contr. Syst. **8**(2), 245–291 (2002). https://doi.org/10.1023/A:1015321627073
9. Giorgadze, G.: On holomorphic principal bundles on a Riemann surface. Bull. Georgian Acad. Sci. **166**(1), 27–31 (2002)
10. Giorgadze, G., Khimshiashvili, G.: Factorization of loops in loop groups. Bull.Georgian Natl.Acad. Sci. **5**(3), 35–38 (2011)
11. Giorgadze, G., Khimshiashvili, G.: The Riemann-Hilbert problem in loop spaces. Doklady Math. **73**(2), 258–260 (2006)
12. Giorgadze, G., Khimshiashvili, G.: Riemann-Hilbert problems with coefficients in compact Lie groups. In: Andrianov, I., Gluzman, S., Mityushev, V. (eds.) Mechanics and Physics of Structured Media. Academic Press, Cambridge (2022). https://doi.org/10.1016/B978-0-32-390543-5.00020-7
13. Gluzman, S., Mityushev, V., Nawalaniec, W.: Computational Analysis of Structured Media. Academic Press, Cambridge (2017)
14. Onishchik, A.: Some concepts and applications of non-abelian cohomology theory. Trans. Mosc. Math. Soc. **17**, 49–97 (1967)
15. Plemelj, J.: Problems in the Sense of Riemann and Klein. Intersience Publishers/A division of J.Wiley & Sons, New York/London/Sidney (1964)
16. Pressley, A., Segal, G.: Loop Groups. Clarendon Press, Oxford (1984)
17. Ramanathan, A.: Stable principal bundles on a compact Riemann surface. Math. Ann. **213**, 129–152 (1975)
18. Röhrl, H.: Holomorphic vector bundles over Riemann surfaces. Bull. Am. Math.Soc. **68**(3), 125–160 (1962)
19. Shatz, S.S.: On subbundles of vector bundles over $\mathbf{CP}^1$. J. Pure Appl. Algebra, **10**, 315–322 (1977)

Introduction to Neoclassical Theory of Composites

Simon Gluzman

Abstract Brief review of main tenets of the neoclassical theory of composites is given. Several examples of its application are given, including two-dimensional conductivity of regular composites, three-dimensional superconductivity of random composites, conductivity of liquid foams in two-and-three dimensions, and permeability for the viscous flow in three-dimensional channels.

1 Introduction

Classical theory of composites amounts to the celebrated Maxwell formula, also known as Clausius–Mossotti approximation. Actually all modern self-consistent methods are justified only for a dilute composites when interactions among inclusions are neglected. Careful analysis shows their restriction to the first- or second-order approximations in concentration. In the same time, exact and high-order formulae for special regular composites which go beyond self-consistent methods were derived, starting with Rayleigh and continued in particular by McPhedran et al.

We are primarily concerned here with the effective properties of deterministic and random composites and porous media. The analysis leads to accurate analytical approximate solutions to the problems when it is impossible to find their exact solutions. Certain problems of micromechanics and their analogs such as boundary value problems for Laplace's equation and bi-harmonic two-dimensional (2D) elasticity equations can be solved in analytical form. At least for an arbitrary 2D multiply connected domain with circular inclusions there are methods which yield analytical formulae for most of the important effective properties, such as conductivity, permeability, effective shear modulus and effective viscosity [6, 16, 28]. Randomness for such problems is introduced through random locations of non-overlapping disks.

S. Gluzman (✉)
Materialica+ Research Group, Toronto, ON, Canada

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_27

Discrete numerical solutions such as finite elements rather powerful and their application makes sense when the geometries and the physical parameters are fixed. In this case the researcher can be fully satisfied with numerical solutions to various boundary value problems. But various numerical packages ought not to be viewed as a universal remedy, since a sackful of numbers is not as useful as an accurate analytical formulae. Pure numerical procedures fail as a rule for the situations with criticality, and analytical matching with asymptotic solutions can be useful even for the numerical computations.

In other words, there is an unlimited belief in numerical methods and regretful underestimation of constructive analytical and asymptotic methods. The situation has to be drastically reconsidered. There are three major neoclassical developments which warrant such a view.

1. Recent mathematical results devoted to explicit solutions to the Riemann–Hilbert and $\mathbb{R}-$linear problems for multiply connected domains [6, 16].
2. Significant progress in symbolic computations greatly extends our computational capacities. Symbolic computations operate on the meta-level of numerical computing. They transform pure analytical constructive formulae into computable objects. Such an approach results in symbolic algorithms which often require optimization and detailed analysis from the computational point of view. Moreover, symbolic and numeric computations can be integrated [6, 16].

   The former two developments allow to obtain the expressions for various physical quantities in the form of truncated series which could be treated as polynomials. They are supposed to reflect accurately enough on their respective infinite expansions, so that with the help of some additional resummation procedure one can extrapolate to the whole series. But even long truncated power series in concentration and contrast parameters are not sufficient because they won't allow us to cover the high-concentration regime. Sometimes the series are short, in other cases they do not converge fast enough, or even diverge in the most interesting regime. Your typical answer to the challenges is to apply additional methods powerful enough to extract information from the series. But in addition to a traditional Padé approximants [2] applied in such cases, the practitioner would require a
3. New post-Padé approximants for analysis of the divergent or poorly convergent series, including different asymptotic regimes as suggested in [6, 11, 16].

As to the engineering needs we recognize the need for an additional fourth step. We can safely assume that the engineer would like to have a convenient formula but also to incorporate in it all available information on the system, with a particular attention to the results of numerical simulations or known experimental values by applying, for instance, the method of "regression on approximants" suggested in book [16]. We present below several typical examples of neoclassical developments where the main neoclassical ideas and methods can be seen in action. The neoclassical approach dwells on classical Maxwellian, but adds three relatively modern ideas just presented above.

2 Example of Crossover in Physics

The techniques and ideas briefly discussed above are geared towards faithful description of various crossover phenomena. For instance, the low-concentration regimes are described by a truncated, sometimes long power series. In the high-concentration regime one often encounters the power laws. In many problems of material sciences one encounters the so-called crossover phenomena, when a physical quantity qualitatively changes its behavior in different domains of its variable. To be more precise, one can specify a crossover as follows. Let a function represent a physical quantity of interest, with a variable x running through the interval $x_1 \leq x \leq x_2$. Let also the behavior of this function be essentially different near the boundary points x_1 and x_2. Assume that the function varies continuously as x changes from x_1 to x_2. Then one may say that the function in the interval $[x_1, x_2]$ undergoes a crossover between the two limiting behaviors. Brilliant work by Koiter [18] should be mentioned here. It warrants a fresh look in connection with recent advances. The approach advanced in [6, 16] allows, in addition to interpolation, also to calculate the critical indices, amplitudes and relaxation times [12].

Even when there are two known expressions at different boundaries, it may be not clear to connect them, say, with conventional splines. Yet, the approximants based on the requirements of asymptotic equivalence with the truncated series are able to smoothly connect the two apparently disconnected truncated expansions, as demonstrated by the example below. Lieb and Liniger [23] have considered a one-dimensional Bose gas with contact interactions. The ground-state energy of the gas can be written as a weak-coupling expansion, with respect to the coupling parameter g [29, 31], as

$$E(g) \simeq g - \frac{4}{3\pi} g^{3/2} + \frac{1.29}{2\pi^2} g^2 - 0.017201 g^{5/2} , \tag{1}$$

as $g \to 0$. In the strong-coupling limit, as $g \to \infty$, there is the following expression [29, 31]

$$E(g) \simeq \frac{\pi^2}{3}\left(1 - \frac{4}{g} + \frac{12}{g^2}\right) . \tag{2}$$

In what follows the approximant $E^*_{3+3}(g)$ assimilates the three coefficients from weak and strong coupling expansions, while $E^*_{4+3}(g)$ is based on all four terms from the weak-coupling side.

The accuracy of the root approximant [6, 16]

$$E^*_{3+3}(g) = \frac{\pi^2}{3\sqrt[5]{\frac{385.383}{g^5} + \left(\frac{388.171}{g^4} + \left(\frac{164.914}{g^3} + \left(\frac{37.3454}{g^2} + \left(\frac{8.12698}{g} + 1\right)^{3/2}\right)^{5/4}\right)^{7/6}\right)^{9/8}}}, \tag{3}$$

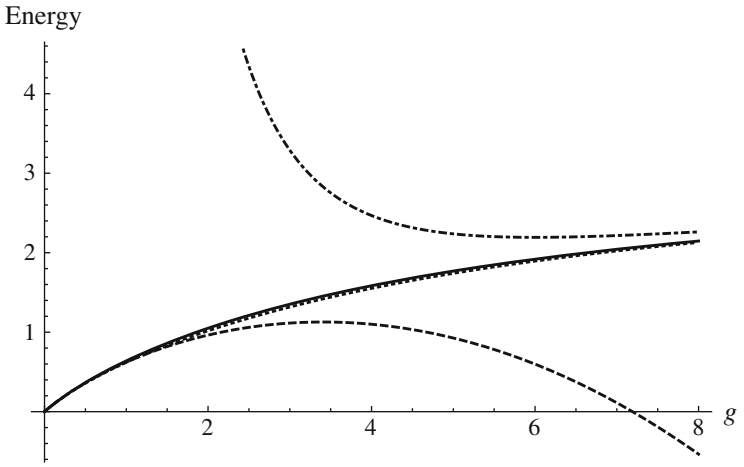


Fig. 1 The interpolation with root approximant (3) is shown with solid line, while the Padé approximant given by the formula A22 from [12], is shown with dotted line. The weak (dashed) and strong-coupling (dot-dashed) expansions are shown as well

turns out to be good as it connected a known asymptotic expansion at the right boundary of the interval with a known asymptotic form at the left boundary.

It should become completely clear from observing Fig. 1, that the problem of interpolation is neither simple, nor superficial. The asymptotic expressions for small and large couplings have little in common with each other. Although the expansions (1) and (2) appear to work only for very small and very large coupling constants, the deduced approximant works rather well everywhere.

3 2D Conductivity: Dependence on Contrast Parameter

Consider a classical problem of the effective conductivity (thermal, electric et.) of a 2D regular composite. An accurate approximate formula can be deduced for a 2D, two-component composite made from a collection of non-overlapping, identical, ideally conducting circular disks, embedded regularly in an otherwise uniform, locally isotropic host. Let σ denote the ratio of the conductivity of inclusions to the matrix conductivity. Usually, the conductivity of the matrix σ_0 is normalized to unity. Introduce the contrast parameter

$$\varrho = \frac{\sigma - 1}{\sigma + 1}, \tag{4}$$

so that $|\varrho| \leq 1$.

The exact formula for the effective conductivity tensor of an arbitrary regular array was written in the most general form [16, eq.(4.2.28)] as a power series in ϱ and concentration (volume fraction) of inclusions f. In the case of a square array of inclusions this series diverges as $f \to f_c = \frac{\pi}{4} \approx 0.7854$ and $\varrho \to 1$. It is difficult to analytically investigate such singular behavior. However, the effective conductivity is known in the form of the other asymptotic formulas [16, 19, Chapter 6]. Consider the truncated expansion for the effective conductivity of the square array

$$\sigma_e \approx \tfrac{1+f\varrho}{1-f\varrho} + 0.611654 f^5 \varrho^3 + 1.22331 f^6 \varrho^4 + 1.83496 f^7 \varrho^5 + 2.44662 f^8 \varrho^6. \tag{5}$$

For many years it was thought that Maxwell's and Clausius-Mossotti approximation for the effective conductivity of 2D (3D) composites

$$\sigma_e = \frac{1+\varrho f}{1-\varrho f} + O(f^2), \tag{6}$$

can be systematically and rigorously extended to higher orders in f by taking into account interactions between pairs of spheres, triplets of spheres, and so on. However, it was recently demonstrated by Mityushev that the field around a finite cluster of inclusions can yield a correct formula for the effective conductivity only for non-interacting clusters. The higher order term can be properly found only after a subtle study of the conditionally convergent series. The coefficients depend only on the parameter ϱ. The expression (5) is expressed as a correction to the celebrated classical Maxwell's, or Clausius-Mossotti formula (6). It is valid for small concentrations but respects the phase interchange symmetry [20], $[\sigma_e(\sigma)]^{-1} = \sigma_e\left(\sigma^{-1}\right)$.

We are interested in the case of highly conducting disks, with finite but large $\sigma \gg 1$. Let us formulate some starting approximation in the vicinity of f_c to satisfy some known critical behaviours.

In particular, as $\sigma \to \infty$, we have a singularity

$$\sigma_e \sim \frac{1}{\sqrt{1-\frac{f}{f_c}}}. \tag{7}$$

While for highly-conducting disks, as $f = f_c$, it is assumed that

$$\sigma_e \sim \sqrt{\sigma}, \tag{8}$$

following [9]. In order to perform calculations we start by choosing the starting approximation based on the two limit-cases (7) and (8),

$$\sigma_e(f,\sigma) \approx \left(\frac{2\sqrt{\frac{\pi}{4} - f}}{\sqrt{\pi}} + \frac{1}{\sqrt{\sigma}} \right)^{-1},$$

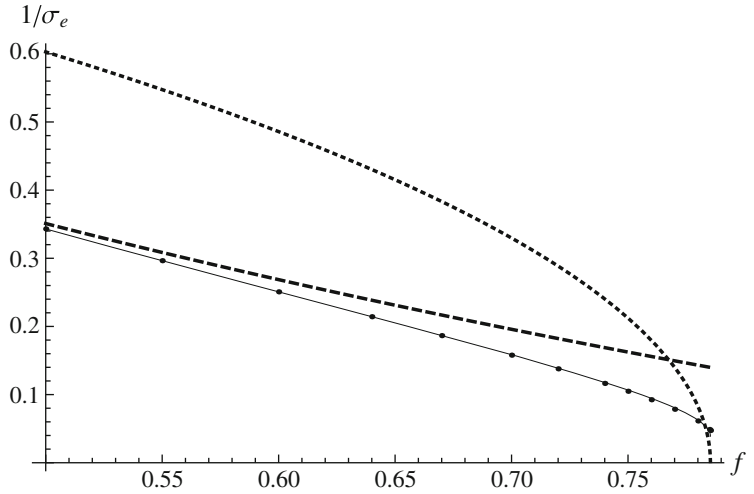


Fig. 2 Case of $\sigma = 50$. The results are shown for the resistance, inverse to the effective conductivity. Our suggestion (9) is shown with a solid line. The Clausius-Mossotti approximation (6) is shown with a dashed line. Naive extrapolation of the formula (7) to the whole region is shown with a dotted line, for comparison. Numerical data from [26] are shown with dots

which respects both cases of critical behaviour. Subsequently, it is going to be corrected by means of a diagonal Padé approximant, by achieving asymptotic equivalence with the expansion (5).

In the high orders orders one can obtain closed-form expressions. However, they are too long to be brought up here. But for concrete parameters their derivation and final form are pretty simple. Assuming the form $P_{4,4}$ for the correcting Padé approximant, we obtain an accurate formula for $\sigma = 50$, in excellent agreement with the numerical data of [26],

$$\sigma_e \approx \frac{f(f(f(1.22393f+0.277621)+1.23975)-2.08275)-3.25763}{(f(f(f(f-0.13819)-0.034015)+2.33313)-3.22041)\left(\sqrt{0.785398-f}+0.125331\right)}. \tag{9}$$

Various approximations are compared in Fig. 2. It is clear that correct qualitative incorporation of the critical regimes holds the key to accurate formula (9). In addition, formulas obtained in the same way as (9) work for $\sigma > 20$, while for $\sigma < 20$ plain Padé approximant becomes more accurate, signalling diminishing influence of the critical regime.

Already at finite but large $\sigma = 10^6$, the effective conductivity is well approximated by the following formula:

$$\sigma_e \approx \frac{f(f(f(1.06191f+0.165179)+1.05256)-1.81366)-2.70478}{(f(f(f(f-0.203144)-0.0950324)+2.11438)-3.04897)\left(\sqrt{0.785398-f}+0.000886227\right)}. \tag{10}$$

In fact such a system is very close to a perfect conductor, as shown in Fig. 3. We can also see that Keller's formula from [19], [16, Chapter 6], is asymptotic and doesn't

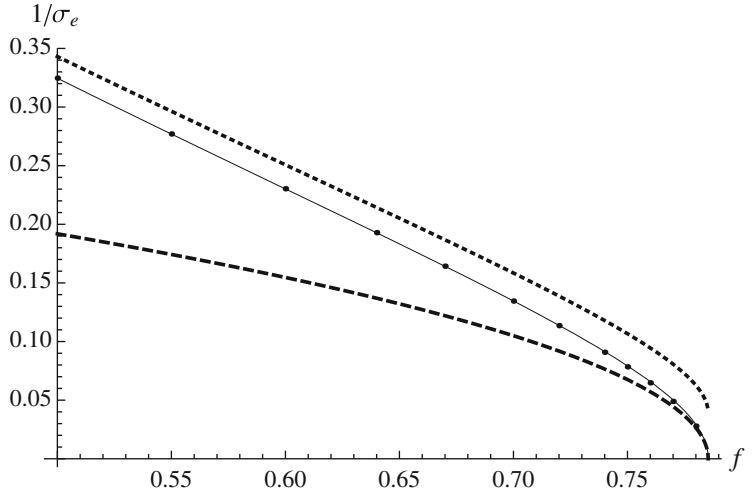


Fig. 3 Our suggestion (9) for $\sigma = 50$ is shown with dotted line, and similarly obtained results (10) for $\sigma = 10^6$ are shown with a solid line. Keller's formula from [19], [16, Chapter 6] is shown with a dashed line. Comparison is made with the numerical data from [26], as $\sigma \to \infty$

cover concentrations beyond an immediate vicinity of f_c. One can also conclude that silver (or copper) inclusions, embedded into a very weakly conducting matrix (vacuum, air, water, poly foam) can be considered as perfect conductors.

4 3D Superconductivity Critical Index of Random Composite

It was demonstrated in the book [6] that the classical Jeffrey formula for the effective conductivity of random composites [17] contains wrong f^2 terms. Proper expansion in volume fraction of inclusions f for the random composite with superconducting (perfectly conducting) inclusions were obtained in [6]. The terms f^2 and f^3 could be written explicitly. In particular, the f^3 term depends on the deterministic and random locations of inclusions. In the limiting case of a perfectly conducting inclusions, the effective conductivity σ_e is expected to tend to infinity as a power-law, as the concentration of inclusions f tends to f_c, the maximal value in 3D.

General methodology of [6] can be applied to the numerical estimation of the effective conductivity of random macroscopically isotropic composites. For samples generation the Random Sequential Adsorption (RSA) protocol was employed. The consecutive objects were placed randomly in the cell, rejecting those that overlap with previously absorbed one. For macroscopically isotropic composites the expansion for scalar effective conductivity takes the following form

$$\sigma_e = 1 + 3f + 3f^2 + 4.80654 f^3 + O(f^{\frac{10}{3}}). \tag{11}$$

It appears to be possible to extrapolate (11) to all f of interest, and calculate the critical index and amplitude from the truncated series. In addition, let us assume that the threshold is known and corresponds to random close packing (RCP), with the typical estimate $f_c = 0.637$ [28]. Then in the vicinity of RCP it is widely assumed that

$$\sigma_e \simeq A(f_c - f)^{-\mathsf{s}}, \tag{12}$$

where the superconductivity critical index s is expected to have the value of 0.73 ± 0.01 [5]. There is also a slightly larger estimate, $\mathsf{s} \approx 0.76$ [3].

Let us estimate the value of s based on asymptotic information encapsulated in (11). There is a possibility to obtain for σ_e, the simplest factor approximant [6, 16] with fixed position of singularity and floating critical index, from the requirement of asymptotic equivalence with (11). After some simple calculations we obtain the following result for the effective conductivity

$$\sigma^*(f) = \mathcal{F}_4^*(f) = \frac{(2.48123 f + 1)^{0.766996}}{(1 - 1.56986 f)^{0.698732}}. \tag{13}$$

The expression (13) suggests the value of 0.7 for the superconductivity critical index.

Assume that in the vicinity of threshold,

$$\sigma(f) \sim (0.637 - f)^{-(1+\mathsf{s}')},$$

with unity to be expected from the usual contribution from the radial distribution function at the particles contact $G(2, f)$, [4, 24], and the value of s' coming from the particle interactions in the composite. One can suggest a simple root approximant [6, 16]

$$r^*(z) = 1 + b_1 z (1 + b_2 z)^{\mathsf{s}'},$$

$z = \frac{f}{f_c - f}$, which is able to take the unity contribution into account explicitly. After imposing the asymptotic equivalence with the truncated series, one can find all three parameters b_1, b_2, s' with the final result for the effective conductivity

$$\sigma^*(f) = 1 + \frac{1.911 f}{\left(\frac{0.709259 f + 0.637}{0.637 - f}\right)^{0.212373} (0.637 - f)}. \tag{14}$$

The expression (14) allows us to estimate $\mathsf{s}' \approx -0.212$. Total value of the critical index, $\mathsf{s} \approx 0.788$ is still close enough to the expected values.

Let us employ the more systematic methodology used to construct the table of indices extracted from the root approximants. It is based on considering iterated root approximants [6, 16] as functions of the critical index by itself. The index can be found by imposing optimization conditions in the form of minimal differences

Table 1 Critical indices for the superconductivity sk obtained from the optimization conditions $\Delta_{kn}(\mathsf{s}_k) = 0$

s_k	$\Delta_{k,k+1}(\mathsf{s}_k) = 0$	$\Delta_{k3}(\mathsf{s}_k) = 0$
s_1	0.725	0.721
s_2	0.715	0.715

$\Delta_{kn}(s_k)$ imposed on critical amplitudes [15]. The series (11) allows to get only three estimates for the critical index s, see Table 1. All three results are fairly close to 0.72 and to the expected value of 0.73.

For possible applications, one can simply adjust the iterated root approximants to the most plausible value 0.73 for the critical exponent,

$$\begin{aligned} \mathcal{R}_2^*(f) &= \left(\frac{2.56706f^2+2.06109f+0.405769}{(0.637-f)^2}\right)^{0.365}, \\ \mathcal{R}_3^*(f) &= \left(\left(\frac{2.56706f^2+2.06109f+0.405769}{(0.637-f)^2}\right)^{3/2} - \frac{0.372504f^3}{(0.637-f)^3}\right)^{0.243333}. \end{aligned} \tag{15}$$

The two expressions are very close numerically, and the critical amplitude can be found from (15). From the former approximant it follows that $A \approx 1.449$, and from the latter approximant one obtains $A \approx 1.456$, giving practically the same result. The 4th order coefficient found from $\mathcal{R}_3^*(f)$, equals 7.48.

Thus, various techniques bring very close results, especially for the critical amplitude. For instance, one can simply extract the singularity first, and then apply the Padé technique, which brings the value of $A \approx 1.44$ close to other estimates, but the 4th order coefficient appears to be different and equal to 6.59.

The error in Jeffreys series estimate of the second order coefficient manifests itself in the value of the critical index. In terms of the variable z defined above, after standard calculations, we obtain the two estimates for the critical index, $\mathsf{s}_1 = 0.96$, $\mathsf{s}_2 = 1.12$, and $\mathsf{s} \approx 1.04 \pm 0.08$. The estimate is close to the effective medium result $\mathsf{s} = 1$.

5 Liquid Foams

Maxwell formula can still be useful for applications, in particular when some additional asymptotic information is available, It can be applied for various cases of highly conducting and non-conducting inclusions conditioned on percolating asymptotic behavior [1]. One of the successful applications is to the liquid foams used in their solidified form for electrical (thermal) insulation.

Because of the modest proportion of liquid in a foam and the large fraction of gas which has a much lower (thermal) conductivity the effective conductivity of the foam is much less than that of a liquid body made of the same material. The gas bubbles are pressed together to form the foam and are separated by thin films.

Where films meet, there is a liquid-filled interstitial channel called a Plateau border. In a real foam some liquid will collect within the edges at which the films meet. The amount of liquid available for these borders depends on the total amount of liquid left in the foam.

Consider first the two-dimensional foam. For a dispersion of 2D bubbles of a non-conducting gas in a continuous liquid phase of very high liquid fraction (very-wet regime) of $\epsilon \to 1$ ($f = 1 - \epsilon \to 0$), Maxwell's expression could be written as follows (see, for example, Andrianov et al [1]),

$$\sigma_e(\epsilon) = \frac{\epsilon}{2-\epsilon} \simeq 1 - 2(1-\epsilon) + 2(1-\epsilon)^2, \tag{16}$$

with $\sigma_e = \frac{\sigma_{sample}}{\sigma_{liquid}}$. At the other end of a very-dry regime, as $\epsilon \to 0$, the foam structure is two-dimensional polygonal, and in the entire condensed phase comprises a network of slender randomly oriented channels, or Plateau borders.

The conductivity of such a network as $\epsilon \to 0$, is given by a simple power-law

$$\sigma_e(\epsilon) = \frac{\epsilon}{2} + O(\sqrt{\epsilon}), \tag{17}$$

as explained in [7]. The leading correction term to the power-law (17) may be calculated more systematically. To this end let us construct and check all possible two-point Padé approximants of the type $\frac{\epsilon}{2} P_{n,m}(\sqrt{\epsilon})$. Just as later in the 3D case, we construct the Padé approximant $P_{3,2}(\sqrt{\epsilon})$, with the following result for the conductivity

$$\sigma_e(\epsilon) = \frac{\epsilon\left(\epsilon - 3\sqrt{\epsilon} + 5\right)}{10 - 7\sqrt{\epsilon}} \tag{18}$$

which behaves at small ϵ as

$$\sigma_e(\epsilon) \simeq \frac{\epsilon}{2} + \frac{\epsilon^{3/2}}{20}.$$

Remarkably, the conductivity formula corresponding to the approximant $P_{2,3}(\sqrt{\epsilon})$ appears to be identical with Maxwell's formula (16). And the numerical difference with another formula (18) is minuscule. We have here a quite unique case when the classic Maxwell theory and the neoclassical crossover formula interpolating between two regimes gives practically the same results for all volume fractions.

Consider the corresponding 3D case. For a dispersion of bubbles of a non-conducting gas in a continuous liquid phase of very high volume fraction (very-wet regime) $\epsilon \to 1$, Maxwell's expression could be adapted [17],

$$\sigma_e(\epsilon) = \frac{2\epsilon}{3-\epsilon} \simeq 1 - \frac{3(1-\epsilon)}{2} + \frac{3}{4}(1-\epsilon)^2. \tag{19}$$

Here ϵ stands for the volume fraction of the continuous liquid phase.

In the very-dry regime, as $\epsilon \to 0$, the foam structure is polyhedral, and the entire condensed phase comprises a network of slender randomly oriented channels (Plateau borders). The model network is one of straight borders of uniform cross-section, isotropic, i.e., uniformly distributed in orientation, and meeting at symmetric tetrahedral vertices [21, 27]. The conductivity in the limit of $\epsilon \to 0$ is given as follows,

$$\sigma_e(\epsilon) = \frac{\epsilon}{3} + O(\sqrt{\epsilon}), \tag{20}$$

as explained in [10, 21, 22, 27]. The leading term in (20) is shown to be an upper bound for the conductivity [8].

Between the two extremes the bubble shape varies from spherical to polyhedral as ϵ decreases. The electrical conductivity in the intermediate regime can be deduced from the two asymptotic expressions. In fact, the data in the wet and dry regimes match smoothly, and can be described by a simple empirical formulae [10], which also respects (19) and (20),

$$\sigma_e(\epsilon) = \frac{3.8\epsilon^{3/2} + \epsilon}{-2.8\epsilon + 4.6\sqrt{\epsilon} + 3}. \tag{21}$$

It expands at small ϵ as

$$\sigma_e(\epsilon) \simeq \frac{\epsilon}{3} + 0.76\epsilon^{3/2}.$$

The sign of the leading correction to the linear term appears to be positive, while another estimate based on formulae from [10, 27] gives negative sign.

The leading correction to the linear term may be calculated more systematically. To this end let us construct and check all possible two-point Padé approximants of the type $\frac{\epsilon}{3} P_{n,m}(\sqrt{\epsilon})$. It turns out that all of them give positive leading corrections to the linear term. The following approximant is closest to (21), with maximal percentage error around 2%,

$$\sigma_e(\epsilon) = \frac{\left(-4\epsilon + 3\sqrt{\epsilon} + 3\right)\epsilon}{9 - 7\sqrt{\epsilon}}, \tag{22}$$

and it behaves at small ϵ as

$$\sigma_e(\epsilon) \simeq \frac{\epsilon}{3} + \frac{16\epsilon^{3/2}}{27}.$$

Formula (22) is also in good agreement with some other fit from [10], which includes only integer powers of ϵ. Thus we constructed a good approximation to various experimental data relying only on asymptotic expressions (19) and (20), without resorting to fitting.

6 Permeability of a Symmetric Sinusoidal Three-Dimensional Channel

The effective quantity called permeability quantifies the amount of viscous fluid flow through a porous medium when a macroscopic pressure gradient is applied to the system. Precise definitions and general discussion of the critical properties of permeability in porous media, including flow in various channels can be found in Chapter 7 of the book [6] and in the paper [12].

Let us consider the three-dimensional channel restricted by the surfaces

$$z = \pm b\Big(1 + \frac{1}{2}\epsilon\big(\cos(x+y) + \cos(x-y)\big)\Big), \tag{23}$$

with $b = 0.3$ as formulated in the paper [25]. The permeability is found as the expansion in ϵ up to $O(\epsilon^{14})$

$$K_{14}(\epsilon) = 1 - 0.465674\epsilon^2 + 0.329218\epsilon^4 - 0.261666\epsilon^6 - 0.004467\epsilon^8 - 0.0386987\epsilon^{10} - 0.0177808\epsilon^{12} - 0.0239319\epsilon^{14}. \tag{24}$$

The case appears to be different from all two-dimensional examples studied in great detail in [6, 12]. For $\epsilon = \epsilon_c = 1$, the surfaces (23) start touching but the permeability remains finite at ϵ_c. The truncated series for permeability (24) is obtained with numerical precision of 10^{-3} for the values of ϵ up to 0.61. The permeability at ϵ_c remains quite significant, $K_{14}(\epsilon_c) = 0.517$, as is simply estimated from the series (24).

One can simply apply the technique of diagonal Padé approximants to the polynomial (24). The Padé approximants bring the following close results

$$P_{6,6}(\epsilon_c) = 0.51277,\ P_{8,8}(\epsilon_c) = 0.490636.$$

The higher order Padé approximants are readily obtained as well,

$$\begin{aligned} P_{6,6}(\epsilon) &= \frac{-0.272534\epsilon^6 + 0.22825\epsilon^4 - 0.657553\epsilon^2 + 1}{-0.0363255\epsilon^6 - 0.190321\epsilon^4 - 0.191879\epsilon^2 + 1}, \\ P_{8,8}(\epsilon) &= \frac{-0.266547\epsilon^8 - 0.131478\epsilon^6 - 0.363105\epsilon^4 + 0.256413\epsilon^2 + 1}{-0.0832011\epsilon^8 - 0.273346\epsilon^6 - 0.356065\epsilon^4 + 0.722087\epsilon^2 + 1}. \end{aligned} \tag{25}$$

One can deduce a reasonable bounds for the solution, such as the upper and lower Padé bounds [2]. They are are given by the non-diagonal Padé approximants [2],

$$\begin{aligned} P_{6,4}(\epsilon) &= \frac{-0.25985\epsilon^6+0.27733\epsilon^4-0.664548\epsilon^2+1}{-0.144498\epsilon^4-0.198874\epsilon^2+1}; \\ P_{6,8}(\epsilon) &= \frac{-0.354713\epsilon^6+0.280003\epsilon^4-0.721617\epsilon^2+1}{-0.0476736\epsilon^8-0.0872062\epsilon^6-0.168401\epsilon^4-0.255943\epsilon^2+1}. \end{aligned} \tag{26}$$

With such guidance we can construct and evaluate the two factor approximants [6, 16]. $\mathcal{F}^*_{12}(\epsilon)$, is standard, while the second, $\mathcal{F}^*_{12,s}(\epsilon)$, is "shifted". The shift also can be calculated and employed to estimate the sought value,

$$\begin{aligned} \mathcal{F}^*_{12}(\epsilon) &= \left(1-0.867964\epsilon^2\right)^{0.474676} \times \\ &\left(1+(0.0821614+0.533783i)\epsilon^2\right)^{1.35488+0.258822i} \times \\ &\left(1+(0.0821614-0.533783i)\epsilon^2\right)^{1.35488-0.258822i}; \\ \mathcal{F}^*_{12,s}(\epsilon) &= 0.481814+0.518186\left(1-\epsilon^2\right)^{0.766642} \times \\ &\left(1-(0.074165+0.649541i)\epsilon^2\right)^{1.46148+0.0652476i} \times \\ &\left(1-(0.074165-0.649541i)\epsilon^2\right)^{1.46148-0.0652476i}. \end{aligned} \tag{27}$$

Both approximants consume up to twelve-order terms from the expansion.

We are looking for the permeability at $\epsilon = 1$. Thus, one can obtain three estimates for the sought quantity,

$$P_{6,6}(1) = 0.51277, \quad F^*_{12}(1) = 0.50195, \quad F^*_{12,s}(1) = 0.481814,$$

all satisfying the expected bounds. Their average K_{av} is equal to 0.498845, and corresponding margin of error can be estimated through the variance, which equals 0.0128272. Different formulas for the permeability together with bounds, are compared in Fig. 4. Close to ϵ_c one finds

$$P_{6,6}(\epsilon) \simeq 0.51277 + 2.30175(1-\epsilon),$$

and the correction to constant is linear. As well, one can calculate from the shifted factor approximant, that

$$\mathcal{F}^*_{12,s}(\epsilon) \simeq 0.481814 + 1.36825(1-\epsilon)^{0.766642}.$$

Possibly, there is an indication here of a non-trivial subcritical index with the value of 0.767.

To elaborate further, we would like to study in more detail the behavior of permeability in the vicinity of ϵ_c. Let us also assume some deviations from linearity, motivated by the shifted factor approximant.

Let us start with general-type initial approximation for the permeability, which holds in the vicinity of $\epsilon_c = 1$,

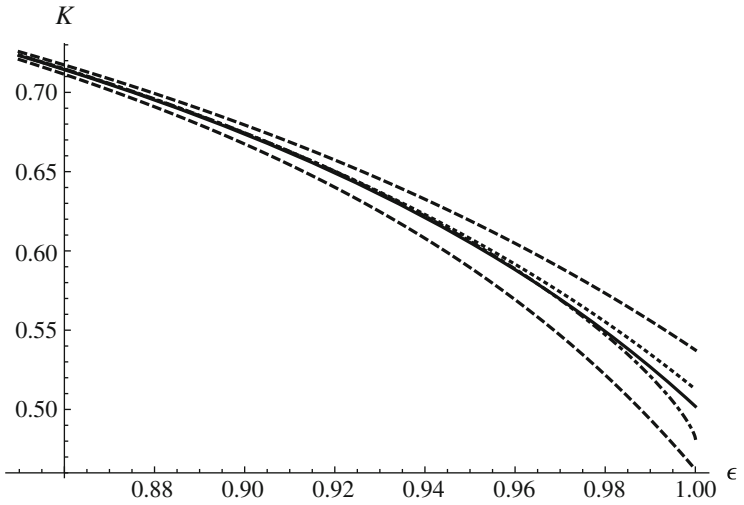


Fig. 4 Bounds (26) for the permeability are shown with dashed lines. Comparison of the formulas in the vicinity of ϵ_c: Padé approximant $P_{6,6}$ is shown with with dotted line, factor approximant $\mathcal{F}^*_{12}$ from (27) is shown with solid line, and shifted factor approximant $\mathcal{F}^*_{12,s}$ from (27) is shown with dot-dashed line

$$K_0(\epsilon) \simeq A_0 + A_1(\epsilon_c^2 - \epsilon^2)^{\lambda_0}, \tag{28}$$

To simplify the procedure of finding the unknowns, let us set from the start $A_0 = K_{av}$. Now, to obtain the remaining unknowns, one can try to satisfy the expansion (24) in the second order. Then, it appears that $A_1 = 0.501155$, $\lambda_0 = 0.929201$. The expression (28) can be understood as a two-fluid model, reflecting on the fact that there are two components in the flow. One which is getting blocked by the obstacles to flow, and another, which can not be blocked.

One ought to appreciate that (28) with its parameters is only a crude approximation. In what follows let us attempt to correct the formula $K_0(\epsilon)$, even further. To this end let us assume in place of λ_0, some more general functional dependence $\Lambda^*(\epsilon)$. As $\epsilon \to \epsilon_c$, $\Lambda^*(\epsilon) \to \lambda_c$, the sought corrected value. The function $\Lambda^*(\epsilon)$ will be designed in such a way, that it smoothly interpolates between the initial value λ_0 valid at small ϵ, and the sought value λ_c valid as $\epsilon \to \epsilon_c$. The permeability $K^*(\epsilon)$ is getting “dressed" in such a way. It is now given as follows:

$$K^*(\epsilon) = A_0 + A_1(\epsilon_c^2 - \epsilon^2)^{\Lambda^*(\epsilon)}. \tag{29}$$

It is valid now for all ϵ. From (29) one can express $\Lambda^*(\epsilon)$ formally, bearing in mind that we do not have the expression for $K^*(\epsilon)$. All we can do is to use its asymptotic form (24), then express $\Lambda^*(\epsilon)$ as a truncated series for small ϵ. And then we can apply to such obtained series some resummation procedure (e.g. Padé technique). Such resummation is expected to extend the series to the whole region of ϵ. Finally,

we are in a position to calculate the limit of the approximants as $\epsilon \to \epsilon_c$, and find the corrected value as $\lambda_c = \Lambda^*(\epsilon_c)$.

Let $p(\epsilon) = K_{14}(\epsilon)$ stand for an asymptotic form of $K^*(\epsilon)$ for small ϵ. Corresponding asymptotic expression for Λ^*, just called $\Lambda(\epsilon)$, can be made explicit from the following relation,

$$\Lambda(\epsilon) \simeq -\frac{\log\left(\frac{(A_0 - p(\epsilon))}{A_1}\right)}{\log(\epsilon_c^2 - \epsilon^2)}. \tag{30}$$

$\Lambda(\epsilon)$ can be presented as expansion in powers of ϵ around the value of λ_0,

$$\Lambda(\epsilon) = \lambda_0 + \Lambda_1(\epsilon). \tag{31}$$

And only now one can construct a sequence of diagonal Padé approximants [2]

$$\Lambda_n(\epsilon) = \lambda_0 + PadeApproximant[\Lambda_1[\epsilon], n, n]), \tag{32}$$

and find the sought limit $\Lambda^*(\epsilon)$. Finally, we estimate the critical index $\lambda_c = \Lambda^*(\epsilon_c)$ and also find the complete formula for permeability, returning to the expression (29).

There is a good convergence within the approximations for the λ_c generated by the sequence of Padé approximants,

$$\lambda_{c,1} = 0.929201, \quad \lambda_{c,2} = 0.402904, \quad \lambda_{c,4} = 0.631631,$$

$$\lambda_{c,6} = 0.630229, \quad \lambda_{c,8} = 0.702766, \quad \lambda_{c,10} = 0.698385 \quad \lambda_{c,12} = 0.702563.$$

Remarkably, in the highest orders the value of index remains practically the same. The final estimate for λ_c can be conjectured to be rational $\frac{2}{3}$.

The function $\Lambda^*(\epsilon)$ is needed to reconstruct the permeability. It can be straightforwardly expressed as the Padé approximant. The approximant corresponding to $\lambda_{c,6}$ has the following form,

$$K_6^*(\epsilon) = 0.498845 + 0.501155\left(1 - \epsilon^2\right)^{\frac{-4.15886\epsilon^6 + 6.957\epsilon^4 - 7.18244\epsilon^2 + 0.929201}{-1.56421\epsilon^6 + 2.06925\epsilon^4 - 6.98732\epsilon^2 + 1}}. \tag{33}$$

Formula (33), as well as the higher-order approximant (34), corresponding to $\lambda_{c,8}$,

$$K_8^*(\epsilon) = 0.498845 + 0.501155\left(1 - \epsilon^2\right)^{\frac{0.134578\epsilon^8 - 0.22113\epsilon^6 + 0.650924\epsilon^4 - 0.904689\epsilon^2 + 0.929201}{-0.0295078\epsilon^8 - 0.199491\epsilon^6 + 0.298201\epsilon^4 - 0.23125\epsilon^2 + 1}}, \tag{34}$$

are confidently located within the Padé-bounds (26). Formulas for the permeability including the subcritical regime, are shown together with the bounds in Fig. 5.

We conclude that in various physical problems it is both possible and also quite handy to possess a general mathematical toolbox to derive asymptotic, typically

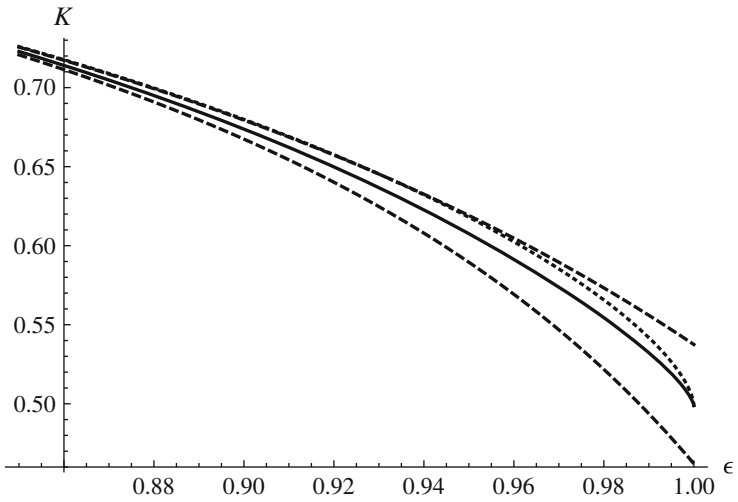


Fig. 5 Bounds (26) for the permeability are shown with dashed lines. Comparison of the formulas in the vicinity of ϵ_c: $K_6^*(\epsilon)$ is shown with dotted line, $K_8^*(\epsilon)$ is shown with solid line

power laws, as well as to obtain explicit crossover formulas. The former, extrapolation problems are more difficult to solve than the latter, interpolation problems. Various transformations of the original truncated expressions were suggested to enhance existing methods, based on Padé or self-similar approximants. Power transformation of the original series was developed in [13, 14]. Method of Borel summation was further developed recently and applied to direct calculation of critical indices at infinity [30]. Many cases of interpolation and extrapolation are presented in [6, 11, 12, 16].

References

1. Andrianov, I.V., Danishevskyy, V.V., Kalamkarov, A.L.: Analysis of the effective conductivity of composite materials in the entire range of volume fractions of inclusions up to the percolation threshold. Compos. Part B Eng. **41**, 503–507 (2010)
2. Baker, G.A., Jr., Graves-Moris, P.: Padé Approximants. Cambridge University, Cambridge (1996)
3. Bergman, D.J., Stroud, D.: Physical properties of macroscopically inhomogeneous media. Solid State Phys. **46**, 148–270 (1992)
4. Brady, J.F.: The rheological behavior of concentrated colloidal dispersions. J. Chem. Phys. **99**, 567–581 (1993)
5. Clerc, J.P., Giraud, G., Laugie, J.M., Luck, J.M.: The electrical conductivity of binary disordered systems, percolation clusters, fractals and related models. Adv. Phys. **39**, 191–309 (1990)
6. Drygaś, P., Gluzman, S., Mityushev, V., Nawalaniec, W.: Applied Analysis of Composite Media. Elsevier (Woodhead), Cambridge (2020)

7. Durand, M.: Low-density cellular materials with optimal conductivity and bulk modulus. In: CongrÃĺs FranÃğais de MÃl'canique, pp. 1–6. Grenoble (2007). https://doi.org/10.4267/2042/15750
8. Durand, M., Sadoc, J.F., Weaire, D.: Maximum electrical conductivity of a network of uniform wires: the Lemlich law as an upper bound. Proc. R. Soc. Lond. A **460**, 1269–1285 (2004)
9. Efros, A.L., Shklovskii, B.I.: Critical behaviour of conductivity and dielectric constant near the metal-non-metal transition threshold. Phys. Status Solidi B. **76**, 475–485 (1976)
10. Feitosa, M., Marze, S., Saint-Jalmes, A., Durian, D.J.: Electrical conductivity of dispersions: from dry foams to dilute suspensions. J. Phys. Condensed Matter **17**, 6301–6305 (2005)
11. Gluzman, S.: Padé and post-padé approximations for critical phenomena. Symmetry **12**(10), 1600 (2020)
12. Gluzman, S.: Nonlinear approximations to critical and relaxation processes. Axioms **9**(4), 126 (2020)
13. Gluzman, S.: Critical indices and self-similar power transform. Axioms **10**(3), 162 (2021)
14. Gluzman, S.: Continued roots, power transform and critical properties. Symmetry **13**, 1525 (2021)
15. Gluzman, S., Yukalov, V.I.: Critical indices from self-similar root approximants. Eur. Phys. J. Plus **132**, 535 (2017)
16. Gluzman, S., Mityushev, V., Nawalaniec, W.: Computational analysis of structured media. Elsevier (Academic Press), Cambridge (2018)
17. Jeffrey, D.J.: Conduction through a random suspension of spheres. Proc. R. Soc. Lond. A **335**, 355–367 (1973)
18. Koiter, W.T.: Solution of some elasticity problems by asymptotic methods. In: Proc. IUTAM Symposium on Applications of the Theory of Functions in Continuum Mechanics (Tbilisi 1963), vol. 1, pp. 15–31 Nauka Publ. House, Moscow (1965)
19. Keller, J.B.: Conductivity of a medium containing a dense array of perfectly conducting spheres or cylinders or nonconducting cylinders. J. Appl. Phys. **34**, 991–993 (1963)
20. Keller, J.B.: A theorem on the conductivity of a composite medium. J. Math. Phys. **5**, 548–549 (1964)
21. Lemlich, R.: Theory for limiting conductivity of polyhedral foam at low-density. J. Colloid Interface Sci. **64**, 107–110 (1978)
22. Lemlich, R.: Semi-theoretical equation to relate conductivity to volumetric foam density. Ind. Eng. Chem. Process Des. Dev. **24**, 686–687 (1985)
23. Lieb, E.H., Liniger, S.: Exact analysis of an interacting Bose gas: the general solution and the ground state. Phys. Rev. **13**, 1605–1616 (1963)
24. Losert, W., Bocquet, L., Lubensky, T.C., Gollub, J.P.: Particle dynamics in sheared granular matter. Phys. Rev. Lett. **85**, 1428–1431 (2000)
25. Malevich, A.E., Mityushev, V.V., Adler, P.M.: Stokes flow through a channel with wavy walls. Acta Mech. **182**, 151–182 (2006)
26. Perrins, W.T., McKenzie, D.R., McPhedran, R.C.: Transport properties of regular array of cylinders. Proc. R. Soc. A **1369**, 207–225 (1979)
27. Phelan, R., Weaire, D., Peters, E.A.J.F., Verbist, G.: The conductivity of a foam. J. Phys. Condensed Matter **8**, L475–L482 (1996)
28. Torquato, S.: Random Heterogeneous Materials: Microstructure and Macroscopic Properties. Springer-Verlag, New York (2002)
29. Yukalov, V.I., Girardeau, M.D.: Fermi-Bose mapping for one-dimensional Bose gases. Laser Phys. Lett. **2**, 375–382 (2005)
30. Yukalov, V.I., Gluzman, S.: Methods of retrieving large-variable exponents. Symmetry **14**, 332 (2022)
31. Yukalov, V.I., Yukalova, E.P., Gluzman, S.: Extrapolation and interpolation of asymptotic series by self-similar approximants. J. Math. Chem. **47**, 959–983 (2010)

Analogues the Kolosov-Muskhelishvili Formulas for Isotropic Materials with Double Voids

Bakur Gulua

Abstract Analogues of the well-known Kolosov-Muskhelishvili formulas for homogeneous equations of statics in the case of elastic materials with double voids are obtained. It is shown that in this theory the displacement and stress vector components are represented by two analytic functions of a complex variable and two solutions of Helmholtz equations. The constructed general solution enables one to solve analytically a sufficiently wide class of plane boundary value problems of the elastic equilibrium with double voids.

1 Introduction

Theories of porous media are applied in many branches of engineering, technology, geomechanics and biomechanics. The theory of elasticity of porous materials with voids (empty pores) essentially differs from the classical theory of elasticity in that the volume fraction, which corresponds to the volume of empty pores, is meant as an independent variable.

The nonlinear theory of porous materials with voids is described in Nunziato and Cowin [1], and the linear theory of porous materials with voids is presented by Cowin and Nunziato [2]. By using the mechanics of materials with voids the theories of elasticity and thermoelasticity for materials with double-porosity structure are presented by Ieşan and Quintanilla [3]. The basic equations of this theory involve the displacement vector field and the volume fraction fields associated with the pores and the fissures. Monograph of Ieşan [4] gives the main objectives of thermoelasticity for various models, which contain empty pores, as well as basic results obtained in the Cowin–Nunziato theory and felds of their application. The book by Straughan [5] gives the major results for the porous bodies of different

B. Gulua (✉)
Sokhumi State University, Tbilisi, Georgia

I. Vekua Institute of Applied Mathematics, I. Javakhishvili Tbilisi State University, Tbilisi, Georgia

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_28

models and bibliographical data. Some results of the 2D and 3D theories of elasticity for materials with microstructures can be seen in [6–10].

The present paper deals with plane strain problem for linear elastic materials with double voids. In the spirit of N.I. Muskhelishvili the governing system of equations of the plane strain is rewritten in the complex form and its general solution is represented by means of two analytic functions of the complex variable and two solutions of Helmholtz equations. The constructed general solution enables us to solve analytically the problems for a circle.

2 Basic Equations for Materials with Double Voids of the 3D Model

Let $x = (x_1; x_2; x_3)$ be a point of the Euclidean three dimensional space R^3. We assume that the subscripts preceded by a comma denote partial differentiation with respect to the corresponding Cartesian coordinate, repeated indices are summed over the range (1; 2; 3).

In what follows we consider an isotropic and homogeneous elastic solid with double voids occupying a region of $\Omega \in R^3$. The governing equations of the theory of elastic materials with double voids can be expressed in the following form [3]:

- Equations of equilibrium

$$\begin{aligned} &t_{ji,j} + \rho_0 f_i = 0, \quad i, j = 1, 2, 3, \\ &\sigma_{j,j} + \xi + \rho_0 g = 0, \\ &\tau_{j,j} + \zeta + \rho_0 l = 0, \end{aligned} \tag{1}$$

  where t_{ij} is the symmetric stress tensor, f_i is the body force per unit mass, ρ_0 is the mass density, σ_i and τ_i are the equilibrated stress vectors, ξ and ζ are the intrinsic equilibrated body forces, g is the extrinsic equilibrated body force per unit mass associated to macro pores, l is the extrinsic equilibrated body force per unit mass associated to fissures.
- Constitutive equations

$$\begin{aligned} t_{ij} &= \lambda e_{kk}\delta_{ij} + 2\mu e_{ij} + b\delta_{ij}\varphi + d\delta_{ij}\psi, \\ \sigma_i &= \alpha\varphi_{,i}, + b_1\psi_{,i}, \\ \tau_i &= b_1\varphi_{,i}, + \gamma\psi_{,i}, \\ \xi &= -be_{kk} - \alpha_1\varphi - \alpha_3\psi, \\ \zeta &= -de_{kk} - \alpha_3\varphi - \alpha_2\psi, \end{aligned} \tag{2}$$

  where λ and μ are the Lamé constants, $\alpha, b, d, b_1, \alpha_1, \alpha_2$ and α_3 are the constants characterizing the body porosity, δ_{ij} is the Kronecker delta, φ is a changes of volume fraction corresponding to pores, ψ is a a changes of volume fraction

corresponding to fissures, e_{ij} is the strain tensor and

$$e_{ij} = \frac{1}{2}\left(u_{i,j} + u_{j,i}\right), \tag{3}$$

where u_i, $i = 1, 2, 3$ are the components of the displacement vector.

The constitutive equations also meet some other conditions, following from physical considerations

$$\begin{array}{l} \mu > 0, \quad 3\lambda + 2\mu > 0, \quad \alpha_2 > 0, \quad \alpha_1\alpha_2 - \alpha_3^2 > 0, \\ (3\lambda + 2\mu)(\alpha_1\alpha_2 - \alpha_3^2) > 3(\alpha_1 d^2 + \alpha_2 b^2 - 2\alpha_3 bd), \quad \alpha > 0, \quad \alpha\gamma > b_1^2. \end{array} \tag{4}$$

Substituting (2) and (3) into (1) we obtain equations with respect to the components of the displacement and the functions φ and ψ

$$\begin{array}{l} \mu\tilde{\Delta}u_i + (\lambda + \mu)\partial_i\Theta + b\partial_i\varphi + d\partial_i\psi = 0, \quad j = 1, 2, 3 \\ (\alpha\tilde{\Delta} - \alpha_1)\varphi + (b_1\tilde{\Delta} - \alpha_3)\psi - b\Theta = 0, \\ (b_1\tilde{\Delta} - \alpha_3)\varphi + (\gamma\tilde{\Delta} - \alpha_2)\psi - d\Theta = 0, \end{array}$$

where $\partial_i \equiv \frac{\partial}{\partial x_i}$, $\Theta = \partial_k u_k$, $\Delta \equiv \partial_{11} + \partial_{22} + \partial_{33}$ is the three-dimensional Laplace operator.

3 Basic (Governing) Equations of the Plane Strain

From the basic three-dimensional equations we obtain the basic equations for the case of plane strain. Let Ω be a sufficiently long cylindrical body with generatrix parallel to the Ox_3-axis. Denote by V the crosssection of this cylindrical body, thus $V \subset R^2$. In the case of plane deformation $u_3 = 0$ while the functions u_1, u_2, φ and ψ do not depend on the coordinate x_3.

As it follows from formulas (2) and (3), in the case of plane strain

$$t_{k3} = t_{3k} = 0, \quad \sigma_3 = 0, \quad \tau_3 = 0, \quad k = 1, 2.$$

Assuming $\Phi_i \equiv 0$ and $\Psi \equiv 0$. Therefore the system of equilibrium Eq. (1) takes the form

$$\begin{array}{l} \partial_1 t_{11} + \partial_2 t_{21} = 0, \\ \partial_1 t_{12} + \partial_2 t_{22} = 0, \\ \partial_k \sigma_k + \xi = 0, \\ \partial_k \tau_k + \zeta = 0. \end{array} \tag{5}$$

Now, Relations (2) are rewritten as

$$\begin{aligned}
&t_{11} = \lambda\theta + 2\mu\partial_1 u_1 + b\varphi + d\psi,\\
&t_{22} = \lambda\theta + 2\mu\partial_2 u_2 + b\varphi + d\psi,\\
&t_{12} = t_{21} = \mu(\partial_1 u_2 + \partial_2 u_1),\\
&t_{33} = \sigma(t_{11} + t_{22}),\\
&\sigma_k = \alpha\partial_k\varphi + b_1\partial_k\psi, \quad k = 1, 2,\\
&\tau_k = b_1\partial_k\varphi + \gamma\partial_k\psi, \quad k = 1, 2,\\
&\xi = -b\theta - \alpha_1\varphi - \alpha_3\psi,\\
&\zeta = -d\theta - \alpha_3\varphi - \alpha_2\psi,
\end{aligned} \tag{6}$$

where σ is the Poisson ratio, $\theta = \partial_1 u_1 + \partial_2 u_2$.

If relations (6) are substituted into system (5) then we obtain the following system of governing equations of statics with respect to the functions u_1, u_2 and φ, ψ

$$\begin{aligned}
&\mu\Delta u_k + (\lambda + \mu)\partial_k\theta + b\partial_k\varphi + d\partial_k\psi = 0, \quad k = 1, 2\\
&(\alpha\Delta - \alpha_1)\varphi + (b_1\Delta - \alpha_3)\psi - b\theta = 0,\\
&(b_1\Delta - \alpha_3)\varphi + (\gamma\Delta - \alpha_2)\psi - d\theta = 0,
\end{aligned} \tag{7}$$

Note that $\Delta \equiv \partial_{11} + \partial_{22}$ is the two-dimensional Laplace operator.

On the plane Ox_1x_2, we introduce the complex variable $z = x_1 + ix_2 = re^{i\vartheta}$, $(i^2 = -1)$ and the operators $\partial_z = 0.5(\partial_1 - i\partial_2)$, $\partial_{\bar{z}} = 0.5(\partial_1 + i\partial_2)$, $\bar{z} = x_1 - ix_2$, and $\Delta = 4\partial_z\partial_{\bar{z}}$.

To write system (5) in the complex form, the second equation of this system we multiplied by i and sum up with the first equation

$$\begin{aligned}
&\partial_z(t_{11} - t_{22} + 2it_{12}) + \partial_{\bar{z}}(t_{11} + t_{22}) = 0,\\
&\partial_z\sigma_+ + \partial_{\bar{z}}\bar{\sigma}_+ + \xi = 0,\\
&\partial_z\tau_+ + \partial_{\bar{z}}\bar{\tau}_+ + \zeta = 0,
\end{aligned} \tag{8}$$

where $\sigma_+ = \sigma_1 + i\sigma_2$, $\tau_+ = \tau_1 + i\tau_2$ and formulas (6) we rewrite as follows

$$\begin{aligned}
&t_{11} - t_{22} + 2it_{12} = 4\mu\partial_{\bar{z}}u_+,\\
&t_{11} + t_{22} = 2(\lambda + \mu)\theta + 2b\varphi + 2d\psi,\\
&\sigma_+ = 2\alpha\partial_{\bar{z}}\varphi + 2b_1\partial_{\bar{z}}\psi,\\
&\tau_+ = 2b_1\partial_{\bar{z}}\varphi + 2\gamma\partial_{\bar{z}}\psi,\\
&\xi = -b\theta - \alpha_1\varphi - \alpha_3\psi,\\
&\zeta = -d\theta - \alpha_3\varphi - \alpha_2\psi,\\
&\theta = \partial_z u_+ + \partial_{\bar{z}}\bar{u}_+, \quad u_+ = u_1 + iu_2.
\end{aligned} \tag{9}$$

Substituting relations (9) into system (8), we rewrite system (7) in the complex form

$$\begin{aligned} &2\mu\partial_{\bar z}\partial_z u_+ + (\lambda+\mu)\partial_{\bar z}\theta + b\partial_{\bar z}\varphi + d\partial_{\bar z}\psi = 0,\\ &(\alpha\Delta-\alpha_1)\varphi + (b_1\Delta-\alpha_3)\psi - b\theta = 0,\\ &(b_1\Delta-\alpha_3)\varphi + (\gamma\Delta-\alpha_2)\psi - d\theta = 0. \end{aligned} \tag{10}$$

4 Kolosov-Muskhelishvili Formulas for (10) System

Now we construct the analogues to the Kolosov-Muskhelishvili formulas for system (10) [11–13].

We take the operator $\partial_{\bar z}$ out of the brackets in the left-hand part of the first equation of system (10)

$$\partial_{\bar z}(2\mu\partial_z u_+ + (\lambda+\mu)\theta + b\varphi + d\psi) = 0. \tag{11}$$

Since (11) is a system of Cauchy-Riemann equations, we have

$$2\mu\partial_z u_+ + (\lambda+\mu)\theta + b\varphi + d\psi = Af'(z), \tag{12}$$

where $f(z)$ is an arbitrary analytic function of z and A an arbitrary constant.

A conjugate equation to (12) has the form

$$2\mu\partial_{\bar z}\bar u_+ + (\lambda+\mu)\theta + b\varphi + d\psi = A\overline{f'(z)}, \tag{13}$$

Summing up Eqs. (12) and (13) and taking into account that

$$\theta = \partial_z u_+ + \partial_{\bar z}\bar u_+$$

we obtain

$$\theta = \frac{A}{2(\lambda+\mu)}(f'(z)+\overline{f'(z)}) - \frac{b}{\lambda+2\mu}\varphi - \frac{d}{\lambda+2\mu}\psi. \tag{14}$$

Substituting formula (14) into the second and third equations of system (10), we have

$$\begin{aligned} &\left(\alpha\Delta-\alpha_1+\tfrac{b^2}{\lambda+2\mu}\right)\varphi + \left(b_1\Delta-\alpha_3+\tfrac{bd}{\lambda+2\mu}\right)\psi = \tfrac{Ab}{2(\lambda+\mu)}(f'(z)+\overline{f'(z)}),\\ &\left(b_1\Delta-\alpha_3+\tfrac{bd}{\lambda+2\mu}\right)\varphi + \left(\gamma\Delta-\alpha_2+\tfrac{d^2}{\lambda+2\mu}\right)\psi = \tfrac{Ad}{2(\lambda+\mu)}(f'(z)+\overline{f'(z)}). \end{aligned} \tag{15}$$

Equation (15) system rewrite in matrix form

$$\Delta\Psi - C\Psi = DF, \tag{16}$$

where

$$C = \begin{pmatrix} \alpha & b_1 \\ b_1 & \gamma \end{pmatrix}^{-1} \cdot \begin{pmatrix} \alpha_1 - \frac{b^2}{\lambda+2\mu} & \alpha_3 - \frac{bd}{\lambda+2\mu} \\ \alpha_3 - \frac{bd}{\lambda+2\mu} & \alpha_2 - \frac{d^2}{\lambda+2\mu} \end{pmatrix},$$

$$D = \begin{pmatrix} \alpha & b_1 \\ b_1 & \gamma \end{pmatrix}^{-1} \cdot \begin{pmatrix} \frac{Ab}{2(\lambda+2\mu)} & 0 \\ 0 & \frac{Ad}{2(\lambda+2\mu)} \end{pmatrix},$$

$$\Psi = \begin{pmatrix} \varphi \\ \psi \end{pmatrix}, \quad F = \begin{pmatrix} f'(z) + \overline{f'(z)} \\ f'(z) + \overline{f'(z)} \end{pmatrix}.$$

The general solutions of system (15) we may write in the form

$$\begin{aligned} \varphi &= l_{11}\chi_1(z,\bar{z}) + l_{12}\chi_2(z,\bar{z}) - AE_1(f'(z) + \overline{f'(z)}), \\ \psi &= l_{21}\chi_1(z,\bar{z}) + l_{22}\chi_2(z,\bar{z}) - AE_2(f'(z) + \overline{f'(z)}), \end{aligned} \tag{17}$$

where $\chi_1(z,\bar{z})$ and $\chi_2(z,\bar{z})$ are a general solutions of the Helmholtz equations

$$\Delta\chi(z,\bar{z}) - \kappa_1\chi(z,\bar{z}) = 0, \quad \Delta\chi(z,\bar{z}) - \kappa_1\chi(z,\bar{z}) = 0.$$

where κ_α are eigenvalues and $(l_{11},\ l_{21})$, $(l_{12},\ l_{22})$ are eigenvectors of the matrix C and from (4) its are positive numbers,

$$E_1 = \frac{b\alpha_2 - d\alpha_3}{2((\alpha_1\alpha_2 - \alpha_3^2)(\lambda+2\mu) - \alpha_1 d^2 - \alpha_2 b^2 + 2\alpha_3 bd)},$$

$$E_2 = \frac{d\alpha_1 - b\alpha_3}{2((\alpha_1\alpha_2 - \alpha_3^2)(\lambda+2\mu) - \alpha_1 d^2 - \alpha_2 b^2 + 2\alpha_3 bd)}.$$

Substituting formulas (14) and (17) into Eq. (12), we obtain

$$2\mu\partial_z u_+ = \frac{\lambda+3\mu}{2(\lambda+2\mu)}(1+bE_1+dE_2)Af'(z) - \frac{\lambda+\mu-(\lambda+3\mu)(bE_1+dE_2)}{2(\lambda+2\mu)}\overline{f'(z)}$$

$$- \frac{(bl_{11}+dl_{21})(\lambda+3\mu)}{2(\lambda+2\mu)}\chi_1(z,\bar{z}) - \frac{(bl_{12}+dl_{22})(\lambda+3\mu)}{2(\lambda+2\mu)}\chi_2(z,\bar{z}).$$

Now, let $A := \frac{2(\lambda+2\mu)}{\lambda+\mu-(\lambda+3\mu)(bE_1+dE_2))}$ than we get

$$2\mu u_+ = \kappa f(z) - z\overline{f'(z)} - \overline{g(z)} - p_1\partial_{\bar{z}}\chi_1(z,\bar{z}) - p_2\partial_{\bar{z}}\chi_2(z,\bar{z}),$$

where $\kappa = \frac{(\lambda+3\mu)(1+bE_1+dE_2)}{\lambda+\mu-(\lambda+3\mu)(bE_1+dE_2)}$, $p_1 = \frac{2(bl_{11}+dl_{21})(\lambda+3\mu)}{\kappa_1(\lambda+2\mu)}$, $p_2 = \frac{2(bl_{12}+dl_{22})(\lambda+3\mu)}{\kappa_2(\lambda+2\mu)}$, $g(z)$ is an arbitrary analytic function of z.

Thus, we have proved

Theorem 1 *The general solution of the system (10) is represented as follows:*

$$\begin{aligned}
2\mu u_+ &= \kappa f(z) - z\overline{f'(z)} - \overline{g(z)} - p_1\partial_{\bar{z}}\chi_1(z,\bar{z}) - p_2\partial_{\bar{z}}\chi_2(z,\bar{z}),\\
\varphi &= l_{11}\chi_1(z,\bar{z}) + l_{12}\chi_2(z,\bar{z}) - \overline{E}_1(f'(z) + \overline{f'(z)}),\\
\psi &= l_{21}\chi_1(z,\bar{z}) + l_{22}\chi_2(z,\bar{z}) - \overline{E}_2(f'(z) + \overline{f'(z)}).
\end{aligned} \tag{18}$$

where $\overline{E}_1 = AE_1$, $\overline{E}_2 = AE_2$.

From (9) we have

$$\begin{aligned}
&t_{11} - t_{22} + 2it_{12} = -2z\overline{f''(z)} - 2\overline{g'(z)} - p_1\partial_{\bar{z}}\partial_{\bar{z}}\chi_1(z,\bar{z}) - p_2\partial_{\bar{z}}\partial_{\bar{z}}\chi_2(z,\bar{z}),\\
&t_{11} + t_{22} = \frac{A(\lambda+\mu) - 2b\mu\overline{E}_1 - 2d\mu\overline{E}_2}{\lambda+2\mu}\left(f'(z) + \overline{f'(z)}\right)\\
&+\frac{2\mu(bl_{11}+dl_{21})}{\lambda+2\mu}\chi_1(z,\bar{z}) + \frac{2\mu(bl_{12}+dl_{22})}{\lambda+2\mu}\chi_2(z,\bar{z}),\\
&\sigma_+ = 2(dl_{11}+b_1l_{21})\partial_{\bar{z}}\chi_1(z,\bar{z}) + 2(dl_{12}+b_1l_{22})\partial_{\bar{z}}\chi_2(z,\bar{z})\\
&\quad -2(d\overline{E}_1 + b_1\overline{E}_2)\overline{f''(z)},\\
&\tau_+ = 2(b_1l_{11}+\gamma l_{21})\partial_{\bar{z}}\chi_1(z,\bar{z}) + 2(b_1l_{12}+\gamma l_{22})\partial_{\bar{z}}\chi_2(z,\bar{z})\\
&\quad -2(b_1\overline{E}_1 + \gamma\overline{E}_2)\overline{f''(z)},\\
&\xi = \left(\frac{b^2l_{11}+bdl_{21}}{\lambda+2\mu} - \alpha_1l_{11} - \alpha_3l_{21}\right)\chi_1(z,\bar{z})\\
&+\left(\frac{b^2l_{12}+bdl_{22}}{\lambda+2\mu} - \alpha_1l_{12} - \alpha_3l_{22}\right)\chi_2(z,\bar{z})\\
&+\left(\alpha_1\overline{E}_1 + \alpha_3\overline{E}_2 - \frac{Ab + 2b^2\overline{E}_1 + 2bd\overline{E}_2}{2(\lambda+2\mu)}\right)(f'(z) + \overline{f'(z)}),
\end{aligned}$$

$$\begin{aligned}
\zeta &= \left(\frac{bdl_{11}+d^2l_{21}}{\lambda+2\mu} - \alpha_3l_{11} - \alpha_2l_{21}\right)\chi_1(z,\bar{z})\\
&+\left(\frac{bdl_{12}+d^2l_{22}}{\lambda+2\mu} - \alpha_3l_{12} - \alpha_2l_{22}\right)\chi_2(z,\bar{z})\\
&+\left(\alpha_3\overline{E}_1 + \alpha_2\overline{E}_2 - \frac{Ad + 2bd\overline{E}_1 + 2d^2\overline{E}_2}{2(\lambda+2\mu)}\right)(f'(z) + \overline{f'(z)}).
\end{aligned}$$

5 A Problem for a Circle

Let the origin of coordinates be at the centre of the circle with radius R Fig. 1. On the boundary of the considered domain the values of φ, ψ and the displacement vector are given. Analogous problems of plane elasticity for materials with single voids are considered in [13–15].

We consider the following problem

$$\begin{aligned} 2\mu u_+|_{r=R} &= 2\mu(G_1 + iG_2) = \sum_{-\infty}^{+\infty} A_n e^{in\vartheta}, \\ \varphi|_{r=R} &= G_3 = \sum_{-\infty}^{+\infty} B_n e^{in\vartheta}, \\ \psi|_{r=R} &= G_4 = \sum_{-\infty}^{+\infty} C_n e^{in\vartheta}. \end{aligned} \tag{19}$$

The analytic functions $f(z)$, $g(z)$ and the metaharmonic functions $\chi_1(z, \bar{z})$ and $\chi_2(z, \bar{z})$ are represented as the series

$$f(z) = \sum_{n=1}^{\infty} a_n z^n, \quad g(z) = \sum_{n=0}^{\infty} b_n z^n,$$

$$\chi_1(z, \bar{z}) = \sum_{-\infty}^{+\infty} \alpha_n I_n(\sqrt{\kappa_1} r) e^{in\vartheta}, \quad \chi_2(z, \bar{z}) = \sum_{-\infty}^{+\infty} \beta_n I_n(\sqrt{\kappa_2} r) e^{in\vartheta}$$

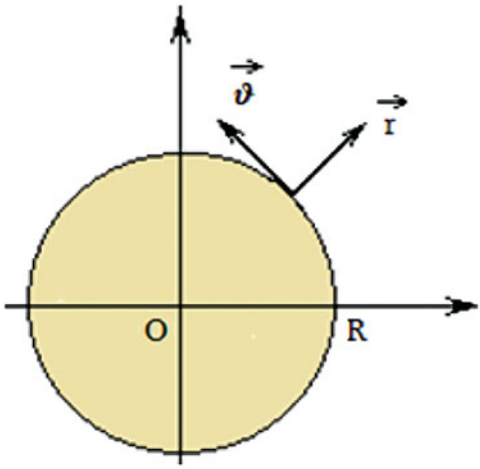


Fig. 1 The circle with radius R

where $I_n(\sqrt{\kappa_1}r)$ and $I_n(\sqrt{\kappa_2}r)$ are the modified Bessel function of the first kind of n-th order. After substituting into the boundary conditions (19) we have

$$\kappa\sum_{n=1}^{\infty}a_nR^ne^{in\vartheta}-\bar{a}_1Re^{i\vartheta}-\sum_{n=0}^{\infty}(n+2)\bar{a}_{n+2}R^{n+2}e^{-in\vartheta}-\sum_{n=0}^{\infty}\bar{b}_nR^ne^{-in\vartheta}$$

$$-\frac{p_1\sqrt{\kappa_1}}{2}\sum_{-\infty}^{+\infty}\alpha_nI_{n+1}(\sqrt{\kappa_1}R)e^{i(n+1)\vartheta}$$

$$-\frac{p_2\sqrt{\kappa_2}}{2}\sum_{-\infty}^{+\infty}\alpha_nI_{n+1}(\sqrt{\kappa_2}R)e^{i(n+1)\vartheta}=\sum_{-\infty}^{+\infty}A_ne^{in\vartheta},$$

$$l_{11}\sum_{-\infty}^{+\infty}\alpha_nI_n(\sqrt{\kappa_1}R)e^{in\vartheta}+l_{12}\sum_{-\infty}^{+\infty}\beta_nI_n(\sqrt{\kappa_2}R)e^{in\vartheta}$$

$$-\overline{E}_1\sum_{n=1}^{\infty}\left(na_nR^{n-1}e^{i(n-1)\vartheta}+n\bar{a}_nR^{n-1}e^{-i(n-1)\vartheta}\right)=\sum_{-\infty}^{+\infty}B_ne^{in\vartheta},$$

$$l_{21}\sum_{-\infty}^{+\infty}\alpha_nI_n(\sqrt{\kappa_1}R)e^{in\vartheta}+l_{22}\sum_{-\infty}^{+\infty}\beta_nI_n(\sqrt{\kappa_2}R)e^{in\vartheta}$$

$$-\overline{E}_2\sum_{n=1}^{\infty}\left(na_nR^{n-1}e^{i(n-1)\vartheta}+n\bar{a}_nR^{n-1}e^{-i(n-1)\vartheta}\right)=\sum_{-\infty}^{+\infty}C_ne^{in\vartheta}.$$

Comparing the coefficients of members with equal degrees, we obtain the following systems of equation

$$\begin{aligned}
&\kappa Ra_1-R\bar{a}_1-\frac{p_1\sqrt{\kappa_1}}{2}I_1(\sqrt{\kappa_1}R)\alpha_0-\frac{p_2\sqrt{\kappa_2}}{2}I_1(\sqrt{\kappa_2}R)\beta_0=A_1,\\
&\kappa R^na_n-\frac{p_1\sqrt{\kappa_1}}{2}I_n(\sqrt{\kappa_1}R)\alpha_{n-1}-\frac{p_2\sqrt{\kappa_2}}{2}I_n(\sqrt{\kappa_2}R)\beta_{n-1}=A_n,\quad n>1,\\
&-(n+2)R^{n+2}\bar{a}_{n+2}-R^n\bar{b}_n-\frac{p_1\sqrt{\kappa_1}}{2}I_n(\sqrt{\kappa_1}R)\alpha_{-n-1}\\
&-\frac{p_2\sqrt{\kappa_2}}{2}I_n(\sqrt{\kappa_2}R)\beta_{-n-1}=A_{-n},\quad n\geq 0,
\end{aligned}$$

$$\begin{aligned}
l_{11}I_0(\sqrt{\kappa_1}R)\alpha_0+l_{12}I_0(\sqrt{\kappa_2}R)\beta_0-\overline{E}_1(a_1+\bar{a}_1)=B_0,\\
l_{21}I_0(\sqrt{\kappa_1}R)\alpha_0+l_{22}I_0(\sqrt{\kappa_2}R)\beta_0-\overline{E}_2(a_1+\bar{a}_1)=C_0,
\end{aligned}$$

$$\begin{aligned}
&l_{11}I_n(\sqrt{\kappa_1}R)\alpha_n+l_{12}I_n(\sqrt{\kappa_2}R)\beta_n-\overline{E}_1(n+1)R^na_{n+1}=B_n,\quad n>0,\\
&l_{21}I_n(\sqrt{\kappa_1}R)\alpha_n+l_{22}I_n(\sqrt{\kappa_2}R)\beta_n-\overline{E}_2(n+1)R^na_{n+1}=C_n,\quad n>0.
\end{aligned}$$

All coefficients are determined by these equations.

It is easily seen the absolute and uniform convergence of the series obtained in the circle (including the contours) when the functions prescribed on the boundary are sufficiently smooth.

The procedure of solving a boundary value problem remains the same when stresses and change in volume fractions on the domain boundary are given, but the condition that the principal vector and the principal moment of external forces are equal to zero is fulfilled.

References

1. Nunziato, G.W., Cowin, S.C.: A nonlinear theory of elastic materials with voids. Arch Ration. Mech. Anal. **72**, 175–201 (1979)
2. Cowin, S.C., Nunziato, G.W.: Linear theory of elastic materials with voids. J. Elast. **13**, 125–147 (1983)
3. Ieşan, D., Quintanilla, R.: On a theory of thermoelastic materials with a double porosity structure. J. Therm. Stress. **37**, 1017–1036 (2014)
4. Ieşan, D.: Thermoelastic Models of Continua. Kluwer, Boston (2004) https://doi.org/10.1007/978-1-4020-2310-1
5. Straughan, B.: Mathematical Aspects of Multi-Porosity Continua. Advances in Mechanics and Mathematics, vol 38. Springer, Berlin (2017)
6. Svanadze, M.: Steady vibration problems in the theory of elasticity for materials with double voids. Acta Mech. **229**, 1517–1536 (2017). https://doi.org/10.1007/s00707-017-2077-z
7. Bitsadze, L.: Explicit solutions of boundary value problems of elasticity for circle with a double-voids structure. J. Braz. Soc. Mech. Sci. Eng. **41**, 383 (2019). https://doi.org/10.1007/s40430-019-1888-3
8. Tsagareli, I.: Explicit solution of elastostatic boundary value problems for the elastic circle with voids. Adv. Math. Phys. (2018). https://doi.org/10.1155/2018/6275432
9. Tsagareli, I.: Solution of boundary value problems of thermoelasticity for a porous disk with voids. J. Porous Media **23**(2) 177–185 (2020)
10. Janjgava, R., Gulua, B., Tsotniashvili, S.: Some boundary value problems for a micropolar porous elastic body. Arch. Mech. **72**(6), 485–509 (2020)
11. Muskhelishvili, N.I.: Some basic problems of the mathematical theory of elasticity. Fundamental Equations, Plane Theory of Elasticity, Torsion and Bending (Russian), 5th revised and enlarged edn. Nauka, Moscow (1966)
12. Gulua, B., Janjgava, R.: On construction of general solutions of equations of elastostatic problems for the elastic bodies with voids. PAMM **18**, 1 (2018). https://doi.org/10.1002/pamm.201800306
13. Gulua, B., Janjgava, R.: Boundary value problems of the plane theory of elasticity for materials with voids. In: International Conference on Applications of Mathematics and Informatics in Natural Sciences and Engineering, pp. 227–236. Springer (2019)
14. Gulua, B.: Basic boundary value problems for circular ring with voids. Trans. A Razmadze Math. Inst. **175**, 3, 437–441 (2021)
15. Gulua B., Kasrashvili T.: Some basic problems of the plane theory of elasticity for materials with voids. Seminar of I. Vekua Institute of Applied Mathematics, REPORTS, vol. 46, pp. 27–36 (2020)

Schwarz-Christoffel Mapping and Generalised Modulus of a Quadrilateral

Giorgi Kakulashvili

Abstract In this work we analyze Schwarz-Christoffel (SC) transformation for polygonal quadrilaterals and by its geometric properties introduce generalised modulus. The function holds many interesting properties and allows to algorithmically solve parameter problem for quadrilaterals.

1 Introduction

Parameters problem for Schwarz-Christoffel (SC) mapping and computation of modulus of quadrilaterals are the most famous and old problems in geometric theory of complex analysis. In last decade, investigation of these problems had led to many deep and interesting works from different point of view [1–3]. In [4–6] is given an overview of the modern theoretical and numerical achievements in this field.

A conformal modulus of a generalised quadrilateral [1, 6] is a non-negative real number which divides quadrilaterals into conformal equivalence classes. For a rectangle conformal modulus is ratio of its neighbors sides or its proportion. Therefore, to find conformal modulus of polygonal quadrilateral, we need to map it conformally to rectangle and for this we use SC mapping.

2 Schwarz-Christoffel Mapping for Quadrilaterals

The classical Schwarz-Christoffel formula allows us to conformally map half-plane onto domains whose boundaries consist of a finite number of line segments. In our considerations we use the following version of the SC-theorem.

Theorem 1 *[7] Let P be the interior of a polygon Γ with clockwise ordered vertices $w_1, \ldots, w_n$ in the complex plane and with external angles $\pi\beta_1, \ldots, \pi\beta_n$.*

G. Kakulashvili (✉)
Ivane Javakhishvili, Tbilisi State University, Tbilisi, Georgia

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_29

Let f be a conformal map from the lower half-plane $\mathbb{H}_-$ to P with $f(\infty) = w_n$. Then

$$f(x) = A + C \int_{-\infty}^{x} \prod_{j=1}^{n-1} \left(1 - \frac{\zeta}{z_j}\right)^{-\beta_j} d\zeta, \tag{1}$$

for some complex constants A and C, where $w_k = f(z_k)$ and the preimages of vertices z_j, for $j = 1, \ldots, n-1$ satisfy the conditions $1 = z_1 < z_2 < \cdots < z_{n-1} < z_n = +\infty$.

Consider particular case of this theorem. Namely, let Q be a simple quadrilateral with inner angles $\pi\tau_j$ enumerated in clockwise order. Suppose

$$z_1 = 1 \quad z_2 = 1 + \theta \quad z_3 = 1 + \theta + r\theta,$$

where $\theta, r > 0$ parametrization of preimages of vertices of quadrilateral Q, i.e., the inequality

$$1 = z_1 < z_2 < z_3$$

is satisfied without loss of generality. Then the Schwarz-Christoffel transformation is given by

$$f(x) = A + C \int_{-\infty}^{x} (1-\zeta)^{\tau_1 - 1} \left(1 - \frac{\zeta}{1+\theta}\right)^{\tau_2 - 1} \left(1 - \frac{\zeta}{1+\theta+r\theta}\right)^{\tau_3 - 1} d\zeta. \tag{2}$$

If $A = 0$ and $C = 1$, then f function from expression 2 maps lower half plane onto clockwise oriented quadrilateral, with $w_4 = 0$ and $w_1 > 0$ vertices.

Orientation of the quadrilateral depends how we define argument of a complex number. For example, we use

$$\arg\left(1 - \frac{\zeta}{z}\right) = \begin{cases} 0, & \zeta < z, \\ \pi, & \zeta > z \end{cases} \quad \text{and} \quad \arg\left(1 - \frac{\zeta}{z}\right)^{-\beta} = \begin{cases} 0, & \zeta < z, \\ -\pi\beta, & \zeta > z \end{cases}$$

therefore

$$\arg\left(\prod_{k=1}^{3} \left(1 - \frac{\zeta}{z_k}\right)^{-\beta_k}\right) = \begin{cases} 0, & \zeta < z_1, \\ -\pi\beta_1, & z_1 < \zeta < z_2, \\ -\pi(\beta_1 + \beta_2), & z_2 < \zeta < z_3, \\ -\pi(\beta_1 + \beta_2 + \beta_3), & \zeta > z_3. \end{cases}$$

Theorem 2 *Let Q be a simple quadrilateral with inner angles $\pi\tau_1, \pi\tau_2, \pi\tau_3$ such that*

$$f(x) = \int_{-\infty}^{x} (1-\zeta)^{\tau_1-1} \left(1 - \frac{\zeta}{1+\theta}\right)^{\tau_2-1} \left(1 - \frac{\zeta}{1+\theta+r\theta}\right)^{\tau_3-1} d\zeta$$

maps the lower half-plane onto Q. Then the side lengths of Q are

$$\begin{aligned}
l_1 &= cB(\tau_4, \tau_1)\,{}_2F_1\,(\tau_4, 1-\tau_3; \tau_4+\tau_1; -r)\,,\\
l_2 &= r^{\tau_3-1} cB(\tau_1, \tau_2)\,{}_2F_1\left(\tau_2, 1-\tau_3; \tau_1+\tau_2; -\frac{1}{r}\right),\\
l_3 &= r^{\tau_2+\tau_3-1} cB(\tau_2, \tau_3)\,{}_2F_1\,(\tau_2, 1-\tau_1; \tau_2+\tau_3; -r)\,,\\
l_4 &= r^{-\tau_4} cB(\tau_3, \tau_4)\,{}_2F_1\left(\tau_4, 1-\tau_1; \tau_3+\tau_4; -\frac{1}{r}\right)
\end{aligned}$$

where

$$c = \theta^{-\tau_4}(1+\theta)^{1-\tau_2}(1+\theta+r\theta)^{1-\tau_3}$$

and the ratio of the adjacent sides of a quadrilateral are independent of θ.

For proof of this theorem we use the following proposition.

Proposition 1 *We have:*

1. *The following equality*

$$\int_0^1 u^{a-1}(1-u)^{b-1}(1-zu)^{c-1}du = B(a,b)\,{}_2F_1\,(a, 1-c; a+b; z) \tag{3}$$

holds for $z < 1$, when $a, b > 0$, $c \in \mathbb{R}$, and for $z = 1$, when $a > 0$ and $b+c > 0$.

2. *If $r > 0$ and $k = \sqrt{1/(1+r)}$ then we have the following equalities*

$$\begin{aligned}
&\int_0^{+\infty} u^{a-1}\,(1+u)^{b-1}\,(1+r+u)^{c-1}\,du\\
&\quad = B(2-a-b-c, a)\,{}_2F_1\,(2-a-b-c, 1-c; 2-b-c; -r)
\end{aligned} \tag{4}$$

and

$$\int_0^{+\infty} u^{a-1}(1+u)^{b-1}\left(1+k^2u\right)^{c-1}du = B(2-a-b-c,a)\,{}_2F_1\left(a,1-c;2-b-c;1-k^2\right), \tag{5}$$

when $a+b+c<2$, $a>0$, *and* $c\in\mathbb{R}$.

Proof The equality (3) can be obtained by Euler's integral representation

$$B(a,b)\,{}_2F_1(a,1-c;a+b;z) = B(a,(a+b)-a)\,{}_2F_1(1-c,a;a+b;z) = \int_0^1 u^{a-1}(1-u)^{b-1}(1-zu)^{c-1}du,$$

when $a+b>a>0$.

To prove second part of the proposition we use the change of variable $u=t-1$ and $t=1/v$

$$\int_0^{+\infty} u^{a-1}(1+u)^{b-1}(1+r+u)^{c-1}du = \int_0^1 v^{1-a-b-c}(1-v)^{a-1}(1+rv)^{c-1}dv$$

$$= B(2-a-b-c,a)\,{}_2F_1(2-a-b-c,1-c;2-b-c;-r).$$

If we multiply both sides by $(1+r)^{1-c}$ and apply Pfaff's transform we obtain

$$\int_0^{+\infty} u^{a-1}(1+u)^{b-1}\left(1+\frac{u}{1+r}\right)^{c-1}du = B(2-a-b-c,a)\,{}_2F_1\left(a,1-c;2-b-c;\frac{r}{1+r}\right).$$

The Proposition is proved. □

Now we can prove Theorem 2.

Proof By changing the variable of integration $\zeta = 1 + u\theta$ we get

$$
\begin{aligned}
l_2 &= |f(1+\theta) - f(1)| \\
&= \left| \int_1^{1+\theta} (1-\zeta)^{\tau_1 - 1} \left(1 - \frac{\zeta}{1+\theta}\right)^{\tau_2 - 1} \left(1 - \frac{\zeta}{1+\theta+r\theta}\right)^{\tau_3 - 1} d\zeta \right| \\
&= \int_1^{1+\theta} (\zeta - 1)^{\tau_1 - 1} \left(1 - \frac{\zeta}{1+\theta}\right)^{\tau_2 - 1} \left(1 - \frac{\zeta}{1+\theta+r\theta}\right)^{\tau_3 - 1} d\zeta du \\
&= \frac{(1+r)^{\tau_3 - 1} \theta^{-\tau_4}}{(1+\theta)^{\tau_2 - 1} (1+\theta+r\theta)^{\tau_3 - 1}} \int_0^1 u^{\tau_1 - 1} (1-u)^{\tau_2 - 1} \left(1 - \frac{u}{1+r}\right)^{\tau_3 - 1}.
\end{aligned}
$$

Below, for computing l_3 and l_4 we use change of variables $\zeta = 1 + \theta + ur\theta$ and $\zeta = 1 + \theta + vr\theta$, respectively:

$$
\begin{aligned}
l_3 &= |f(1+\theta+r\theta) - f(1+\theta)| \\
&= \int_{1+\theta}^{1+\theta+r\theta} (\zeta - 1)^{\tau_1 - 1} \left(\frac{\zeta}{1+\theta} - 1\right)^{\tau_2 - 1} \left(1 - \frac{\zeta}{1+\theta+r\theta}\right)^{\tau_3 - 1} d\zeta \\
&= \frac{r^{\tau_2 + \tau_3 - 1} \theta^{-\tau_4}}{(1+\theta)^{\tau_2 - 1} (1+\theta+r\theta)^{\tau_3 - 1}} \int_0^1 u^{\tau_2 - 1} (1-u)^{\tau_3 - 1} (1+ru)^{\tau_1 - 1} du
\end{aligned}
$$

$$
\begin{aligned}
l_4 &= |f(+\infty) - f(1+\theta+r\theta)| \\
&= \int_{1+\theta+r\theta}^{+\infty} (\zeta - 1)^{\tau_1 - 1} \left(\frac{\zeta}{1+\theta} - 1\right)^{\tau_2 - 1} \left(\frac{\zeta}{1+\theta+r\theta} - 1\right)^{\tau_3 - 1} d\zeta \\
&= \frac{r^{\tau_2 + \tau_3 - 1} \theta^{-\tau_4}}{(1+\theta)^{\tau_2 - 1} (1+\theta+r\theta)^{\tau_3 - 1}} \int_1^{+\infty} v^{\tau_2 - 1} (v-1)^{\tau_3 - 1} (1+vr)^{\tau_1 - 1} dv \\
&= \frac{r^{-\tau_4} \theta^{-\tau_4}}{(1+\theta)^{\tau_2 - 1} (1+\theta+r\theta)^{\tau_3 - 1}} \int_0^1 u^{\tau_4 - 1} (1-u)^{\tau_3 - 1} \left(1 + \frac{u}{r}\right)^{\tau_1 - 1} du.
\end{aligned}
$$

By (3) from Proposition 1 we have

$$l_2 = (1+r)^{\tau_3-1} cB(\tau_1,\tau_2)\,{}_2F_1\left(\tau_1, 1-\tau_3; \tau_1+\tau_2; \frac{1}{1+r}\right),$$

$$l_3 = r^{\tau_2+\tau_3-1} cB(\tau_2,\tau_3)\,{}_2F_1\left(\tau_2, 1-\tau_1; \tau_2+\tau_3; -r\right),$$

$$l_4 = r^{-\tau_4} cB(\tau_3,\tau_4)\,{}_2F_1\left(\tau_4, 1-\tau_1; \tau_3+\tau_4; -\frac{1}{r}\right),$$

where

$$c = \theta^{-\tau_4}(1+\theta)^{1-\tau_2}(1+\theta+r\theta)^{1-\tau_3}.$$

If we use the Pfaff transformation for expression of l_2 we obtain

$$l_2 = r^{\tau_3-1} cB(\tau_1,\tau_2)\,{}_2F_1\left(\tau_2, 1-\tau_3; \tau_1+\tau_2; -\frac{1}{r}\right).$$

For l_1 we use formula (4) from Proposition 1 and the change of variables $t = 1-\zeta$:

$$\begin{aligned} l_1 &= \int_{-\infty}^{1} (1-\zeta)^{\tau_1-1}\left(1-\frac{\zeta}{1+\theta}\right)^{\tau_2-1}\left(1-\frac{\zeta}{1+\theta+r\theta}\right)^{\tau_3-1} d\zeta \\ &= \int_{0}^{\infty} t^{\tau_1-1}\left(1-\frac{1-t}{1+\theta}\right)^{\tau_2-1}\left(1-\frac{1-t}{1+\theta+r\theta}\right)^{\tau_3-1} dt \\ &= \frac{\theta^{(\tau_1-1)+(\tau_2-1)+(\tau_3-1)}}{(1+\theta)^{\tau_2-1}(1+\theta+r\theta)^{\tau_3-1}}\times \\ &\times \int_{0}^{\infty}\left(\frac{t}{\theta}\right)^{\tau_1-1}\left(1+\frac{t}{\theta}\right)^{\tau_2-1}\left(1+r+\frac{t}{\theta}\right)^{\tau_3-1} dt. \end{aligned}$$

Change of variables $t = \theta v$ and $w = 1/u$ gives

$$\begin{aligned} &\frac{\theta^{(\tau_1-1)+(\tau_2-1)+(\tau_3-1)+1}}{(1+\theta)^{\tau_2-1}(1+\theta+r\theta)^{\tau_3-1}} \int_{0}^{\infty} v^{\tau_1-1}(1+v)^{\tau_2-1}(1+r+v)^{\tau_3-1} dv \\ &= \frac{\theta^{-\tau_4}}{(1+\theta)^{\tau_2-1}(1+\theta+r\theta)^{\tau_3-1}} \int_{1}^{\infty} w^{\tau_2-1}(w-1)^{\tau_1-1}(r+w)^{\tau_3-1} dw \\ &= \frac{\theta^{-\tau_4}}{(1+\theta)^{\tau_2-1}(1+\theta+r\theta)^{\tau_3-1}} \int_{0}^{1} u^{\tau_4-1}(1-u)^{\tau_1-1}(1+ru)^{\tau_3-1} du. \end{aligned}$$

The first part of the theorem is proved.

The proof of the second statement directly follows from calculation. Indeed,

$$\frac{l_1}{l_2} = r^{1-\tau_3}\frac{B(\tau_4,\tau_1)}{B(\tau_1,\tau_2)}\frac{{}_2F_1\left(\tau_4, 1-\tau_3; \tau_4+\tau_1; -r\right)}{{}_2F_1\left(\tau_2, 1-\tau_3; \tau_1+\tau_2; -1/r\right)},$$
$$\frac{l_2}{l_3} = r^{-\tau_2}\frac{B(\tau_1,\tau_2)}{B(\tau_2,\tau_3)}\frac{{}_2F_1\left(\tau_2, 1-\tau_3; \tau_1+\tau_2; -1/r\right)}{{}_2F_1\left(\tau_2, 1-\tau_1; \tau_2+\tau_3; -r\right)},$$
$$\frac{l_3}{l_4} = r^{1-\tau_1}\frac{B(\tau_2,\tau_3)}{B(\tau_3,\tau_4)}\frac{{}_2F_1\left(\tau_2, 1-\tau_1; \tau_2+\tau_3; -r\right)}{{}_2F_1\left(\tau_4, 1-\tau_1; \tau_3+\tau_4; -1/r\right)},$$
$$\frac{l_4}{l_1} = r^{-\tau_4}\frac{B(\tau_3,\tau_4)}{B(\tau_4,\tau_1)}\frac{{}_2F_1\left(\tau_4, 1-\tau_1; \tau_3+\tau_4; -1/r\right)}{{}_2F_1\left(\tau_4, 1-\tau_3; \tau_4+\tau_1; -r\right)}.$$

The theorem is proved. □

The SC-mapping f from Theorem 2 defines quadrilateral Q with $w_4 = 0$ and $w_1 > 0$ vertices and the positive number r from the expression we call r-invariant of the quadrilateral Q.

Remark 1 Since translation, rotation, and uniform scaling are conformal transformations, by adding and multipling by A, C complex numbers any arbitrary quadrilateral can be transformed, so that it vertices are mapped like $w_4 = 0$ and $w_1 > 0$.

3 Generalised Modulus of Quadrilaterals

In this section we define generalised modulus φ of quadrilateral, which holds important properties on how conformal modulus changes when quadrilateral propositions are changed.

Definition 1 For given

$$\tau_1, \tau_2, \tau_3 > 0, \quad \tau_1+\tau_2+\tau_3 < 2, \quad \tau_4 = 2-(\tau_1+\tau_2+\tau_3)$$

we define the function φ as

$$\varphi(x;\tau_1,\tau_2,\tau_3) = x^{1-\tau_1}\frac{B(\tau_2,\tau_3)}{B(\tau_3,\tau_4)}\frac{{}_2F_1\left(\tau_2, 1-\tau_1; \tau_2+\tau_3; -x\right)}{{}_2F_1\left(\tau_4, 1-\tau_1; \tau_3+\tau_4; -1/x\right)} \tag{6}$$

for $x > 0$ and call it the generalised modulus of quadrilateral with inner angles $\pi\tau_j$, $j = 1, 2, 3, 4$.

Theorem 3 *Let Q be a simple quadrilateral with side lengths l_j and inner angles $\pi\tau_j$, $j = 1, 2, 3, 4$. Then*

$$\varphi(r; \tau_1, \tau_2, \tau_3) = \frac{l_3}{l_4}, \qquad \varphi\left(\frac{1}{r}; \tau_2, \tau_3, \tau_4\right) = \frac{l_4}{l_1},$$

$$\varphi(r; \tau_3, \tau_4, \tau_1) = \frac{l_1}{l_2}, \qquad \varphi\left(\frac{1}{r}; \tau_4, \tau_1, \tau_2\right) = \frac{l_2}{l_3}.$$

where r is r-invariant of Q.

Proof By Remark 1 we can map Q without changing its properties and apply Theorem 2. We prove only the last identity, the other ones are obtained by similar way. By the Euler's transformations we obtain

$$\begin{aligned}\varphi\left(\frac{1}{r}; \tau_4, \tau_1, \tau_2\right) &= r^{\tau_4-1}\frac{B(\tau_1, \tau_2)}{B(\tau_2, \tau_3)}\frac{{}_2F_1(\tau_1, 1-\tau_4; \tau_1+\tau_2; -1/r)}{{}_2F_1(\tau_3, 1-\tau_4; \tau_2+\tau_3; -r)}\\ &= r^{\tau_4-1}\frac{B(\tau_1, \tau_2)}{B(\tau_2, \tau_3)}\\ &\quad\times\frac{(1+1/r)^{\tau_2+\tau_4-1}{}_2F_1(\tau_2, 1-\tau_3; \tau_1+\tau_2; -1/r)}{(1+r)^{\tau_2+\tau_4-1}{}_2F_1(\tau_2, 1-\tau_1; \tau_2+\tau_3; -r)}\\ &= r^{-\tau_2}\frac{B(\tau_1, \tau_2)}{B(\tau_2, \tau_3)}\frac{{}_2F_1(\tau_2, 1-\tau_3; \tau_1+\tau_2; -1/r)}{{}_2F_1(\tau_2, 1-\tau_1; \tau_2+\tau_3; -r)} = \frac{l_2}{l_3}.\end{aligned}$$

The theorem is proved. □

Corollary 1 *Suppose $\tau_1 = \tau_2 = \tau_3 = 1/2$, then*

$$\varphi\left(x; \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right) = \frac{\mathcal{K}'\left(\sqrt{1/(1+x)}\right)}{\mathcal{K}\left(\sqrt{1/(1+x)}\right)}, \tag{7}$$

where $\mathcal{K}(x)$ is complete elliptic integral and $\mathcal{K}'(x) = \mathcal{K}\left(\sqrt{1-x^2}\right)$.

Proof The proof of (7) we obtain from the Pfaff's transformation and from the chain of following identities:

$${}_2F_1\left(\frac{1}{2}, \frac{1}{2}; 1; \frac{1}{1+x}\right) = \left(1+\frac{1}{x}\right)^{1/2}{}_2F_1\left(\frac{1}{2}, \frac{1}{2}; 1; -\frac{1}{x}\right),$$

$$\mathcal{K}\left(\sqrt{\frac{1}{1+x}}\right) = \frac{\pi}{2}{}_2F_1\left(\frac{1}{2}, \frac{1}{2}; 1; \frac{1}{1+x}\right) = \frac{\pi}{2}\left(1+\frac{1}{x}\right)^{1/2}{}_2F_1\left(\frac{1}{2}, \frac{1}{2}; 1; -\frac{1}{x}\right),$$

$$\mathcal{K}'\left(\sqrt{\frac{1}{1+x}}\right) = \mathcal{K}\left(\sqrt{\frac{x}{1+x}}\right) = \frac{\pi}{2}{}_2F_1\left(\frac{1}{2},\frac{1}{2};1;\frac{x}{1+x}\right)$$
$$= \frac{\pi}{2}{}_2F_1\left(\frac{1}{2},\frac{1}{2};1;\frac{1}{1+1/x}\right) = \frac{\pi}{2}(1+x)^{1/2}{}_2F_1\left(\frac{1}{2},\frac{1}{2};1;-x\right).$$

Finally,

$$\frac{\mathcal{K}'\left(\sqrt{1/(1+x)}\right)}{\mathcal{K}\left(\sqrt{1/(1+x)}\right)} = \sqrt{x}\frac{{}_2F_1\left(1/2,1/2;1;-x\right)}{{}_2F_1\left(1/2,1/2;1;-1/x\right)}.$$

and corollary is proved. □

From Theorem 3 we find that for given Q quadrilateral there exists only one $r > 0$ variable and by formula (7) we get connection with MQ conformal modulus

$$MQ = \frac{\mathcal{K}'\left(\sqrt{1/(1+r)}\right)}{\mathcal{K}\left(\sqrt{1/(1+r)}\right)}.$$

Theorem 4 *Generalised modulus φ for given inner angles $\pi\tau_i$ of simple quadrilateral is increasing when $\tau_1 < 1$ and decreasing $\tau_1 > 1$.*

Proof By using Pfaff transform we get the following equality

$$\varphi(r;\tau_1,\tau_2,\tau_3) = \frac{B(\tau_2,\tau_3)}{B(\tau_3,\tau_4)}\frac{{}_2F_1\left(1-\tau_1,\tau_3;\tau_2+\tau_3;1-k^2\right)}{{}_2F_1\left(1-\tau_1,\tau_3;\tau_3+\tau_4;k^2\right)}$$

where $k = \sqrt{1/(1+r)}$. By Eq. (3) we can find that

$${}_2F_1\left(1-\tau_1,\tau_3;\tau_2+\tau_3;1-k^2\right) \text{ is decreasing } \tau_1 < 1 \text{ and increasing } \tau_1 > 1,$$

$${}_2F_1\left(1-\tau_1,\tau_3;\tau_3+\tau_4;k^2\right) \text{ is increasing } \tau_1 < 1 \text{ and decreasing } \tau_1 > 1,$$

and hense

$$\varphi\left(\frac{1}{k^2}-1;\tau_1,\tau_2,\tau_3\right)$$

as respect to k is decreasing when $\tau_1 < 1$ and increasing $\tau_1 > 1$. Therefore, φ respect r is increasing and decreasing accordingly. □

Corollary 2 *For Q clockwise oriented simple quadrilateral with sides l_j and inner angles τ_j conformal modulus is*

$$\frac{\mathcal{K}'\left(\sqrt{1/(1+r)}\right)}{\mathcal{K}\left(\sqrt{1/(1+r)}\right)}$$

where r is solution to

$$\varphi\left(r;\tau_1,\tau_2,\tau_3\right)=\frac{l_3}{l_4}.$$

References

1. Heikkala, V., Vamanamurthy, M.K., Vuorinen, M.: Generalized elliptic integrals. Comput. Meth. Funct. Theory **9**, 75–109 (2009). https://doi.org/10.1007/BF03321716
2. Mityushev, V.: Schwarz-Christoffel Formula for Multiply Connected Domains. https://doi.org/10.1007/BF03321837
3. Crowdy, D.G.: The Schwarz-Christoffel mapping to bounded multiply connected polygonal domains. Proc. R. Soc. A **461**, 2653–2678 (2005)
4. Driscoll, T.A., Trefethen L.N.: Schwarz-Christoffel Mapping. Oxford University, Oxford (2009). https://doi.org/10.1017/CBO9780511546808
5. Papamichael N., Stylianopoulos N.: Numerical Conformal Mapping, World Scientific, Singapore (2010)
6. Nasser Muhamed, M.S., Rainio, O., Vuorinen, M. : Condenser capacity and hyperbolic perimeter (2021). ArXiv:2103.10237/v2. https://arxiv.org/abs/2103.10237
7. Lehto, O., Virtanen, K.I.: Quasiconformal Mappings in the Plane. Springer, Berlin (1973)

Dimension Reduction in the Periodicity Cell Problem for Plate Reinforced by a Unidirectional System of Fibers

Alexander G. Kolpakov and Sergei I. Rakin

Abstract We reduce 3-D periodicity cell problem (PCP) for plate reinforced by a unidirectional system of fibers to several 2-D problems. Numerical solutions to some 2-D problems are presented. Numerical solutions to some 2-D problems are presented.

1 Introduction

We consider a plate reinforced by a periodic system of fibers. Let us assume that the fibers are parallel to the Oy-axis and form a periodic structure with the periodicity cell (PC) as displayed in Fig. 1.

The structure of the plate is invariant with respect to translation in the direction along the fibers. Them there is a reason to look for a two-dimensional model for this plate. The procedure of dimension reduction is well known for the solids reinforced by periodic systems of fibers (for discussion of the pertinent literature before the 1970s see [1], for the recent literature see [2–4]) and plates of complex geometry made of homogeneous materials [5, 6]. In this paper, we consider an inhomogeneous plate. The characteristic features of plate PC are the free surfaces and the bending/torsion modes of deformation. These features drastically distinguish the plates from the solids.

An attempt of dimension reduction in a model problem of bending of the fiber-reinforced plate was done in [7] by using the double periodic function of complex variables. We call the problem considered in [7] the model because it corresponds to the bending of a plate of infinite thickness. The method of [7] may be useful to compute the strain-stress state (SSS) inside the plate but not near the free surfaces.

A. G. Kolpakov (✉)
SysAn, Novosibirsk, Russia
e-mail: algk@ngs.ru

S. I. Rakin
Siberian Transport University, Novosibirsk, Russia
e-mail: rakinsi@ngs.ru

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_30

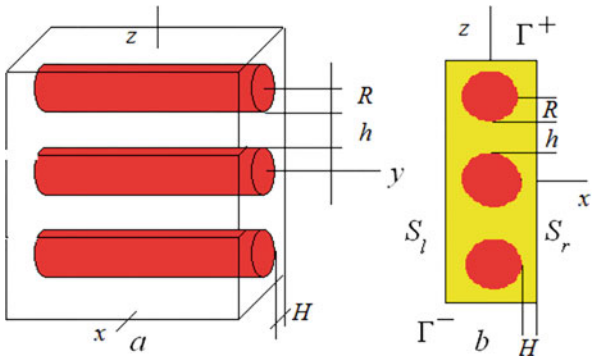


Fig. 1 Periodicity cell of 3-D of fiber-reinforced plate and its 2-D cross-section

We use the homogenization theory [8, 9] as the starting point of the research. The homogenization theory [10] periodicity cell problem (PCP) for plate is

$$\begin{cases} (a_{ijkl}(x,z)N^{AB\mu}_{(k,l)} + (-1)^{\mu}a_{ijAB}(x,z)z^{\mu})_{,j} = 0 & \text{in } P, \\ (a_{ijkl}(x,z)N^{AB\mu}_{(k,l)} + (-1)^{\mu}a_{ijAB}(x,z)z^{\mu})n_j = 0) & \text{on } \Gamma, \\ \mathbf{N}^{AB\mu}(\mathbf{y}) & \text{periodic in } x, y. \end{cases} \tag{1}$$

The variables notation correspondence is the following: $x \leftrightarrow 1$, $y \leftrightarrow 2$, $z \leftrightarrow 3$. The index $\mu = 0.1$.

The PC is invariant with respect to translation in the direction and the elastic constants $a_{ijkl}(x,z)$ in (1) do not depend on variable. One can assume that solution to (1) has the form $bf N^{AB\mu} = \mathbf{N}^{AB\mu}(x,z)$. Substituting into (1), we arrive at the following boundary-value problem:

$$\begin{cases} (a_{i\alpha k\beta}(x,z)N^{AB\mu}_{(k,\beta)} + (-1)^{\mu}a_{i\alpha AB}(x,z)z^{\mu})_{,\alpha} = 0 & \text{in } P, \\ (a_{i\alpha k\beta}(x,z)N^{AB\mu}_{(k,\beta)} + (-1)^{\mu}a_{i\alpha AB}(x,z)z^{\mu})n_{\alpha} = 0) & \text{on } \Gamma, \\ \mathbf{N}^{AB\mu}(\mathbf{y}) & \text{periodic in } x, y. \end{cases} \tag{2}$$

Hereafter $\alpha, \beta = 1, 3$; $A, B = 1, 1; 2, 2; 1, 2; 2, 1$. In (2)

$$a_{i\alpha k\beta}(y)N^{AB\mu}_{k,\beta}(y) + (-1)^{\mu}a_{i\alpha AB}(x,z)z = \\ = a_{i\alpha\theta\beta}(x,z)N^{AB\mu}_{\theta,\beta}(x,z) + a_{i\alpha 2\beta}(x,z)N_{2,\beta})^{AB\mu}(x,z) + (-1)^{\mu}a_{i\alpha AB}(x,z)z^{\mu} \tag{3}$$

We consider the problem (2) for $i = 2$ and $i = 1, 3$ separately.

2 Problem (2) with Index $i = 2$

We assume the fibers and matrix are isotropic. In this case, $a_{2\alpha\theta\beta} = 0$, $a_{2\alpha AB} = 0$ [11] and expression in RHP (3) takes the form ($\alpha = 1, 3$)

$$a_{2\alpha 2\alpha}(y)N^{AB\mu}_{(2,\alpha}(x,z) + \begin{cases} (-1)^{\mu}a_{2121}(x,z)z^{\mu} \text{ if } AB = 21, 12, \\ 0 \qquad\qquad\qquad \text{else.} \end{cases}$$

If $AB \neq 21$, then $N_2^{AB\nu}(x,z) = 0$. For $AB = 21$, problem (2) takes the following form:

$$\begin{cases} (a_{2\alpha 2\alpha}(x,z)N^{21\mu}_{2,\alpha} + (-1)^{\mu}a_{2121}(x,z)z^{\mu}\delta_{\alpha 1})_{,\alpha} = 0 & \text{in } P, \\ (a_{2\alpha 2\alpha}(x,z)N^{21\mu}_{2,\alpha} + (-1)^{\mu}a_{2121}(x,z)z^{\mu}\delta_{\alpha 1})n_{\alpha} = 0 & \text{on } \Gamma, \\ N_2^{21\mu}(x,z) & \text{periodic in } x. \end{cases} \tag{4}$$

Problem (4) corresponds to in-plane shift (if $\mu = 0$) or torsion (if $\mu = 1$), see Fig. 2a.

2.1 *Investigation of the Problem (4)*

We eliminate the mass and surface forces in (4). We demonstrate that there exists a function w, such that ($\nu = 0, 1$)

$$a_{2\delta 2\delta}w_{,\delta} = (-1)^{\nu}a_{2121}z^{\nu}. \tag{5}$$

For $\delta = 2$ and $\delta = 3$, we obtain from (5) $a_{2121}w_{,1} = (-1)^{\nu}a_{2121}z^{\nu}$ and $a_{2323}w_{,3} = 0$. From these equalities, we obtain the following system of ordinary differential equations:

$$w_{,1} = (-1)^{\nu}z^{\nu}, \quad w_{,3} = 0. \tag{6}$$

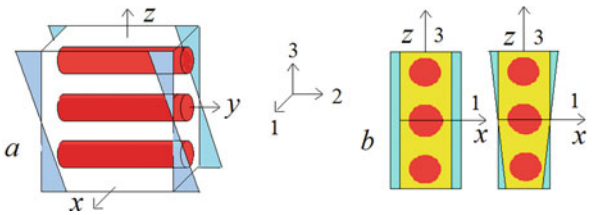


Fig. 2 Shift along the fibers—(**a**), tension and bending perpendicular to the fibers—(**b**)

2.2 In-Plane Shift

For $\nu = 0$, the system (6) takes the form $w_{,1} = 1$, $w_{,3} = 0$. This system is integrable, and its solution is $w(x, z) = x$. Introducing function $M(x, z) = N_1^{120}(x, z) + x$, we rewrite (4) in the form of the boundary-value problem without mass and surface forces (corresponding to $a_{2121}(x, z)z^\nu \delta_{\alpha 1}$ in (4)):

$$\begin{cases} (\Delta M = 0 & \text{in } P_0, \\ \frac{\partial M}{\partial \mathbf{n}} = 0 & \text{on } \Gamma_0, \\ M(x, z) - x \text{ periodic in } x \in [-L, L]. \end{cases} \tag{7}$$

2.3 Torsion

For $\nu = 1$ the system (6) takes form $w_{,1} = -z$, $w_{,3} = 0$. This system is not integrable. Really, the necessary integrability condition is not satisfied for this system because $w_{,13} = -z_{,3} = -1 \neq w_{,31} = 0$.

For $\nu = 1$, (4) takes the form

$$\begin{cases} (a_{2\alpha 2\alpha}(x, z)N_{2,\alpha}^{211} - a_{2121}(x, z)z\delta_{\alpha 1})_{(,\alpha)} = 0 & \text{in } P, \\ (a_{2\alpha 2\alpha}(x, z)N_{2,\alpha}^{211} - a_{2121}(x, z)z\delta_{\alpha 1})n_\alpha = 0 & \text{on } \Gamma, \\ N_2^{211}(x, z) & \text{periodic in } x. \end{cases} \tag{8}$$

We introduce function $\varphi(x, z)$ as

$$\varphi_{,3} = a_{2121}(x, z)(N_{2,1}^{211} - z), \quad \varphi_{,1} = -a_{2323}(x, z)N_{2,3}^{211}. \tag{9}$$

Equations (9) are similar to the formulas introducing the conjugate function [12]. The equality $\varphi_{,31} - \varphi_{,13} = (a_{2121}(x, z)(N_{2,1}^{211} - z))_{,1} + (a_{2323}(x, z)N_{2,3}^{211})_{,3} = 0$ follows from (8) and ensures the existence of this function $\varphi(x, z)$. Express $\mathbf{N}_2^{211}(x, z)$ from (9)

$$N_{2,1}^{211} = \frac{1}{a_{2121}(x, z)}\varphi_{,3} + z, \quad N_{2,3}^{211} = -\frac{1}{a_{2121}(x, z)}\varphi_{,1}. \tag{10}$$

Differentiation of (10) yields $0 = N_{2,13}^{211} - N_{2,31}^{211} = \left(\frac{1}{a_{2121}(x,z)}\varphi_{,3)z}\right)_{,3} + \left(\frac{1}{a_{2121}(x,z)}\varphi_{,1}\right)_{,1}$. Thus

$$\left(\frac{1}{a_{2121}(x, z)}\varphi_{,3}\right)_{,3} + \left(\frac{1}{a_{2121}(x, z)}\varphi_{,1}\right)_{,1} = 1 \tag{11}$$

Consider the boundary conditions on the free (upper and lower) surfaces Γ^+ and Γ^- in (4). With the use of the function $\varphi(x, z)$, these conditions can be written as

$$\begin{aligned}&(a_{2121}(x,z)N_{2,1}^{21\nu} - a_{2121}(x,z)z)n_1 + a_{2323}(x,z)N_{2,3}^{21\nu}n_3\\ &= \varphi_{,3}n_1 - \varphi_{,1}n_3 = \frac{\partial\varphi}{\partial s} = 0\end{aligned} \tag{12}$$

on Γ, where $\frac{\partial}{\partial s}$ is the derivative along the upper or lower boundaries Γ^+ and Γ^-.

Because of (12), the function $\varphi(x, z)$ is constant on the upper and lower boundaries Γ^+ and Γ^-. Without the loss of generality, we can assume that at the lower boundary Γ^-, $\varphi(x, z) = 0$. Function $\frac{\partial N_2^{211}}{\partial n}(x, z) = \frac{\partial N_2^{211}}{\partial x}(x, z)$ is periodic in . By virtue of (10), $\frac{\partial N_2^{211}}{\partial x} - \frac{1}{a_{2121}(x,z)}\varphi_{,3} = z$. Since z is periodic in x, $\frac{\partial N_2^{211}}{\partial x} - \frac{1}{a_{2121}(x,z)}\varphi_{,3}$ is periodic in x. Integrating the last equality over S_l and using (9), we can write

$$\begin{aligned}\varphi(z,-L) &= \varphi(-h,-L) + \int_{-h}^{z}\varphi_{,3}dz = \varphi(-h,-L)\\ &+ \int_{-h}^{z} a_{2121}(x,z)(N_{2,1}^{211} - z)dz\end{aligned} \tag{13}$$

For $z = h$ (on Γ^+)

$$\begin{aligned}\varphi(h,-L) &= \varphi(-h,-L) + \int -h^h a_{2121}(x,z)N_{2,1}^{211}dz\\ &- \int_{-h}^{h} z dz = S_{2121}^1 - \int_{-h}^{h} z dz\end{aligned} \tag{14}$$

S_{2121} is the asymmetric (out-of-plane) stiffness.

As a result, we arrive at the following boundary-value problem:

$$\begin{cases}\left(\frac{1}{a_{2121}(x,z)}\varphi_{,3}\right)_{,3} + \left(\frac{1}{a_{2121}(x,z)}\varphi_{,1}\right)_{,1} = 1 & \text{in } P_0,\\ \varphi = 0 & \text{on } \Gamma^-,\\ \varphi = S_{2121}^1 + \int_{-h}^{h} z dz & \text{on } \Gamma^+,\\ \varphi(x,z) & \text{periodic in } x \in [-L, L].\end{cases} \tag{15}$$

3 Problem (2) with Indices $i = 1, 3$

We change notation i for ξ. In this case, $a_{\xi\alpha 2\beta}(\mathbf{y}) = 0$ and (3) takes the form $(\alpha, \beta, \theta, \xi = 1, 3)$

$$a_{i\alpha k\beta}(\mathbf{y}) N^{AB\mu}_{k,\beta}(\mathbf{y}) + (-1)^{\mu} a_{\xi\alpha AB}(\mathbf{y}) z^{\mu} = a_{\xi\alpha\theta\beta}(\mathbf{y}) N^{AB\mu}_{\theta,\beta}(\mathbf{y}) + (-1)^{\mu} a_{\xi\alpha AB}(\mathbf{y}) z^{\mu}.$$

Here $A, B = 1, 1; 2, 2; 1, 2; 2, 1$. Then the PCP (2) takes the form

$$\begin{cases} \left(a_{\xi\alpha\theta\beta}(x, z) N^{A}_{\theta,\beta} B\mu + (-1)^{\mu} a_{\xi\alpha AB}(x, z) z^{\mu}\right)_{,\alpha} = 0 & \text{in } P, \\ \left(a_{\xi\alpha\theta\beta}(x, z) N^{AB\mu}_{\theta,\beta} + (-1)^{\mu} a_{\xi\alpha AB}(x, z) z^{\mu}\right) n_{\alpha} = 0 & \text{on } \Gamma, \\ \left(N_1^{AB\mu}, N_3^{AB\mu}\right)(x, z) & \text{periodic in } x. \end{cases} \tag{16}$$

Note that $a_{\xi\alpha 12} = 0$ and $a_{\xi\alpha 21} = 0$ for $i = \xi = 1, 3$ [11], then $\left(N_1^{21B\mu}, N_3^{21\mu}\right) = \left(N_1^{12B\mu}, N_3^{12\mu}\right) = 0$. For $AB = 1, 1; 2, 2$, the problem (16) has non-trivial solutions.

In some cases, there exist deformations $e^{AB\mu}_{\theta\beta} = v_(\theta, \beta)^{AB\mu}$ $(\mu = 0, 1)$, such that

$$a_{\xi\alpha AB}(x, z) z^{\nu} = a_{\xi\alpha\theta\beta}(x, z) e^{AB\nu}_{\theta\beta}. \tag{17}$$

3.1 *Index $AB = 22$: Tension-Compression and Bending Along the Fibers (in the Oyz-Plane)*

Equation (17) takes the form $a_{\xi\alpha\theta\beta} e_{\theta\beta} = a_{\xi\alpha 22}(x, z) z^{\nu}$. Having written out coordinatewise, we get

$$\begin{aligned} &a_{1111} e_{11} + a_{1133} e_{33} = -a_{1122} z^{\mu}, \\ &a_{3311} e_{11} + a_{3333} e_{33} = -a_{3322} z^{\mu}, \\ &a_{1313} e_{13} = -a_{1322} z^{\mu} = 0, \\ &a_{3131} e_{31} = -a_{3122} z^{\mu} = 0. \end{aligned} \tag{18}$$

Write the elastic constants in the terms of Young E modulus and Poisson ratio ν [11]

$$a_{1111} = a_{3333} = \frac{E(1-\nu)}{(1+\nu)(1-2\nu)}, \; a_{1133} = a_{3311} = a_{1122} = a_{3322} = \frac{E\nu}{(1+\nu)(1-2\nu)}$$

Substituting into (18), we obtain

$$\begin{cases} (1-\nu)e_{11} + \nu e_{33} = -\nu(x,z)z^{\mu}, \\ \nu e_{11} + (1-\nu)e_{33} = -\nu(x,z)z^{\mu}. \end{cases} \tag{19}$$

Solution to (19) is $e_{11} = e_{33} = -\nu(x,z)z^{\nu}$. In addition $e_{13} = e_{31} = 0$. In many cases the difference between the Poisson ratios of the fibers and the matrix may be ignored: $\nu(x,z) = \nu = const$. Then

$$\frac{\partial v_1}{\partial x} = -\nu z^{\mu}, \quad \frac{\partial v_3}{\partial z} = -\nu z^{\mu}, \quad \frac{\partial v_1}{\partial z} + \frac{\partial v_3}{\partial x} = 0. \tag{20}$$

The problem (20) may be solved in the explicit form. Index $\mu = 0$. From the first two equations in (20), we have $v_1 = -\nu x + f(z)$ and $v_3 = -\nu z + g(x)$. Substituting into the third equation in (20), we arrive at $f'(z) + g'(x) = 0$, then $f(z) = 0$ and $g(x) = 0$. Index $\mu = 1$. We have $v_1 = -\nu z x + f(z)$ and $v_3 = -\nu/2z^2 + g(x)$. Substituting into the third equation in (20), we arrive at $-\nu x + f'(z) + g'(x) = 0$, and obtain $f'(z) = 0,\ g'(x) = \nu x$. Then $f(z) = 0$ and $g(x) = \frac{\nu}{2}x^2$. Finally,

$$v_1^{22} = \begin{cases} -\nu x & \text{if } \mu = 0, \\ -\nu z x & \text{if } \mu = 1, \end{cases} \quad v_3^{22} = \begin{cases} -\nu x & \text{if } \mu = 0, \\ -\nu/2z^2 + \nu/2x^2 & \text{if } \mu = 1 \end{cases} . \tag{21}$$

Introduce $(M_1^{AB\mu}, M_3^{AB\mu}) = (N_1^{AB\mu}, N_3^{AB\mu}) + (v_1^{AB\mu}, v_3^{AB\mu})$. For $(M_1^{22\mu}, M_3^{22\mu})$, the third condition in (16) may be written in the form $[M_1^{22\nu}]_x = -\nu z^{\mu}[x]_x$, $[M_3^{22\nu}]_x = 0$, where $\mu = 0, 1$. Then the problem (16) takes the form

$$\begin{cases} (a_{\xi\alpha\theta\beta}(x,z)M_{\theta,\beta}^{22\nu})_{,\alpha} = 0 & \text{in } P, \\ a_{\xi\alpha\theta\beta}(x,z)M_{\theta,\beta}^{22\nu}n_{\alpha} = 0 & \text{on } \Gamma, \\ \left[M_1^{22\nu}\right]_x = -\nu z^{\mu}\,[x]_x\,, \left[M_3^{22\nu}\right]_x = 0. \end{cases} \tag{22}$$

3.2 *Index $AB = 11$: Tension-Compression and Bending Perpendicular to the Fibers (in the Oxz-Plane)*

In this case, Eq. (17) takes the form $a_{\xi\alpha\theta\beta}e_{\theta\beta} = a_{\xi\alpha 11}(x,z)z^{\nu}$ or, in the coordinate form

$$\begin{aligned} &a_{1111}e_{11} + a_{1133}e_{33} = -a_{1111}z^{\nu}, \\ &a_{3311}e_{11} + a_{3333}e_{33} = -a_{3311}z^{\nu}, \\ &a_{1313}e_{13} = -a_{1311}z^{\nu} = 0, \\ &a_{3131}e_{31} = -a_{3111}z^{\nu} = 0. \end{aligned} \tag{23}$$

Express in (23) the elastic tensor components in the terms of the Young modulus and Poisson ratio. The first and the second equations form the following system:

$$\begin{cases} (1-\nu)e_{11} + \nu e_{33} = -(1-\nu)z^{\nu}, \\ \nu e_{11} + (1-\nu)e_{33} = -\nu z^{\nu}. \end{cases}$$

Its solution is $e_{11} = -z^{\nu}$, $e_{33} = 0$. In addition $e_{13} = e_{31} = 0$. Then, we arrive at

$$\frac{\partial v_1}{\partial x} = -z^{\mu}, \quad \frac{\partial v_3}{\partial z} = 0, \quad \frac{\partial v_1}{\partial z} + \frac{\partial v_3}{\partial x} = 0. \tag{24}$$

From the first two equations in (24), we have $v_1 = -z^{\mu}x + f(z)$, $v_3 = g(x)$. Substituting into the third equation in (24), we have $-\mu z^(\mu-1)x + f'(z) + g'(x) = 0$. Then $f'(z) = 0$, $g'(x) = \mu z^(\mu-1)x$ and $f(z) = 0$, $g(x) = \mu z^(\mu-1)x^2/2$. Finally,

$$v_1^{22} = \begin{cases} -x & \text{if } \mu = 0, \\ -zx & \text{if } \mu = 1, \end{cases} \quad v_3^{22} = \begin{cases} 0 & \text{if } \mu = 0, \\ x^2/2 & \text{if } \mu = 1 \end{cases} \tag{25}$$

The third condition for $(M_1^{11\nu}, M_3^{11\nu})$ in (16) takes the form: $(M_1^{11\nu} - v_1^{11}, M_3^{11\nu} - v_2^{11})$ periodic in x. With regard to (25), it can be written as $[M_1^{11\nu}]_x = -\nu z^{\mu}[x]_x$, $[M_3^{11\nu}]_x = 0$. Then (16) takes the form

$$\begin{cases} (a_{\xi\alpha\theta\beta}(x,z)M_{\theta,\beta}^{11\nu})_{,\alpha} = 0 & \text{in } P, \\ a_{\xi\alpha\theta\beta}(x,z)M_{\theta,\beta}^{11\nu}n_{\alpha} = 0 & \text{on } \Gamma, \\ \left[M_1^{11\nu}\right]_x = -z^{\mu}\,[x]_x\,, \left[M_3^{11\nu}\right]_x = 0. \end{cases} \tag{26}$$

The boundary displacements in (26) are similar to one displayed in Fig. 2b.

3.3 *Index AB* = 12; 21

For $AB = 12$, Eq. (17) takes the form $a_{\xi\alpha 12}(x,z)z^{\nu} = 0$, $\xi, \alpha = 1, 2$. Its solution is $e_{\theta\beta} = 0$. Then $v_1^{12} = v_3^{12} = 0$ and solution to (16) is $(M_1^{12\nu}, M_3^{12\nu}) = 0$.

We have reduced the original 3-D plate PCP to 2-D problems (7), (15), (22), (26). The problems (7) and (15) are anti-plane elasticity problems, and (22) and (26) are planar elasticity theory problems. The advantage of dimension reduction is evident. For example, the full processing power of a typical engineering computer is used to solve 3-D PCP for 5–10 layer plate if a fine finite element mesh is used. The corresponding 2-D PCPs may be solved with fine mesh for 100 and more layers.

4 Some Numerical Computations

We present several illustrative examples interesting from the mechanical point of view. We present solution to the problem (26), corresponding to the tension and bending perpendicular to the fibers. In computations, the fibers Young's modulus GPa and Poisson's ratio ; the matrix Young's modulus GPa and Poisson's ratio. These values correspond to carbon/epoxy composite. The Young ratio. Such type composites are referred to as stiff fibers in the soft matrix. Most models of the fiber-reinforced materials were developed for composites of this type (see, e.g., [1, 10, 13, 14] and references in these books).

In the computations, the periodicity cell dimensions are $h_1 = 1.1$, $h_2 = 2$, $h_3 = 1.1$; $h = 0.1$ and $2H = 0.1$; $R = 0.45$. The values are indicated in the non-dimensional "fast" variables $\mathbf{y}$. The corresponding dimensional values are computed by multiplying by the characteristic size ε . The programs we developed by using the APDL programming language of the ANSYS FEM software [15]. The finite elements PLANE183 are used both for the fibers and the matrix. The characteristic size of the finite elements is 0.03. The total number of finite elements is about 11,000.

4.1 Fibers

Solutions to PCPs for the periodic system of free fibers may be found in [10]. The local SSS in such the system are uniform for the PCPs corresponding to in-plane deformations along the fibers. For the PCPs corresponding to bending/torsion deformations, the SSS in the fibers is similar to the SSS in the classical plate theory, and cannot be accepted as uniform. If the radii of the fibers are small as compared with the thickness of the plate, the SSS in the fibers may be accepted as (approximately) uniform. In the lastr case the method developed in [16] may be useful. For the tension of composite in the direction perpendicular to the fibers (this deformation mode is impossible in the periodic system of free fibers) the local SSS is not homogeneous (von Mises stress variation is about 50%), see Fig. 3a.

4.2 The Matrix and Multi-Continuum Behavior of Composite

We see that the local SSS in the matrix is vary not uniform for all PCPs. Also, we observe significant differences between the SSSs in the fiber and the matrix. The difference is so large that the fibers and the matrix may be qualified as two different media. Based on our numerical computations, we conclude that the stiff fibers in the soft matrix composites demonstrate the multi-continuum behavior predicted for the high-contrast composites theoretically by Panasenko in [17, 18].

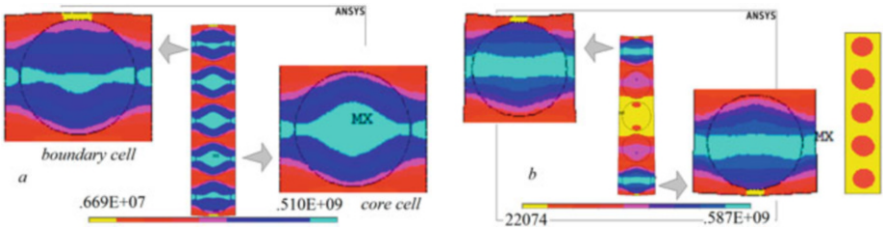


Fig. 3 Local von Mises stress in the tension—(**a**) and the bending—(**b**) modes. 5-layer plate

4.3 The Boundary Layers

We see edge effects near the top and the bottom surfaces of the plate in Fig. 3. The boundary layer thickness is less the thickness of one structural layer $2R + h$ (fiber + surrounding matrix). If the plate is thick, these boundary layers do not influence the effective stiffness of the plate. But they influence the local SSS, as result, the strength of the plate. In the tension mode, the maximum von Mises stress occurs not in the boundary layers, but in the core of the plate, Fig. 3a. In the bending, the maximum von Mises stress occurs in the boundary layers at some distance from the top and the bottom surfaces of the plate, Fig. 3b.

4.4 The Densely Packed Fibers

In our computations, the interparticle distances and is 0.22. The corresponding volume fraction of fibers is 0.53 (the maximum possible value for circular fibers is 0.79). It is the case of densely packed fibers. We meet the effects of the concentration and localization [19] of SSS in the necks between the closely placed fibers. Earlier, these effects were numerically investigated in [20–23] (see [19] for the additional references) for problems close but not identical to the problem investigated in this paper.

5 Conclusion

The original 3-D PCP (1) is reduced to four to 2-D problems: anti-plane elasticity problems (7) and (15) and planar elasticity theory problems (22) and (26). There are open problems, among them the reduction procedure for arbitrary in Sect. 3.1 and detailed numerical analysis of the obtained 2-D problems.

References

1. Sendeckyj, G.P.: Elastic behavior of composites. In: Mechanics of Composite Materials, vol. 2, pp. 45–83. Academic Press, New York/London (1974)
2. Mityushev, V., Rogozin, S.V.: Constructive Methods for Linear and Nonlinear Boundary Value Problems of Analytic Function Theory. Chapman&Hall/CRC, Boca Raton (2000)
3. Gluzman, S., Mityushev, V., Nawalaniec, W.: Computational Analysis of Structured Media. Academic Press, Amsterdam (2018)
4. Drygaś, P., Gluzman, S., Mityushev, V., Nawalaniec, W.: Applied Analysis of Composite Media Analytical and Computational Results for Materials Scientists and Engineers. Elsevier, Amsterdam (2020)
5. Annin, B.D., Kolpakov, A.G., Rakin, S.I.: Homogenization of corrugated plates based on the dimension reduction for the periodicity cell. In: Altenbach, H., et al. (eds.) Problem Mechanics for Materials and Technologies, pp. 30–72. Springer International Publisher, New York (2017)
6. Kolpakov, A.A., Kolpakov, A.G.: On the effective stiffnesses of corrugated plates of various geometries. Int. J. Eng. Sci. **154**, 103327 (2020)
7. Grigolyuk, E.I., Kovalev, Y.D., Fil'shtinskii, L.A. Bending of a layer weakened by through tunnel cuts. Dokl. Akad. Nauk SSSR, **317**(1), 51–53 (1991)
8. Caillerie, D.: Thin elastic and periodic plates. Math. Meth. Appl. Sci. **6**, 159–191 (1984)
9. Kohn, R.V., Vogelius, M.: A new model for thin plates with rapidly varying thickness. Int. J. Solids Struct. **20**, 333–350 (1984)
10. Kalamkarov, A.L., Kolpakov, A.G.: Analysis, Design and Optimization of Composite Structures. Wiley, Chichester (1997)
11. Love, A.E.H.: A treatise on the Mathematical Theory of Elasticity. Cambridge University Press, Cambridge (2013)
12. Sedov, L.I. Continuum Mechanics, vol. 2. Nauka, Moscow (1973). In Russian
13. Agarwal, B.D., Broutman, L.J., Chandrashekhara, K.: Analysis and Performance of Fiber Composites, 4th edn. Wiley, Hoboken (2017)
14. Dvorak, G.: Micromechanics of Composite Materials. Springer, Berlin (2013)
15. Thompson, M.K., Thompson, J.M.: ANSYS Mechanical APDL for Finite Element Analysis. Butterworth-Heinemann, Oxford (2017)
16. Kolpakov, A.G.: Effective rigidities of composite plates. J. Appl. Math. Mech. **46**(4), 529–535 (1982)
17. Panasenko, G.P.: Averaging of processes in strongly inhomogeneous structures. Dokl. Math. **33**(1), 20–22 (1988)
18. Panasenko, G.P.: Multicomponent homogenization for processes in essentially nonhomogeneous structures. Math. USSR-Sb. **69**(1), 143–153 (1991)
19. Kolpakov, A.A., Kolpakov, A.G.: Capacity and Transport in Contrast Composite Structures: Asymptotic Analysis and Applications. CRC Press, Boca Raton (2009)
20. Keller, J.B., Flaherty, J.E.: Elastic behavior of composite media. Commun. Pure. Appl. Math. **26**, 565–580 (1973)
21. Kang, H., Yu, S.: A proof of the Flaherty–Keller formula on the effective property of densely packed elastic composites. Calc. Var. **59**, 22 (2020)
22. Kolpakov, A.A.: Numerical verification of the existence of the energy-concentration effect in a high-contrast heavy-charged composite material. J. Eng. Phys. Thermophys. **80**(4), 812–819 (2007)
23. Rakin, S.I.: Numerical verification of the existence of the elastic energy localization effect for closely spaced rigid disks. J. Eng. Phys. Thermophys. **87**, 246–252 (2014)

Self-Consistent Approximations in the Theory of Composites and Their Limitations

Vladimir Mityushev

Much Ado About Nothing

— William Shakespeare

Abstract Many attempts were undertaken to modify Maxwell's approach in the theory of composites. Self-consistent methods (effective medium approximation, mean field, Mori-Tanaka methods, reiterated homogenization etc) were advanced to determine the effective properties of composites. It is demonstrated by an example that these extensions are methodologically misleading. They lead to a plenty of illusory different formulas reduced to the Maxwell type, lower order estimation for dilute composites.

1 Introduction

Self-consistent (SC) approach has different meanings in various fields of science. We pay a particular attention to Maxwell's approach in the theory of composites [7, p.365]. Many attempts were applied to extend this approach, see [4] and others. Usually, the authors do not associate their study with Schwarz's method [6] and, basing on Eshelby's inclusion problem invent various approximations known as effective medium, Mori-Tanaka method etc. Various justified and unjustified physical arguments and empirical observations have been applying to find a plenty of correct and wrong analytical formulas for the effective properties of composites. It was proved in [10, 12, 13] that such an extension of Maxwell's approach to a finite cluster of inclusions may give at most the effective properties of dilute clusters. In the present paper, we give a rigorous explanation why a SC method does not give more than Maxwell's formula for dilute composites. We consider a 2D conductivity

V. Mityushev (✉)
Cracow University of Technology, Kraków, Poland
e-mail: wladimir.mitiuszew@pk.edu.pl

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_31

problem. Of course, analogous application of SC method to 3D, to elastic and viscous problems has the same shortcomings. The main culprit is the corresponding numerical procedure which contains a conditionally convergent series. Different methods of summations yield different results misleadingly considered as different models.

2 $\mathbb{R}$-Linear Problem

Let $z = x_1 + ix_2$ stand for a complex variable on the plane $\mathbb{R}^2 \equiv \mathbb{C}$. Consider a smooth domain $Q \subset \mathbb{C}$ bounded by a piece-wise Lyapunov simple closed curve ∂Q which divides the complex plane onto two domains. Let non-overlapping disks $D_k = \{|z - a_k| < r_k\}$ $(k = 1, 2, \dots, N)$ lie in Q as shown in Fig. 1. Introduce the disconnected domain, the union of all the disks $D^+ = \cup_{k=1}^n D_k$. Let D denote the complement of $D^+ \cup \partial D^+$ to the domain Q.

Let the disks be occupied by a material of conductivity σ and the conductivity of the host D be normalized to unity. The domain Q is considered as representative volume element (RVE) in the theory of composites. Introduce the complex potentials $\varphi(z)$ and $\varphi_k(z)$ analytic in D and D_k, respectively, and continuously differentiable in the closures of considered domains except at a finite number of points of ∂Q, where the derivatives $\psi(z) = \varphi'(z)$ and $\psi_k(z) = \varphi_k'(z)$ belong to Muskhelishvili's class H [15]. The functions $\varphi_k'(z)$ belong to the Banach space $\mathcal{H}(D^+)$ consisting of functions analytic in $D^+ = \cup_{k=1}^n D_k$ and Hölder continuous in the closure of D^+.

The perfect contact between the components is written as the $\mathbb{R}$-linear problem [3, 11]

$$\varphi(t) = \varphi_k(t) - \rho\overline{\varphi_k(t)}, \quad |t - a_k| = r_k \quad (k = 1, 2, \dots, N). \tag{1}$$

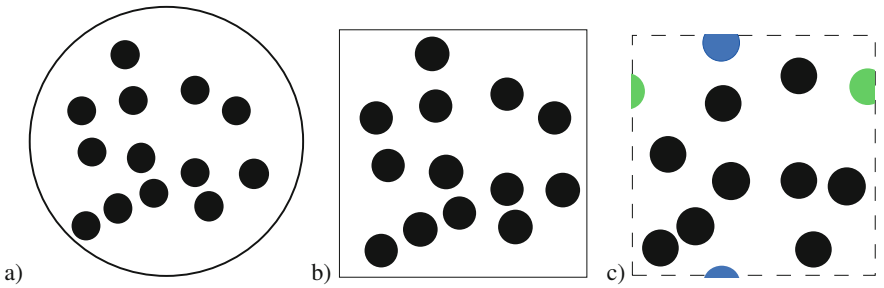


Fig. 1 Various types of cells Q: (**a**) Circular cell; (**b**) Square cell; (**c**) Double periodic square cell. Blue and green pieces of disks form the same two disks, respectively, in the torus topology

The Riemann-Hilbert boundary conditions are given on the boundary of Q

$$\text{Re}\ \overline{\lambda(t)}\varphi(t) = g(t), \quad t \in \partial Q, \tag{2}$$

where $\lambda(t)$ and $g(t)$ are given Hölder continuous functions.

The proper statement of homogenization [1] leads to a periodic boundary value problem for the unit square cell Q as illustrated in Fig. 1c. The cell Q with welded opposite sides is a plane torus without boundary. The function $\varphi(z)$ satisfies the quasi-periodicity conditions with an undetermined real constant d

$$\varphi(z+1) = \varphi(z) + 1, \ \varphi(z+i) = \varphi(z) + id. \tag{3}$$

The stated problems illustrated in Figs. 1b and 1c can be reduced to each other in the following way. Consider for simplicity a macroscopically isotropic composite. Then, the double periodic problem shown in Fig. 1c is equivalent to a boundary value problem for the quadruple cell $\mathcal{Q}$ which consists of a rectangle Q and its three adjusted symmetric with respect to the vertical and horizontal axes shown by dashed lines in Fig. 2. The effective conductivity of $\mathcal{Q}$ coincides with the effective conductivity of the quarters Q, Q^*, Q_*, Q_*^*. Due to symmetry we have the mixed boundary value problem for the domain $\mathcal{Q}$. The Dirichlet conditions $u = \pm 1$ are given on the vertical lines of $\partial\mathcal{Q}$ and the Neumann conditions $\frac{\partial u}{\partial x_2} = 0$ on the

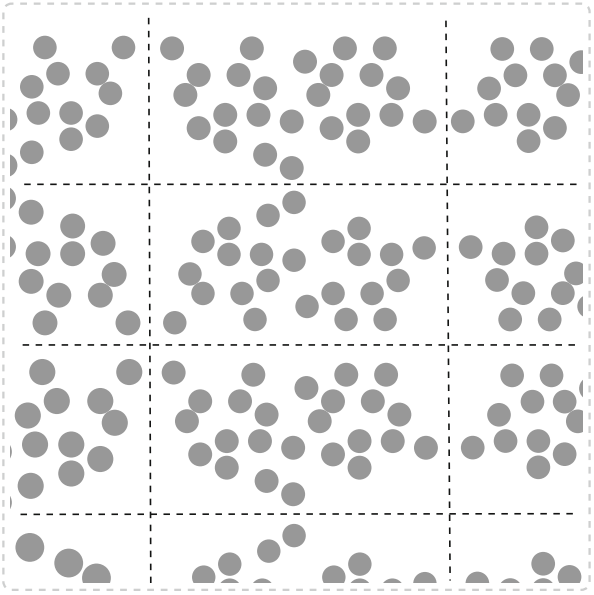


Fig. 2 Rectangular symmetric cells. The whole structure is symmetric with respect to dashed lines

horizontal lines of ∂Q. These conditions on u correspond to the coefficients $\lambda(t) = 1$ and $\lambda(t) = i$ in (2).

3 Self-Consistent Approach

Consider a boundary value problem for medium displayed in Fig. 1a. Let the Dirichlet condition is given on ∂Q

$$\operatorname{Re} \varphi(t) = t, \quad |t| = 1. \tag{4}$$

We are looking for the functions $\varphi(z)$ and $\varphi_k(z)$ analytic in D and D_k, respectively, and continuously differentiable in the closures of the considered domains. These function satisfy the boundary conditions (4) and (1). It is assumed for definiteness that the point $z = 0$ belongs to the domain D.

Such a problem for the infinite plane, i.e., without contour ∂Q, was solved in [8]. The exact solution was obtained for an arbitrary radii and arbitrary contrast parameters in the form of the modified Poincaré series. The problem with vortices in the disk Q was solved in [2]. We now combine the modified Poincaré series [8] and the asymptotic method of functional equations [2] in order to solve the problem (4) and (1) up to $O(r^4)$. It will be sufficient to analyze a self-consistent approximation.

Consider the complex flux introduced in Sect. 2

$$\psi(z) = \varphi'(z) \equiv \frac{\partial u}{\partial x_1} - i\frac{\partial u}{\partial x_2}, \quad z = x_1 + ix_2 \in D, \tag{5}$$

$$\psi_k(z) = \varphi_k'(z) \equiv \frac{\sigma+1}{2}\left(\frac{\partial u}{\partial x_1} - i\frac{\partial u}{\partial x_2}\right), \ |z - a_k| \le r \ (k = 1, 2, \dots, N). \tag{6}$$

The functions (5)–(6) satisfy the $\mathbb{R}$-linear conditions which express the perfect contact between the components [6]

$$\psi(t) = \psi_k(t) + \varrho\left(\frac{r}{t - a_k}\right)^2 \overline{\psi_k(t)}, \quad |t - a_k| = r \quad (k = 1, 2, \dots, N). \tag{7}$$

Introduce the fictitious potential $\psi_0(z)$ analytic in $|z| > R$ and continuous in $|z| \ge R$ following [6] in such a way that

$$\psi(t) = \psi_0(t) + \left(\frac{R}{t}\right)^2 \overline{\psi_0(t)} + 1, \quad |t| = R. \tag{8}$$

Equation (8) is equivalent to (4) as demonstrated in [6].

The $\mathbb{R}$-linear problem (7)–(8) is reduced to the system of functional equations [6]

$$\psi_k(z) = \varrho r^2 \sum_{m\neq k}^{N} \frac{1}{(z-a_m)^2}\overline{\psi_m\left(\frac{r^2}{z-a_m}+a_m\right)} + \left(\frac{R}{z}\right)^2 \overline{\psi_0\left(\frac{R^2}{z}\right)} + 1, \quad |z-a_k| \le r \quad (k=1,2,\dots,N), \tag{9}$$

$$\psi_0(z) = \varrho r^2 \sum_{m=1}^{N} \frac{1}{(z-a_m)^2}\overline{\psi_m\left(\frac{r^2}{z-a_m}+a_m\right)}, \quad |z| \ge R. \tag{10}$$

It follows from [6, 11] that the above system has a unique solution in the space $\mathcal{H}(D^+)$. This solution can be found by a method of successive approximations. The function $\psi(z)$ is expressed through the solution of the system (9)–(10) as follows

$$\psi(z) = \varrho r^2 \sum_{m=1}^{N} \frac{1}{(z-a_m)^2}\overline{\psi_m\left(\frac{r^2}{z-a_m}+a_m\right)} + \left(\frac{R}{z}\right)^2 \overline{\psi_0\left(\frac{R^2}{z}\right)} + 1, \quad z \in D. \tag{11}$$

We find $\psi_k(z)$ and $\overline{\psi_0\left(\frac{R^2}{z}\right)}$ up to $O(r^4)$. The function $\psi_k(z)$ up to $O(r^2)$ has the form

$$\psi_k^{(0)}(z) = 1, \tag{12}$$

since $\psi_0(z)$ is of order $O(r^2)$ by (10). Substitute (12) into (10) to determine $\psi_0(z)$ up to $O(r^4)$

$$\psi_0^{(1)}(z) = \varrho r^2 \sum_{m=1}^{N} \frac{1}{(z-a_m)^2}, \quad |z| \ge R. \tag{13}$$

We now proceed to calculate the next approximation $\psi_k^{(1)}(0)$ using (11). It follows from (13) that

$$\psi_0^{(1)}(z) = \varrho r^2 \sum_{m=1}^{N} \left(\frac{1}{z^2} + \frac{\alpha}{z^3}\dots\right), \quad |z| > R. \tag{14}$$

Hence,

$$\left(\frac{R}{z}\right)^2 \overline{\psi_0^{(1)}\left(\frac{R^2}{z}\right)} = \varrho r^2 \sum_{m=1}^{N} \left(\frac{1}{R^2} + \frac{\overline{\alpha} z}{R^4}\dots\right), \quad |z| > R. \tag{15}$$

Using the approximations (12) and (15) we calculate

$$\psi^{(1)}(0) = 1 + \varrho f + \varrho f \frac{R^2}{N} \sum_{m=1}^{N} \frac{1}{a_m^2}, \tag{16}$$

where $f = \frac{r^2 N}{R^2}$ denote the concentration of inclusions in the cell Q. Let the concentration f be properly defined and oscillate near the same value as N tends to infinity; $R^2 \sim N$ up to a multiplier.

In the framework of a self-consistent method it is suggested that the local flux calculated for a finite N approximates the local flux in the considered infinite composite when $N \to \infty$. However, it is not true. The limit sum from (16) is reduced up to a multiplier to the conditionally convergent series discussed in [9]

$$S_2 = \sum_{m=1}^{\infty} \frac{1}{a_m^2}. \tag{17}$$

It is worth noting that the function $\psi^{(1)}(z)$ is calculated only at one point $z = 0$. It is sufficient to demonstrate a shortage of the SC approach.

In the case of the regular rectangular array, this sum coincides with the famous lattice sum introduced by Rayleigh [16]. The Eisenstein summation [6, 17] was used in order to properly define S_2. The same approach was employed in plane elastic problems in [5, 14].

4 Conclusion and Discussion

Some authors think that by increasing the number of inclusions in a finite cell one can numerically approach to the proper local field in inclusions and to the effective constants. This pipe dream is based on the following misleading assertion: "One can systematically correct such dilute-limit formulas by taking into account interactions between pairs of spheres, triplets of spheres, and so on." However, the conditionally convergent series (17) can lead to nothing, more precisely, to everything, since the series can be arranged in a permutation so that the resulting series will converge to any complex number including infinity. This is the false trick used in SC manipulations which leads to illusory different formulas called models for the effective constants.

References

1. Bakhvalov, N., Panasenko, G.: Homogenisation: Averaging Processes in Periodic Media: Mathematical Problems in the Mechanics of Composite Materials. Kluwer Academic Publishers, Dordrecht (1989)
2. Berlyand, L., Mityushev, V., Ryan, S.D.: Multiple Ginzburg-Landau vortices pinned by randomly distributed small holes. IMA J. Appl. Math. **83**, 977–1006 (2018)
3. Bojarski, B., Mityushev, V.: $\mathbb{R}$-linear problem for multiply connected domains and alternating method of Schwarz. J. Math. Sci. **189**, 68–77 (2013)
4. Choy, T.C.: Effective Medium Theory. Clarendon Press, Oxford (1999)
5. Drygaś, D., Gluzman, S., Mityushev, V., Nawalaniec, W.: Applied Analysis of Composite Media. Analytical and Computational Results for Materials Scientists and Engineers. Elsevier, Duxford (2020)
6. Gluzman, S., Mityushev, V., Nawalaniec, W.: Computational Analysis of Structured Media. Elsevier, Amsterdam (2018)
7. Maxwell, J.C.: A Treatise on Electricity and Magnetism. Clarendon Press Series. Macmillan, Oxford (1873)
8. Mityushev, V.: Plane problem for the steady heat conduction of material with circular inclusions. Arch. Mech. **45**, 211–215 (1993)
9. Mityushev, V.: Transport properties of two-dimensional composite materials with circular inclusions. Proc. R. Soc. Lond. **A455**, 2513–2528 (1999)
10. Mityushev, V., Rylko, N.: Maxwell's approach to effective conductivity and its limitations. Q. J. Mech. Appl. Math. **66**, 241–251 (2013)
11. Mityushev, V.V., Rogosin, S.V.: Constructive methods for linear and non-linear boundary value problems for analytic functions. Chapman & Hall/CRC, Boca Raton (2000)
12. Mityushev, V.: Cluster method in composites and its convergence. Appl. Math. Lett. **77**, 44–48 (2018)
13. Mityushev, V., Drygaś, P.: Effective properties of fibrous composites and cluster convergence. Multiscale Model Simul. SIAM **17**, 696–715 (2019)
14. Mityushev, V., Andrianov, I., Gluzman, S.: L.A. Filshtinsky's contribution to applied mathematics and mechanics of solids. In: Andrianov, I., Gluzman, S., Mityushev, V. (Eds.), Mechanics and Physics of Structured Media. Academic Press, Oxford (2022)
15. Muskhelishvili, N.I.: Singular Integral Equations: Boundary Problems of Function Theory and Their Application to Mathematical Physics. Wolters-Noordhoff, Groningen (1958)
16. Rayleigh, L.: On the influence of obstacles arranged in rectangular order upon the properties of medium. Philos. Mag. **34**, 481–502 (1892)
17. Weil, A.: Elliptic Functions According to Eisenstein and Kronecker. Springer, Berlin (1999)

On Electromagnetic Wave Equations for a Nonhomegenous Microperiodic Medium

Ryszard Wojnar

Abstract We consider the equations of the electromagnetic field in a heterogeneous microperiodic medium, using the representation of the field by the vector potential and the scalar potential. The wave equation for scalar potential is separated from the equation for vector potential, but not vice versa. Thus, the system of equations loses the beautiful symmetry known for the case of a homogeneous medium. The homogenized equations and the expressions for the effective material coefficients are given. A special case of Oersted's experiment in axially nonhomogeneous medium was considered more closely.

1 Introduction

The direct integration of Maxwell's equations is not easy. A serious obstacle is the fact that in these equations the different components of the vectors get mixed up. In the case when the material coefficients of the medium are constant, it is possible to simplify the solution of the electromagnetism problem by introducing potentials, [1, 2]. We will see what a role the potentials play when material coefficients are functions of spatial variables.

There are many works that deal with the interaction of electromagnetic fields with heterogeneous matter, and determine the effective coefficients of the matter. In this paper, we are going to deal with the propagation of electromagnetic waves in a microperiodic material, it is in a medium whose electrical and magnetic susceptibility coefficients are periodic position functions, the size of the period being small in relation to the size of the medium, [3–9].

To determine the effective coefficients of the medium, we will use the asymptotic homogenization method, which is widely applied for similar issues [10].

R. Wojnar (✉)
Institute of Fundamental Technological Research PAS, IPPT PAN, Warszawa, Poland
e-mail: rwojnar@ippt.gov.pl

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_32

It is known that the introduction of scalar φ and vector $\boldsymbol{A}$ potentials allows to separate the equations of electrodynamics, as long as these equations describe the situation in a homogeneous isotropic medium. In a non-homogeneous case, the situation is more difficult, and even the introduction of potentials does not allow for the full separation. Nevertheless, thanks to Lorentz's focus on potentials, it is possible to simplify equations and homogenize the medium.

It turned out that, unlike in the case of a homogeneous medium, the homogenized equations do not completely separate. More precisely, the equation for the scalar potential φ is separated, while the equation for the vector potential $\boldsymbol{A}$ also includes the scalar potential. Apart from that, the wave operators (d'Alembert operators) are not identical for the field $\boldsymbol{A}$ and for the field φ .

As a specific example, the Oersted law for microperiodic inhomogeneous medium was homogenised. Unlike the case of a homogeneous medium (when three Cartesian components of vector $\boldsymbol{A}$ are described by three Poisson equations), in the case of heterogeneous medium equations for the components of vector A, they are not separable.

2 System of Maxwell's Equations

The full system of Maxwell's equations (in Gaussian units) is as follows, [1, 2],

$$\begin{aligned} &\epsilon_{sab} E_{b,a} = -\frac{1}{c}\dot{B}_s \\ &\epsilon_{sab} H_{b,a} = -\frac{1}{c}\dot{D}_s + \frac{4\pi}{c} j_s \\ &(D_a)_{,a} = 4\pi\rho \\ &(B_a)_{,a} = 0 \end{aligned} \tag{1}$$

Here $\boldsymbol{E}$ and $\boldsymbol{H}$ are electric and magnetic fields, ρ and $\boldsymbol{j}$—electric charge and current densities, while $\boldsymbol{D}$ and $\boldsymbol{B}$ are electric and magnetic inductions, respectively. Moreover, we have

$$D_s = \varepsilon E_s \qquad \text{and} \qquad B_s = \mu H_s \tag{2}$$

where respectively, ε and μ are dielectric and magnetic permeabilities. The material coefficients ε and μ are independent of time, but can be any functions of space coordinates. In this paper micro-periodic inhomogeneity wiill be treated. The top dot denotes a partial time derivative, $\partial/\partial t$, and as usualy, the letter c denotes the speed of light in a vacuum.

3 Vector and Scalar Potentials

In order to solve the field problems, electromagnetic potentials are introduced, the vector potential $\boldsymbol{A}$ and the scalar potential φ, as follows

$$\begin{aligned} B_k &= \epsilon_{kab}\, A_{b,a} \\ E_k &= -\frac{1}{c}\,\dot{A}_k - \varphi_{,k} \end{aligned} \tag{3}$$

It can be seen that in time-dependent problems the potentials of electric and magnetic fields are not independent. These potentials are not determined uniquely through the relationships, for example, we can assume the divergence of vector $\boldsymbol{A}$ arbitrarily. Mostly, the Lorentz condition is applied to potentials

$$A_{k,k} + \frac{\varepsilon\mu}{c}\,\dot{\varphi} = 0 \tag{4}$$

By $(2)_2$ and $(3)_1$, the vector $\boldsymbol{H}$ can be expressed as

$$H_k = \frac{1}{\mu}\,\epsilon_{kab}\, A_{b,a}$$

After substituting the potentials, Maxwell's equations read

$$\begin{aligned} &\left(\frac{1}{\mu}\,A_{k,l}\right)_{,l} - \frac{\varepsilon}{c^2}\,\ddot{A}_k + \left(\frac{1}{\mu}\right)_{,k} A_{l,l} - \left(\frac{1}{\mu}\right)_{,l} A_{l,k} + \left(\frac{\varepsilon}{c}\right)_{,k}\dot{\varphi} = -\frac{4\pi\,\mu}{c}\,j_k \\ &\left(\varepsilon\varphi_{,k}\right)_{,k} - \frac{\varepsilon^2\mu}{c^2}\,\ddot{\varphi} + \left(\frac{\varepsilon}{c}\right)_{,k}\dot{A}_k = -4\pi\,\rho \end{aligned} \tag{5}$$

and the following two relationships resulting from the Lorentz condition were used

$$\begin{aligned} &\frac{1}{\mu}\,A_{l,lk} + \frac{\varepsilon}{c}\,\dot{\varphi}_{,k} = -\left(\frac{1}{\mu}\right)_{,k} A_{l,l} - \left(\frac{\varepsilon}{c}\right)_{,k}\dot{\varphi} \\ &\dot{A}_{k,k} = -\frac{\varepsilon\mu}{c}\,\ddot{\varphi} \end{aligned} \tag{6}$$

If the ε and μ material coefficients do not depend on spatial variables, then the system (5) reduces to equations

$$\begin{aligned} A_{k,ll} - \frac{\varepsilon\mu}{c^2}\,\ddot{A}_s &= -\frac{4\pi\,\mu}{c}\,j_k \\ \varphi_{,kk} - \frac{\varepsilon\mu}{c^2}\,\ddot{\varphi} &= -\frac{4\pi}{\varepsilon}\,\rho \end{aligned} \tag{7}$$

in which the fields $\boldsymbol{A}$ and φ are separated.

4 Homogenization

Let $\Gamma = \partial G$ denote the boundary of the domain $G \subset \mathbb{R}^3$. We introduce a parameter

$$\lambda = l/L,$$

where l and L are typical length scales associated with micro-inhomogeneities and the region G, respectively. According to the asymptotic two-scale method, instead of one spatial variable x, we introduce two variables, the macroscopic x and the microscopic y, where $y = x/\lambda$, and instead of the function $f(x)$ we consider the function $f(x, y)$.

Consequently, instead of the domain G, we consider the domain $G \times Y$, where $Y = Y_1 \times Y_2 \times Y_3$ is a basic cell (a rectangular parallelepiped) of micro-periodicity. We use the formula for the total derivative

$$\frac{\partial f(x,y)}{\partial x} \to \frac{\partial f(x,y)}{\partial x} + \frac{1}{\lambda}\frac{\partial f(x,y)}{\partial y} \qquad \text{where} \quad y = \frac{x}{\lambda}$$

In line with the two-scale asymptotic expansions method we assume

$$f^\lambda = f^\lambda(x) = f^{(0)}(x,y) + \lambda^1 f^{(1)}(x,y) + \lambda^2 f^{(2)}(x,y) + \cdots$$

where the functions $f^{(i)}(x, y)$, $i = 0, 1, 2, \ldots$ are Y - periodic. The superscript λ indicates the micro-periodicity of the respective quantities.

It is tacitly assumed that all derivatives appearing in the procedure of asymptotic homogenisation make sense. The effect of micro-structural heterogeneity is described by periodic functions, the so-called *local* functions on the cell.

In our case, according to the method of two-scale homogenization, we assume expansions

$$\begin{aligned} A_k^\lambda &= A_k^{(0)}(x,y) + \lambda A_k^{(1)}(x,y) + \lambda^2 A_k^{(2)}(x,y) + \ldots \\ \varphi^\lambda &= \varphi^{(0)}(x,y) + \lambda\varphi^{(1)}(x,y) + \lambda^2\varphi^{(2)}(x,y) + \ldots \end{aligned} \tag{8}$$

For simplicity, we have omitted the argument t in the expansions.

Performing homogenisation or passing with $\lambda \to 0$ one obtains the homogenised (effective) coefficients of the material.

After substitution expressions (8) into Eq. (5) one obtains for field equations

$$
\begin{aligned}
&\left(\frac{\partial}{\partial x_l}+\frac{1}{\lambda}\frac{\partial}{\partial y_l}\right)\cdot\\
&\left[\frac{1}{\mu}\left(\frac{\partial}{\partial x_l}+\frac{1}{\lambda}\frac{\partial}{\partial y_l}\right)(A_k^{(0)}(x,y)+\lambda A_k^{(1)}(x,y)+\lambda^2 A_k^{(2)}(x,y)+\dots)\right]\\
&-\frac{\varepsilon}{c^2}(\ddot{A}_k^{(0)}(x,y)+\lambda\ddot{A}_k^{(1)}(x,y)+\lambda^2\ddot{A}_k^{(2)}(x,y)+\dots)\\
&+\left(\frac{1}{\mu}\right)_{,k}\left(\frac{\partial}{\partial x_l}+\frac{1}{\lambda}\frac{\partial}{\partial y_l}\right)(A_l^{(0)}(x,y)+\lambda A_l^{(1)}(x,y)+\lambda^2 A_l^{(2)}(x,y)+\dots)\\
&-\left(\frac{1}{\mu}\right)_{,l}\left(\frac{\partial}{\partial x_k}+\frac{1}{\lambda}\frac{\partial}{\partial y_k}\right)(A_l^{(0)}(x,y)+\lambda A_l^{(1)}(x,y)+\lambda^2 A_l^{(2)}(x,y)+\dots)\\
&+\left(\frac{\varepsilon}{c}\right)_{,k}\left(\frac{\partial}{\partial x_k}+\frac{1}{\lambda}\frac{\partial}{\partial y_k}\right)(\dot{\varphi}^{(0)}(x,y)+\lambda\dot{\varphi}^{(1)}(x,y)+\lambda^2\dot{\varphi}^{(2)}(x,y)+\dots)\Big]\\
&\qquad\qquad=-\frac{4\pi}{c}j_k\\
&\left(\frac{\partial}{\partial x_k}+\frac{1}{\lambda}\frac{\partial}{\partial y_k}\right)\\
&\left[\varepsilon\left(\frac{\partial}{\partial x_k}+\frac{1}{\lambda}\frac{\partial}{\partial y_k}\right)(\varphi^{(0)}(x,y)+\lambda\varphi^{(1)}(x,y)+\lambda^2\varphi^{(2)}(x,y)+\dots)\right]\\
&-\frac{\varepsilon^2\mu}{c^2}(\ddot{\varphi}^{(0)}(x,y)+\lambda\ddot{\varphi}^{(1)}(x,y)+\lambda^2\ddot{\varphi}^{(2)}(x,y)+\dots)\\
&+\left(\frac{\varepsilon}{c}\right)_{,k}(\dot{A}_k^{(0)}(x,y)+\lambda\dot{A}_k^{(1)}(x,y)+\lambda^2\dot{A}_k^{(2)}(x,y)+\dots)=-4\pi\rho
\end{aligned}
\tag{9}
$$

and for Lorentz' condition

$$
\begin{aligned}
&\left(\frac{\partial}{\partial x_k}+\frac{1}{\lambda}\frac{\partial}{\partial y_k}\right)\cdot\left(A_k^{(0)}+\lambda A_k^{(1)}+\lambda^2 A_k^{(2)}+\cdots\right)+\\
&\qquad+\frac{\varepsilon\mu}{c}\left(\dot{\varphi}^{(0)}+\lambda\dot{\varphi}^{(1)}+\lambda^2\dot{\varphi}^{(2)}+\cdots\right)=0
\end{aligned}
\tag{10}
$$

Comparing to zero the coefficients at successive negative powers of λ one finds: At λ^{-2}

$$
\frac{\partial}{\partial y_l}\left(\frac{1}{\mu}\frac{\partial}{\partial y_l}\right)A_k^{(0)}(x,y)=0 \quad\text{and}\quad \frac{\partial}{\partial y_k}\left(\frac{1}{\mu}\frac{\partial}{\partial y_k}\right)\varphi^{(0)}(x,y)=0 \tag{11}
$$

Multiply tha last equation by $\varphi^{(0)}(x, y)$ and integrate by over the basic cell Y. One gets

$$\int_Y \frac{\partial}{\partial y_k}\left(\varphi^{(0)}(x, y)\,\frac{1}{\mu}\,\frac{\partial \varphi^{(0)}(x, y)}{\partial y_k}\right) \mathrm{d}Y - \int_Y \frac{1}{\mu}\,\frac{\partial \varphi^{(0)}(x, y)}{\partial y_k}\,\frac{\partial \varphi^{(0)}(x, y)}{\partial y_k}\,\mathrm{d}Y = 0$$

The first integral vanishes by the divergence theorem and the periodic boundary conditions, and, to satisfy the equality one must assume

$$\frac{\partial \varphi^{(0)}(x, y)}{\partial y_k} = 0 \tag{12}$$

In similar manner one gest

$$\frac{\partial\, A_k^{(0)}(x, y)}{\partial y_k} = 0 \tag{13}$$

The last two equations testify that neither $A^{(0)}$ nor $\varphi^{(0)}$ depend on y,

$$A_k^{(0)} = A_k^{(0)}(x) \qquad \text{and} \qquad \varphi^{(0)} = \varphi^{(0)}(x) \tag{14}$$

but they can be also functions of time t.

At λ^{-1} we find

$$\frac{\partial}{\partial y_l}\left[\frac{1}{\mu}\left(\frac{\partial\, A_k^{(0)}}{\partial x_l} + \frac{\partial\, A_k^{(1)}}{\partial y_l}\right)\right] = 0 \quad \text{and} \quad \frac{\partial}{\partial y_k}\left[\varepsilon\left(\frac{\partial\, \varphi^{(0)}}{\partial x_k} + \frac{\partial\, \varphi^{(1)}}{\partial y_k}\right)\right] = 0 \tag{15}$$

Now, we assume

$$A_k^{(1)}(x, y) = a_{kab}(y)\,\frac{\partial\, A_a^{(0)}(x)}{\partial\, x_b} \quad \text{and} \quad \varphi^{(1)}(x, y) = f_a(y)\,\frac{\partial\, \varphi^{(0)}(x)}{\partial\, x_a} \tag{16}$$

and, after substitution into Eqs.(15) we get the equations for *local* functions $a_{kab}(y)$ and $f_a(y)$,

$$\begin{aligned}&\frac{\partial}{\partial y_l}\left[\frac{1}{\mu(y)}\left(\delta_{ak}\delta_{bl} + \frac{\partial a_{kab}(y)}{\partial y_l}\right)\right] = 0 \text{ and}\\ &\frac{\partial}{\partial y_k}\left[\varepsilon(y)\left(\delta_{ak} + \frac{\partial\, f_a(y)}{\partial y_k}\right)\right] = 0\end{aligned} \tag{17}$$

Finally, at λ^0

$$\frac{\partial}{\partial x_l}\left[\frac{1}{\mu}\left(\frac{\partial A_k^{(0)}(x)}{\partial x_l}+\frac{\partial A_k^{(1)}(x,y)}{\partial y_l}\right)\right]$$
$$+\frac{\partial}{\partial y_l}\left[\frac{1}{\mu}\left(\frac{\partial A_k^{(1)}(x,y)}{\partial x_l}+\frac{\partial A_k^{(2)}(x,y)}{\partial y_l}\right)\right]-\frac{\varepsilon}{c^2}\,\ddot{A}_k^{(0)}(x)$$
$$+\left(\frac{1}{\mu}\right)_{,k}\left(\frac{\partial A_l^{(0)}(x)}{\partial x_l}+\frac{\partial A_l^{(1)}(x,y)}{\partial y_l}\right)-\left(\frac{1}{\mu}\right)_{,l}\left(\frac{\partial A_l^{(0)}(x}{\partial x_k}+\frac{\partial A_l^{(1)}(x,y)}{\partial y_k}\right)$$
$$+\left(\frac{\varepsilon}{c}\right)_{,k}\left(\frac{\partial\dot{\varphi}^{(0)}(x)}{\partial x_k}+\frac{\partial\dot{\varphi}^{(1)}(x,y)}{\partial y_k}\right)=-\frac{4\pi}{c}\,j_k$$
$$\frac{\partial}{\partial x_k}\left[\varepsilon\left(\frac{\partial\varphi^{(0)}(x)}{\partial x_k}+\frac{\partial\varphi^{(1)}(x,y)}{\partial y_k}\right)\right]+\frac{\partial}{\partial y_k}\left[\varepsilon\left(\frac{\partial\varphi^{(1)}(x,y)}{\partial x_k}+\frac{\partial\varphi^{(2)}(x,y)}{\partial y_k}\right)\right]$$
$$-\frac{\varepsilon^2\mu}{c^2}\,\ddot{\varphi}^{(0)}(x)+\left(\frac{\varepsilon}{c}\right)_{,k}\dot{A}_k^{(0)}(x)=-4\pi\rho$$

and what concerns Lorentz' condition

$$\frac{\partial A_k^{(0)}(x)}{\partial x_k}+\frac{\partial A_k^{(1)}(x,y)}{\partial y_k}+\frac{\varepsilon\mu}{c}\,\dot{\varphi}^{(0)}(x)=0$$

Introduce the average

$$\langle(\cdots)\rangle=\frac{1}{Y}\int_Y(\cdots)\,\mathrm{d}Y \tag{18}$$

to the last three equations and obtain

$$\frac{\partial}{\partial x_l}\left\langle\frac{1}{\mu}\left(\frac{\partial A_k^{(0)}(x)}{\partial x_l}+\frac{\partial A_k^{(1)}(x,y)}{\partial y_l}\right)\right\rangle-\left\langle\frac{\varepsilon}{c^2}\right\rangle\ddot{A}_k^{(0)}(x)$$
$$+\left\langle\left(\frac{1}{\mu}\right)_{,k}\left(\frac{\partial A_l^{(0)}(x)}{\partial x_l}+\frac{\partial A_l^{(1)}(x,y)}{\partial y_l}\right)\right\rangle-\left\langle\left(\frac{1}{\mu}\right)_{,l}\left(\frac{\partial A_l^{(0)}(x)}{\partial x_k}+\frac{\partial A_l^{(1)}(x,y)}{\partial y_k}\right)\right\rangle$$
$$+\left\langle\left(\frac{\varepsilon}{c}\right)_{,k}\left(\frac{\partial\dot{\varphi}^{(0)}(x)}{\partial x_k}+\frac{\partial\dot{\varphi}^{(1)}(x,y)}{\partial y_k}\right)\right\rangle=-\frac{4\pi}{c}\,\langle j_k\rangle$$
$$\frac{\partial}{\partial x_k}\left\langle\varepsilon\left(\frac{\partial\varphi^{(0)}(x)}{\partial x_k}+\frac{\partial\varphi^{(1)}(x,y)}{\partial y_k}\right)\right\rangle-\left\langle\frac{\varepsilon^2\mu}{c^2}\right\rangle\ddot{\varphi}^{(0)}(x)+\left(\frac{\varepsilon}{c}\right)_{,k}\dot{A}_k^{(0)}(x)=-4\pi\,\langle\rho\rangle \tag{19}$$

and (Lorentz' condition)

$$\left\langle\frac{1}{\mu}\left[\frac{\partial A_k^{(0)}(x)}{\partial x_k}+\frac{\partial A_k^{(1)}(x,y)}{\partial y_k}\right]\right\rangle+\left\langle\frac{\varepsilon}{c}\right\rangle\dot{\varphi}^{(0)}(x)=0 \tag{20}$$

Integrating by parts inside the averaging signs in the second row of formula (19) we observe that both components compensate to zero, because

$$\left\langle \left(\frac{1}{\mu}\right)_{,k} \left(\frac{\partial A_l^{(0)}(x)}{\partial x_l} + \frac{\partial A_l^{(1)}(x,y)}{\partial y_l} \right) \right\rangle = - \left\langle \frac{1}{\mu} \frac{\partial^2 A_k^{(1)}(x,y)}{\partial y_k \partial y_l} \right\rangle$$

$$\left\langle \left(\frac{1}{\mu}\right)_{,l} \left(\frac{\partial A_l^{(0)}(x}{\partial x_k} + \frac{\partial A_l^{(1)}(x,y)}{\partial y_k} \right) \right\rangle = - \left\langle \frac{1}{\mu} \frac{\partial^2 A_k^{(1)}(x,y)}{\partial y_l \partial y_k} \right\rangle$$

Moreover

$$\left\langle \left(\frac{\varepsilon}{c}\right)_{,k} \left(\frac{\partial \dot{\varphi}^{(0)}(x)}{\partial x_k} + \frac{\partial \dot{\varphi}^{(1)}(x,y)}{\partial y_k} \right) \right\rangle = - \frac{1}{c} \left\langle \varepsilon \frac{\partial^2 f_a(y)}{\partial y_k \partial y_k} \right\rangle \frac{\partial \dot{\varphi}^{(0)}(x)}{\partial x_a}$$

and

$$\left\langle \left(\frac{\varepsilon}{c}\right)_{,k} \right\rangle = 0$$

Finally, we insert the functions $A_k^{(1)}$ and $\varphi^{(1)}$, cf. Eqs.(16), into Eqs.(19) and (20), and get

$$\begin{aligned} &\left(\frac{1}{\mu}\right)_{akbl} \frac{\partial^2 A_a^{(0)}}{\partial x_l \partial x_b} - \left\langle \frac{\varepsilon}{c^2} \right\rangle \ddot{A}_k^{(0)}(x) - \frac{1}{c} \left\langle \varepsilon \frac{\partial^2 f_a(y)}{\partial y_k \partial y_k} \right\rangle \frac{\partial \dot{\varphi}^{(0)}(x)}{\partial x_a} = - \frac{4\pi}{c} \langle j_k \rangle \\ &\varepsilon_{ak}^{\text{eff}} \frac{\partial^2 \varphi^{(0)}(x)}{\partial x_k \partial x_a} - \left\langle \frac{\varepsilon^2 \mu}{c^2} \right\rangle \ddot{\varphi}^{(0)}(x) = -4\pi \langle \rho \rangle \end{aligned} \tag{21}$$

and (Lorentz' condition)

$$\left(\frac{1}{\mu}\right)_{akbk}^{\text{eff}} \frac{\partial A_a^{(0)}(x)}{\partial x_b} + \left\langle \frac{\varepsilon}{c} \right\rangle \dot{\varphi}^{(0)}(x) = 0 \tag{22}$$

where

$$\left(\frac{1}{\mu}\right)_{akbl}^{\text{eff}} = \left\langle \frac{1}{\mu} \left(\delta_{ak}\delta_{bl} + \frac{\partial a_{kab}}{\partial y_l} \right) \right\rangle \quad \text{and} \quad \varepsilon_{ak}^{\text{eff}} = \left\langle \varepsilon \left(\delta_{ak} + \frac{\partial f_a}{\partial y_k} \right) \right\rangle \tag{23}$$

In particular,

$$\left(\frac{1}{\mu}\right)_{akbk}^{\text{eff}} = \left\langle \frac{1}{\mu} \left(\delta_{ab} + \frac{\partial a_{kab}}{\partial y_k} \right) \right\rangle \tag{24}$$

The homogenized equations (21) can be compared with the homogeneous medium equations (7). We have still four second order differential equations. One can see that the equation for scalar potential φ is separated from the equation for vector potential $\boldsymbol{A}$, although the separation is not complete: there is a scalar potential in the equation for vector potential. The d'Alembert operator in the equation for the vector $\boldsymbol{A}$, however, differs from the corresponding operator for the scalar φ.

Also Lorentz' condition (22) does not have its original simplicity given by Eq.(4).

5 Oersted's Experiment in a Nonhomogeneous Medium

Oersted's law is stating that an electric current creates a magnetic field in the following manner

$$\epsilon_{sab} H_{b,a} = \frac{4\pi}{c} j_s \tag{25}$$

Maxwell's second equation $(1)_1$ is a generalization of this law. In this case the current density $\boldsymbol{j}$ is steady. The vector $\boldsymbol{H}$ of magnetic field can be expressed by the vector potential $\boldsymbol{A}$ as

$$H_k = \frac{1}{\mu} \epsilon_{kab} A_{b,a} \tag{26}$$

The coefficient of magnetic permeability μ is a microperiodic function of space position. Substituting (26) into (25) we get

$$\left(\frac{1}{\mu} A_{k,l}\right)_{,l} + \left(\frac{1}{\mu}\right)_{,k} A_{l,l} - \left(\frac{1}{\mu}\right)_{,l} A_{l,k} = -\frac{4\pi}{c} j_k$$

We can always choose a vector potential such that

$$A_{k,k} = 0 \tag{27}$$

Then we get

$$\left(\frac{1}{\mu} A_{k,l}\right)_{,l} - \left(\frac{1}{\mu}\right)_{,l} A_{l,k} = -\frac{4\pi}{c} j_k \tag{28}$$

If the mediumt is homogeneous (μ = constant) we get three Poisson equations for three Cartesian components of the vector $\boldsymbol{A}$,

$$A_{k,ll} = -\frac{4\pi\,\mu}{c} j_k \tag{29}$$

Proceeding with Eq.(28) in the manner described in the previous point we get

$$\left(\frac{1}{\mu}\right)_{akbl} \frac{\partial^2 A_a^{(0)}}{\partial x_l \partial x_b} = -\frac{4\pi}{c} \langle j_k \rangle \tag{30}$$

where the tensor $(1/\mu)_{akbl}$ is given by Eq.(24).

Now, the condition (27) reads

$$\left(\frac{1}{\mu}\right)^{\text{eff}}_{akbk} \frac{\partial A_a^{(0)}(x)}{\partial x_b} = 0 \tag{31}$$

Remark In 1802 Gian Domenico Romagnosi (1761–1835) observed in Trento the deviation of the magnetic needle induced by an electric current. Joseph Hamel has pointed out, Romagnosi's discovery was documented in the book by Joseph Izarn, *Manuel du Galvanisme* (1805), where a galvanic current is explicitly mentioned. It was also mentioned on page 340 of the book by Giovanni Aldini, *Essai théorique et expérimental sur le Galvanisme* (1804). Aldini was also communicating with Oersted at the time, Hamel notes, [11, 12] (https://en.wikipedia.org/wiki/Gian_Domenico_Romagnosi).

In cylindrical co-ordinates (r, ϑ, z) the *curl* formula reads

$$\boldsymbol{B} = \nabla \times \boldsymbol{A} = \frac{1}{r} \begin{vmatrix} \boldsymbol{e}_r & r\,\boldsymbol{e}_\vartheta & \boldsymbol{e}_z \\ \dfrac{\partial}{\partial r} & \dfrac{\partial}{\partial \vartheta} & \dfrac{\partial}{\partial z} \\ A_r & A_\vartheta & A_z \end{vmatrix} \tag{32}$$

We consider the case in which the components B_r and B_z vanish. Thus $\boldsymbol{B} = (0, B_\vartheta, 0)$ and Oersted's law takes the form

$$\frac{4\pi}{c} \boldsymbol{j} = \nabla \times \boldsymbol{H} = \nabla \times \left(\frac{1}{\mu}\,\boldsymbol{B}\right) = \frac{\partial B_\vartheta}{\partial r}\,\boldsymbol{e}_z = -\frac{\partial}{\partial r}\left(\frac{1}{\mu}\, r\, \frac{\partial A_z}{\partial r}\right) \boldsymbol{e}_z$$

We have additionally assumed here that neither μ nor B_ϑ and A_r depend on the variable z. Thus we have

$$\frac{\partial}{\partial r}\left(\frac{1}{\mu}\, r\, \frac{\partial A_z}{\partial r}\right) = -\frac{4\pi}{c}\, j \tag{33}$$

We have accepted that the current density $\boldsymbol{j}$ has only one component, $\boldsymbol{j} = (0, 0, j)$.

We are now using the asymptotic homogenization method. We put

$$A_z = A_z^{(0)} + \lambda A_z^{(1)} + \lambda^2 A_z^{(2)} + \cdots$$

and

$$\frac{\partial}{\partial r} \to \left(\frac{\partial}{\partial r} + \frac{1}{\lambda} \frac{\partial}{\partial s} \right)$$

where r is a macroscopic and s is a microscipic independent variable.

To avoid the singularity at λ^{-2}, $\lambda \to 0$, we have to put

$$\frac{\partial}{\partial s} \left(\frac{1}{\mu} \frac{\partial A_z^{(0)}}{\partial s} \right) = 0 \qquad \text{and} \qquad \mu = \mu(s) > 0 \tag{34}$$

Hence, $A_z^{(0)}$ does not depend on s, and depends at most on the macroscipic variable r, $A_z^{(0)} = A_z^{(0)}(r)$ only.

On the other hand, the vanishing of the coefficient at λ^{-1} means that the following equality holds

$$\frac{\partial}{\partial s} \Big[\frac{1}{\mu} \left(\frac{\partial A_z^{(0)}}{\partial r} + \frac{\partial A_z^{(1)}}{\partial s} \right) \Big] = 0$$

which can be satisfied by substituting

$$A_z^{(1)}(r, s) = \psi(s) \frac{\partial A_z^{(0)}(r)}{\partial r} \tag{35}$$

The function $\psi = \psi(s)$ is found from the equation

$$\frac{\mathrm{d}}{\mathrm{d}s} \Big[\frac{1}{\mu(s)} \left(1 + \frac{\mathrm{d}\,\psi(s)}{\mathrm{d}s} \right) \Big] = 0 \tag{36}$$

At λ^0 we get after the averaging

$$\frac{1}{r} \frac{\partial}{\partial r} \left[r \left\langle \frac{1}{\mu(s)} \left(1 + \frac{\mathrm{d}\,\psi(s)}{\mathrm{d}s} \right) \right\rangle \frac{\partial A_z^{(0)}}{\partial r} \right] = - \frac{4\pi}{c} \langle j \rangle \tag{37}$$

or

$$\left(\frac{1}{\mu} \right)^{\text{eff}} \frac{1}{r} \frac{\mathrm{d}}{\mathrm{d}r} \left(r \frac{\mathrm{d} A_z^{(0)}}{\mathrm{d}r} \right) = - \frac{4\pi}{c} \langle j \rangle \tag{38}$$

where the effective (homogenized) coefficient is

$$\left(\frac{1}{\mu}\right)^{\text{eff}} = \left\langle \frac{1}{\mu(s)} \left(1 + \frac{\mathrm{d}\,\psi(s)}{\mathrm{d}s}\right)\right\rangle = \frac{1}{Y}\int_0^Y \frac{1}{\mu(s)} \left(1 + \frac{\mathrm{d}\,\psi(s)}{\mathrm{d}s}\right) \mathrm{d}s \tag{39}$$

For example, consider a medium made of coaxial cylinders, with alternating magnetic properties. Thus, the basic cell has the property

$$\mu = \mu(s) = \begin{cases} \mu_0 \text{ if } 0 \le s < \alpha\, Y \\ \mu_1 \text{ if } \alpha\, Y \le s \le Y \end{cases} \tag{40}$$

The coefficient α is positive and less than 1. Integrating Eq.(39) with the coefficient (40), and remembering about periodic boundary conditions on the function ψ we get

$$\left(\frac{1}{\mu}\right)^{\text{eff}} = \frac{1}{\alpha\,\mu_0 + (1-\alpha)\,\mu_1} = \frac{1}{\mu^{\text{arith}}} \tag{41}$$

where $\mu^{\text{arith}} = \alpha\,\mu_0 + (1-\alpha)\,\mu_1$ is the arithmetic average. Now, Oersted's equation (38) can be written as

$$\frac{1}{r}\frac{\mathrm{d}}{\mathrm{d}r}\left(r\frac{\mathrm{d}A_z^{(0)}}{\mathrm{d}r}\right) = -\frac{4\pi}{c}\,\mu^{\text{arith}}\,\langle j\rangle \tag{42}$$

This is a Poisson equation on a component $A_z^{(0)}$ of the vector potential.

6 Conclusions

Introduction of vector and scalar potentials to Maxwell's equations for a medium with microperiodic heterogeneous material coefficients leads to two partial differential equations of the second order that can still be simplified by Lorentz' condition for potentials. After carrying out the homogenization, the equation for scalar potential is separated.

Homogenization of the equation describing Oersted's experiment in a heterogeneous mediumr leads to the equation similar as in a homogeneous case, except that the homogenized coefficient appears.

References

1. Landau, L.D., Lifshitz, E.M.: Electrodynamics of continuous media. In: Course of Theoretical Physics Volume 8 (2nd English ed.) Pergamon Press, Oxford (1987). Translated from the Russian by J. B. Sykes and W. H. Reid
2. Suffczyński, M.: Elektrodynamika. PWN, Warszawa (1964)
3. Telega, J.J.: Piezoelectricity and homogenization. Application to biomechanics. In: Maugin, G.A. (Ed.), Continuum Models and Discrete Systems, vol. 2, p. 220. Longman, Essex (1991)
4. Gałka, A., Telega, J.J., Wojnar, R.: Homogenization and thermopiezolectricity. Mech. Res. Commun. **19**, 315–324 (1992)
5. Rylko, N.: Effective anti-plane properties of piezoelectric fibrous composites. Acta Mech. **224**, 2719–2734 (2013)
6. Gambin, B.: Influence of microstructure on properties of elastic, piezoelectric and thermoelastic composites. Prace IPPT - IFTR Reports, Warszawa (2006)
7. Malevich, A.E., Mityushev, V.V., Adler, P.M.: Electrokinetic phenomena in wavy channels. J. Colloid Interface Sci. **345** 72–87 (2010)
8. Andrianov, I., Mityushev, V.: Exact and "Exact" Formulae in the theory of composites. In: Drygaś, P., Rogosin, S. (Eds.), Modern Problems in Applied Analysis, Trends in Mathematics. Springer, Berlin (2018)
9. Mityushev, V., Nosov, D., Wojnar, R.: Two-dimensional equations of magneto-electro-elasticity. In: Mechanics and Physics of Structured Media Asymptotic and Integral Equations Methods of Leonid Filshtinsky, chap. 3. Elsevier, London (2022). Coordinators: Andrianov Igor, Gluzman Simon, Mityushev Vladimir
10. Sanchez-Palencia, E.: Non-homogeneous Media and Vibration Theory. Springer, Berlin (1980)
11. Hamel, J.: Historical Account of the Introduction of the Galvanic and Electro-Magnetic Telegraph, p. 34 (1859)
12. Stringari, S., Wilson, R.R.: Romagnosi and the discovery of electromagnetism. Rend. Fis. Acc. Lincei **11**, 115–136 (2000)

Part VII
Generalized Functions and Applications

A Note on Composition Operators Between Weighted Spaces of Smooth Functions

Andreas Debrouwere and Lenny Neyt

Abstract For certain weighted locally convex spaces X and Y of one real variable smooth functions, we characterize the smooth functions $\varphi : \mathbb{R} \to \mathbb{R}$ for which the composition operator $C_\varphi : X \to Y,\ f \mapsto f \circ \varphi$ is well-defined and continuous. This problem has been recently considered for $X = Y$ being the space $\mathscr{S}$ of rapidly decreasing smooth functions (Galbis and Jordá, Rev Mat Iberoam 34:397–412, 2018) and the space $\mathscr{O}_M$ of slowly increasing smooth functions (Albanese et al., J Math Anal Appl 54:126303, 2022). In particular, we recover both these results as well as obtain a characterization for $X = Y$ being the space $\mathscr{O}_C$ of very slowly increasing smooth functions.

1 Introduction

One of the most fundamental questions in the study of composition operators is to characterize when such an operator is well-defined and continuous in terms of its symbol. The goal of this article is to consider this question for weighted locally convex spaces of one real variable smooth functions.

Let $\varphi : \mathbb{R} \to \mathbb{R}$ be smooth. In [1] Galbis and Jordá showed that the composition operator $C_\varphi : \mathscr{S} \to \mathscr{S},\ f \mapsto f \circ \varphi$, with $\mathscr{S}$ the space of rapidly decreasing smooth functions [2], is well-defined (continuous) if and only if

$$\exists N \in \mathbb{Z}_+ \ : \ \sup_{x \in \mathbb{R}} \frac{1 + |x|}{(1 + |\varphi(x)|)^N} < \infty$$

A. Debrouwere (✉)
Department of Mathematics and Data Science, Vrije Universiteit Brussel, Brussels, Belgium
e-mail: andreas.debrouwere@vub.be

L. Neyt
Department of Mathematics, Analysis, Logic and Discrete Mathematics, Ghent University, Ghent, Belgium
e-mail: lenny.neyt@UGent.be

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_33

and

$$\forall p \in \mathbb{Z}_+ \, \exists N \in \mathbb{N} \, : \, \sup_{x \in \mathbb{R}} \frac{|\varphi^{(p)}(x)|}{(1+|\varphi(x)|)^N} < \infty.$$

Albanese et al. [3] proved that the composition operator $C_\varphi : \mathscr{O}_M \to \mathscr{O}_M$, with $\mathscr{O}_M$ the space of slowly increasing smooth functions [2], is well-defined (continuous) if and only if $\varphi \in \mathscr{O}_M$. In [3, Remark 2.6] they also pointed out that the corresponding result for the space $\mathscr{O}_C$ of very slowly increasing smooth functions [2] is false, namely, they showed that $\sin(x^2) \notin \mathscr{O}_C$, while, obviously, $\sin x, x^2 \in \mathscr{O}_C$.

Inspired by these results, we study in this article the following general question: Given two weighted locally convex spaces X and Y of smooth functions, when is the composition operator $C_\varphi : X \to Y$ well-defined (continuous)? We shall consider this problem for X and Y both being Fréchet spaces, (LF)-spaces, or (PLB)-spaces.

We now state a particular instance of our main result that covers many well-known spaces. We need some preparation. Given a positive continuous function v on $\mathbb{R}$, we write $\mathscr{B}^n_v$, $n \in \mathbb{N}$, for the Banach space consisting of all $f \in C^n(\mathbb{R})$ such that

$$\|f\|_{v,n} = \max_{p \leq n} \sup_{x \in \mathbb{R}} \frac{|f^{(p)}(x)|}{v(x)} < \infty.$$

For $v \geq 1$ we consider the following three weighted spaces of smooth functions

$$\mathscr{K}_v = \varprojlim_{N \in \mathbb{N}} \mathscr{B}^N_{1/v^N},$$

$$\mathscr{O}_{C,v} = \varinjlim_{N \in \mathbb{N}} \varprojlim_{n \in \mathbb{N}} \mathscr{B}^n_{v^N},$$

$$\mathscr{O}_{M,v} = \varprojlim_{n \in \mathbb{N}} \varinjlim_{N \in \mathbb{N}} \mathscr{B}^n_{v^N}.$$

Theorem 2 below implies the following result:

Theorem 1 *Let $v, w : \mathbb{R} \to [1, \infty)$ be continuous functions such that*

$$\sup_{x,t \in \mathbb{R}, |t| \leq 1} \frac{v(x+t)}{v^\lambda(x)} < \infty \qquad \text{and} \qquad \sup_{x,t \in \mathbb{R}, |t| \leq 1} \frac{w(x+t)}{w^\mu(x)} < \infty,$$

for some $\lambda, \mu > 0$. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be smooth. Then,

(I) The following statements are equivalent:

(i) $C_\varphi(\mathscr{K}_v) \subseteq \mathscr{K}_w$.
(ii) $C_\varphi : \mathscr{K}_v \to \mathscr{K}_w$ is continuous.
(iii) φ satisfies the following two properties

$$(a)\ \ \exists \lambda > 0\ :\ \sup_{x\in\mathbb{R}} \frac{w(x)}{v^{\lambda}(\varphi(x))} < \infty.$$

$$(b)\ \ \forall p \in \mathbb{Z}_+\ \exists \lambda > 0\ :\ \sup_{x\in\mathbb{R}} \frac{|\varphi^{(p)}(x)|}{v^{\lambda}(\varphi(x))} < \infty.$$

(II) The following statements are equivalent:

(i) $C_\varphi(\mathscr{O}_{C,v}) \subseteq \mathscr{O}_{C,w}$.
(ii) $C_\varphi : \mathscr{O}_{C,v} \to \mathscr{O}_{C,w}$ *is continuous.*
(iii) φ *satisfies the following two properties*

$$(a)\ \ \exists \mu > 0\ :\ \sup_{x\in\mathbb{R}} \frac{v(\varphi(x))}{w^{\mu}(x)} < \infty.$$

$$(b)\ \ \forall p, k \in \mathbb{Z}_+\ :\ \sup_{x\in\mathbb{R}} \frac{|\varphi^{(p)}(x)|}{w^{1/k}(x)} < \infty.$$

(III) The following statements are equivalent:

(i) $C_\varphi(\mathscr{O}_{M,v}) \subseteq \mathscr{O}_{M,w}$.
(ii) $C_\varphi : \mathscr{O}_{M,v} \to \mathscr{O}_{M,w}$ *is continuous.*
(iii) φ *satisfies the following two properties*

$$(a)\ \ \exists \mu > 0\ :\ \sup_{x\in\mathbb{R}} \frac{v(\varphi(x))}{w^{\mu}(x)} < \infty.$$

$$(b)\ \ \forall p \in \mathbb{Z}_+\ \exists \mu > 0\ :\ \sup_{x\in\mathbb{R}} \frac{|\varphi^{(p)}(x)|}{w^{\mu}(x)} < \infty.$$

By setting $v(x) = w(x) = 1 + |x|$ in Theorem 1 we recover the above results about $\mathscr{S}$ and $\mathscr{O}_M$ from [1, 3] as well as the following characterization for the space $\mathscr{O}_C$ of very slowly increasing smooth functions: $C_\varphi : \mathscr{O}_C \to \mathscr{O}_C$ is well defined (continuous) if and only if

$$\exists N \in \mathbb{N}\ :\ \sup_{x\in\mathbb{R}} \frac{|\varphi(x)|}{(1+|x|)^N} < \infty \quad \text{and} \quad \forall p, k \in \mathbb{Z}_+\ :\ \sup_{x\in\mathbb{R}} \frac{|\varphi^{(p)}(x)|}{(1+|x|)^{1/k}} < \infty.$$

For $v = w = 1$, Theorem 1 gives the following result for the Fréchet space $\mathscr{B}$ of smooth functions that are bounded together will all their derivatives [2]: $C_\varphi : \mathscr{B} \to \mathscr{B}$ is well defined (continuous) if and only if $\varphi' \in \mathscr{B}$. Another interesting choice is $v(x) = w(x) = e^{|x|}$, for which Theorem 2 characterizes composition operators on spaces of exponentially decreasing/increasing smooth functions [4, 5]. We leave it to the reader to explicitly formulate this and other examples.

2 Statement of the Main Result

A pointwise non-decreasing sequence $V = (v_N)_{N\in\mathbb{N}}$ of positive continuous functions on $\mathbb{R}$ is called a *weight system* if $v_0 \geq 1$ and

$$\forall N\, \exists M \geq N \ : \ \sup_{x,t\in\mathbb{R},|t|\leq 1} \frac{v_N(x+t)}{v_M(x)} < \infty.$$

We shall also make use of the following condition on a weight system $V = (v_N)_{N\in\mathbb{N}}$:

$$\forall N, M\, \exists K \geq N, M \ : \ \sup_{x\in\mathbb{R}} \frac{v_N(x)v_M(x)}{v_K(x)} < \infty. \tag{1}$$

Example 1 Let $v : \mathbb{R} \to [1,\infty)$ be a continuous function satisfying

$$\sup_{x,t\in\mathbb{R},|t|\leq 1} \frac{v(x+t)}{v^N(x)} < \infty.$$

for some $N \in \mathbb{N}$ (cf. Theorem 1). Then,

$$V_v = (v^N)_{N\in\mathbb{N}}$$

is a weight system satisfying (1). □

Recall that for a positive continuous function v on $\mathbb{R}$ and $n \in \mathbb{N}$, we write $\mathscr{B}^n_v$ for the Banach space consisting of all $f \in C^n(\mathbb{R})$ such that

$$\|f\|_{v,n} = \max_{p\leq n} \sup_{x\in\mathbb{R}} \frac{|f^{(p)}(x)|}{v(x)} < \infty.$$

Let $V = (v_N)_{N\in\mathbb{N}}$ be a weight system. We shall be concerned with the following weighted spaces of smooth functions

$$\mathscr{K}_V = \varprojlim_{N\in\mathbb{N}} \mathscr{B}^N_{1/v_N},$$

$$\mathscr{O}_{C,V} = \varinjlim_{N\in\mathbb{N}} \varprojlim_{n\in\mathbb{N}} \mathscr{B}^n_{v_N},$$

$$\mathscr{O}_{M,V} = \varprojlim_{n\in\mathbb{N}} \varinjlim_{N\in\mathbb{N}} \mathscr{B}^n_{v_N}.$$

Note that $\mathscr{K}_V$ is a Fréchet space, $\mathscr{O}_{C,V}$ is an (LF)-space, and $\mathscr{O}_{M,V}$ is a (PLB)-space. Furthermore, we have the following continuous inclusions

$$\mathscr{D}(\mathbb{R}) \subset \mathscr{K}_V \subset \mathscr{O}_{C,V} \subset \mathscr{O}_{M,V} \subset C^\infty(\mathbb{R}),$$

where $\mathscr{D}(\mathbb{R})$ denotes the space of compactly supported smooth functions. The spaces $\mathscr{K}_V$ were introduced and studied by Gelfand and Shilov [6], while we refer to [7] for more information on the spaces $\mathscr{O}_{C,V}$. For $N, n \in \mathbb{N}$ fixed we will also need the following spaces

$$\mathscr{B}_{v_N} = \varprojlim_{n\in\mathbb{N}} \mathscr{B}^n_{v_N}, \qquad \mathscr{O}^n_{M,V} = \varinjlim_{N\in\mathbb{N}} \mathscr{B}^n_{v_N}.$$

The goal of this article is to show the following result.

Theorem 2 *Let $V = (v_N)_{N\in\mathbb{N}}$ and $W = (w_M)_{M\in\mathbb{N}}$ be two weight systems and let $\varphi : \mathbb{R} \to \mathbb{R}$ be smooth.*

(I) Suppose that V satisfies (1). The following statements are equivalent:

(i) $C_\varphi(\mathscr{K}_V) \subseteq \mathscr{K}_W$.
(ii) $C_\varphi : \mathscr{K}_V \to \mathscr{K}_W$ *is continuous.*
(iii) φ *satisfies the following two properties*

(a) $\forall M\ \exists N\ :\ \sup_{x\in\mathbb{R}} \frac{w_M(x)}{v_N(\varphi(x))} < \infty.$

(b) $\forall p \in \mathbb{Z}_+\ \exists N\ :\ \sup_{x\in\mathbb{R}} \frac{|\varphi^{(p)}(x)|}{v_N(\varphi(x))} < \infty.$

(II) Suppose that W satisfies (1). The following statements are equivalent:

(i) $C_\varphi(\mathscr{O}_{C,V}) \subseteq \mathscr{O}_{C,W}$.
(ii) $C_\varphi : \mathscr{O}_{C,V} \to \mathscr{O}_{C,W}$ *is continuous.*
(iii) $\forall N\ \exists M$ *such that* $C_\varphi : \mathscr{B}_{v_N} \to \mathscr{B}_{w_M}$ *is continuous.*
(iv) φ *satisfies the following two properties*

(a) $\forall N\ \exists M\ :\ \sup_{x\in\mathbb{R}} \frac{v_N(\varphi(x))}{w_M(x)} < \infty.$

(b) $\exists M\ \forall p, k \in \mathbb{Z}_+\ :\ \sup_{x\in\mathbb{R}} \frac{|\varphi^{(p)}(x)|}{w_M^{1/k}(x)} < \infty.$

(III) Suppose that W satisfies (1). The following statements are equivalent:

(i) $C_\varphi(\mathscr{O}_{M,V}) \subseteq \mathscr{O}_{M,W}$.
(ii) $C_\varphi : \mathscr{O}_{M,V} \to \mathscr{O}_{M,W}$ *is continuous.*
(iii) $C_\varphi : \mathscr{O}^n_{M,V} \to \mathscr{O}^n_{M,W}$ *is continuous for all* $n \in \mathbb{N}$.
(iv) φ *satisfies the following two properties*

(a) $\forall N\ \exists M\ :\ \sup_{x\in\mathbb{R}} \frac{v_N(\varphi(x))}{w_M(x)} < \infty.$

(b) $\forall p \in \mathbb{Z}_+\ \exists M\ :\ \sup_{x\in\mathbb{R}} \frac{|\varphi^{(p)}(x)|}{w_M(x)} < \infty.$

The proof of Theorem 2 will be given in the next section. The spaces $\mathscr{K}_v$, $\mathscr{O}_{C,v}$ and $\mathscr{O}_{M,v}$ from the introduction can be written as

$$\mathscr{K}_v = \mathscr{K}_{V_v}, \qquad \mathscr{O}_{C,v} = \mathscr{O}_{C,V_v}, \qquad \mathscr{O}_{M,v} = \mathscr{O}_{M,V_v},$$

where $V_v = (v^N)_{N\in\mathbb{N}}$ is the weight system from Example 1. Hence, Theorem 1 is a direct consequence of Theorem 2 with $V = V_v$ and $W = V_w$.

3 Proof of the Main Result

Throughout this section we fix a smooth symbol $\varphi : \mathbb{R} \to \mathbb{R}$. We need two lemmas in preparation for the proof of Theorem 2. For $n \in \mathbb{N}$ we set

$$\|f\|_n = \|f\|_{1,n} = \max_{p\le n} \sup_{x\in\mathbb{R}} |f^{(p)}(x)|.$$

Lemma 1 *Let $v, \widetilde{v}, w$ be three positive continuous functions on $\mathbb{R}$ such that*

$$C_0 = \sup_{x,t\in\mathbb{R},|t|\le 1} \frac{v(x+t)}{\widetilde{v}(x)} < \infty.$$

Let $p, n \in \mathbb{N}$ be such that

$$\|C_\varphi(f)\|_{w,p} \le C_1 \|f\|_{\widetilde{v},n}, \qquad \forall f \in \mathscr{D}(\mathbb{R}), \tag{2}$$

for some $C_1 > 0$. Then,

$$\sup_{x\in\mathbb{R}} \frac{v(\varphi(x))}{w(x)} < \infty, \tag{3}$$

and, if $p \ge 1$, also

$$\sup_{x\in\mathbb{R}} \frac{v(\varphi(x))|\varphi'(x)|^p}{w(x)} < \infty, \tag{4}$$

and

$$\sup_{x\in\mathbb{R}} \frac{v(\varphi(x))|\varphi^{(p)}(x)|}{w(x)} < \infty. \tag{5}$$

Proof Given $f \in \mathscr{D}(\mathbb{R})$ with $\operatorname{supp} f \subseteq [-1,1]$, we set $f_x = f(\cdot - \varphi(x))$ for $x \in \mathbb{R}$. Note that

$$\|f_x\|_{\widetilde{v},n} \leq \frac{C_0 \|f\|_n}{v(\varphi(x))}, \qquad x \in \mathbb{R}. \tag{6}$$

We first show (3). Choose $f \in \mathscr{D}(\mathbb{R})$ with $\operatorname{supp} f \subseteq [-1,1]$ such that $f(0) = 1$. For all $x \in \mathbb{R}$ it holds that

$$\|C_\varphi(f_x)\|_{w,p} \geq \frac{|C_\varphi(f_x)(x)|}{w(x)} = \frac{1}{w(x)}.$$

Hence, by (2) and (6), we obtain that

$$\frac{v(\varphi(x))}{w(x)} \leq C_0 C_1 \|f\|_n, \qquad \forall x \in \mathbb{R}.$$

Now assume that $p \geq 1$. We prove (4). Choose $f \in \mathscr{D}(\mathbb{R})$ with $\operatorname{supp} f \subseteq [-1,1]$ such that $f^{(j)}(0) = 0$ for $j = 1, \ldots, p-1$ and $f^{(p)}(0) = 1$. Faà di Bruno's formula implies that for all $x \in \mathbb{R}$

$$\|C_\varphi(f_x)\|_{w,p} \geq \frac{|C_\varphi(f_x)^{(p)}(x)|}{w(x)} = \frac{|\varphi'(x)|^p}{w(x)}.$$

Similarly as in the proof of (3), the result now follows from (2) and (6). Finally, we show (5). Choose $f \in \mathscr{D}(\mathbb{R})$ with $\operatorname{supp} f \subseteq [-1,1]$ such that $f'(0) = 1$ and $f^{(j)}(0) = 0$ for $j = 2, \ldots, p$. Faà di Bruno's formula implies that for all $x \in \mathbb{R}$

$$\|C_\varphi(f_x)\|_{w,p} \geq \frac{|C_\varphi(f_x)^{(p)}(x)|}{w(x)} = \frac{|\varphi^{(p)}(x)|}{w(x)}.$$

As before, the result is now a consequence of (2) and (6). □

Lemma 2 *Let v and w be positive continuous functions on $\mathbb{R}$. Then,*

(*i*) *If*

$$\sup_{x \in \mathbb{R}} \frac{v(\varphi(x))}{w(x)} < \infty,$$

then $C_\varphi : \mathscr{B}_v^0 \to \mathscr{B}_w^0$ is well-defined and continuous.

(*ii*) *Let $n \in \mathbb{Z}_+$. If*

$$\sup_{x \in \mathbb{R}} \frac{v(\varphi(x))}{w(x)} \prod_{p=1}^{n} |\varphi^{(p)}(x)|^{k_p} < \infty$$

for all $(k_1, \ldots, k_n) \in \mathbb{N}^n$ with $\sum_{j=1}^{p} j k_j \leq p$ for all $p = 1, \ldots, n$, then $C_\varphi : \mathscr{B}_v^n \to \mathscr{B}_w^n$ is well-defined and continuous.

Proof
(i) Obvious.
(ii) This is a direct consequence of (i) and Faà di Bruno's formula. □

Proof of Theorem 2 (I) $(i) \Rightarrow (ii)$: Since $C_\varphi : C^\infty(\mathbb{R}) \to C^\infty(\mathbb{R})$ is continuous, this follows from the closed graph theorem for Fréchet spaces.
$(ii) \Rightarrow (iii)$: For all $p, M \in \mathbb{N}$ there are $n, L \in \mathbb{N}$ such that

$$\|C_\varphi(f)\|_{p,1/w_M} \leq C\|f\|_{n,1/v_L}, \qquad \forall f \in \mathscr{K}_V.$$

Choose $N \geq L$ such that

$$\sup_{x,t\in\mathbb{R},|t|\leq 1} \frac{v_L(x+t)}{v_N(x)} = \sup_{x,t\in\mathbb{R},|t|\leq 1} \frac{1/v_N(x+t)}{1/v_L(x)} < \infty.$$

Lemma 1 with $w = 1/w_M$, $v = 1/v_N$ and $\widetilde{v} = 1/v_L$ yields that

$$\sup_{x\in\mathbb{R}} \frac{w_M(x)}{v_N(\varphi(x))} < \infty$$

and (recall that $w_M \geq 1$)

$$\sup_{x\in\mathbb{R}} \frac{|\varphi^{(p)}(x)|}{v_N(\varphi(x))} < \infty.$$

$(iii) \Rightarrow (i)$: As V satisfies (1), this follows from Lemma 2.
$(I I)$ $(i) \Rightarrow (ii)$: Since $C_\varphi : C^\infty(\mathbb{R}) \to C^\infty(\mathbb{R})$ is continuous, this follows from De Wilde's closed graph theorem.
$(ii) \Rightarrow (iii)$: This is a consequence of Grothendieck's factorization theorem.
$(iii) \Rightarrow (iv)$: Fix an arbitrary $N \in \mathbb{N}$. Choose $L \geq N$ such that

$$\sup_{x,t\in\mathbb{R},|t|\leq 1} \frac{v_N(x+t)}{v_L(x)} < \infty.$$

Choose $K \in \mathbb{N}$ such that $C_\varphi : \mathscr{B}_{v_L} \to \mathscr{B}_{w_K}$ is continuous. For all $m \in \mathbb{Z}_+$ there are $n \in \mathbb{Z}_+$ and $C > 0$ such that

$$\|C_\varphi(f)\|_{m,w_K} \leq C\|f\|_{n,v_L}, \qquad \forall f \in \mathscr{B}_{v_L}.$$

Lemma 1 with $w = w_K$, $v = v_N$ and $\widetilde{v} = v_L$ yields that

$$\sup_{x\in\mathbb{R}} \frac{v_N(\varphi(x))}{w_K(x)} < \infty \tag{7}$$

and (recall that $v_N \geq 1$)

$$\sup_{x\in\mathbb{R}} \frac{|\varphi'(x)|}{w_K^{1/m}(x)} < \infty \qquad \text{and} \qquad \sup_{x\in\mathbb{R}} \frac{|\varphi^{(m)}(x)|}{w_K(x)} < \infty. \tag{8}$$

Equation (7) shows (a). We now prove (b). To this end, we will make use of the following Landau-Kolmogorov type inequality due to Gorny [8]: For all $j \leq m \in \mathbb{Z}_+$ there is $C > 0$ such that

$$\|g^{(j)}\| \leq C\|g\|^{1-j/m} \left(\max\{\|g\|, \|g^{(m)}\|\}\right)^{j/m}, \qquad \forall g \in C^\infty([-1,1]), \tag{9}$$

where $\|\cdot\|$ denotes the sup-norm on $[-1,1]$. Choose $M \geq K$ such that

$$\sup_{x,t\in\mathbb{R},|t|\leq 1} \frac{w_K(x+t)}{w_M(x)} < \infty.$$

Let $p, k \in \mathbb{Z}_+$ and $x \in \mathbb{R}$ be arbitrary. Equation (8) yields that for all $m \in \mathbb{Z}_+$ there is $C > 0$ such that

$$\|\varphi'(x+\cdot)\| \leq C w_M^{1/m}(x) \qquad \text{and} \qquad \|\varphi^{(m)}(x+\cdot)\| \leq C w_M(x).$$

By applying (9) to $g = \varphi'(x+\cdot)$ and $m \geq p$ such that

$$\left(1 - \frac{p-1}{m}\right)\frac{1}{m} + \frac{p-1}{m} \leq \frac{1}{k}$$

we find that (recall that $w_M \geq 1$)

$$\begin{aligned} |\varphi^{(p)}(x)| &\leq \|\varphi^{(p)}(x+\cdot)\| \\ &\leq C\|\varphi'(x+\cdot)\|^{1-(p-1)/m} \left(\max\{\|\varphi'\|, \|\varphi^{(m+1)}\|\}\right)^{(p-1)/m} \\ &\leq C' w_M^{1/k}(x). \end{aligned}$$

$(iv) \Rightarrow (i)$: As W satisfies (1), this follows from Lemma 2.
(III) $(iii) \Rightarrow (ii) \Rightarrow (i)$: Obvious.
$(i) \Rightarrow (iv)$: Fix arbitrary $p \in \mathbb{Z}_+$ and $N \in \mathbb{N}$. Choose $L \geq N$ such that

$$\sup_{x,t\in\mathbb{R},|t|\leq 1} \frac{v_N(x+t)}{v_L(x)} < \infty.$$

Since $\mathscr{B}_{v_L} \subset \mathscr{O}_{M,V}$ and $\mathscr{O}_{M,W} \subset \mathscr{O}^p_{M,W}$, we obtain that $C_\varphi(\mathscr{B}_{v_L}) \subset \mathscr{O}^p_{M,W}$. As $C_\varphi : C^\infty(\mathbb{R}) \to C^p(\mathbb{R})$ is continuous, De Wilde's closed graph theorem implies

that $C_\varphi : \mathscr{B}_{v_L} \to \mathscr{O}^p_{M,W}$ is continuous. Grothendieck's factorization theorem yields that there is $M \in \mathbb{N}$ such that $C_\varphi : \mathscr{B}_{v_L} \to \mathscr{B}^p_{w_M}$ is well-defined and continuous, and thus that

$$\|C_\varphi(f)\|_{p,w_M} \leq C\|f\|_{n,v_L}, \qquad \forall f \in \mathscr{B}_{v_L},$$

for some $n \in \mathbb{N}$ and $C > 0$. Lemma 1 with $w = w_M$, $v = v_N$ and $\widetilde{v} = v_L$ yields that

$$\sup_{x\in\mathbb{R}} \frac{v_N(\varphi(x))}{w_M(x)} < \infty$$

and (recall that $v_N \geq 1$)

$$\sup_{x\in\mathbb{R}} \frac{|\varphi^{(p)}(x)|}{w_M(x)} < \infty.$$

$(iv) \Rightarrow (iii)$: As W satisfies (1), this follows from Lemma 2. □

Acknowledgments L. Neyt gratefully acknowledges support by FWO-Vlaanderen through the postdoctoral grant 12ZG921N.

References

1. Galbis, A., Jordá, E.: Composition operators on the Schwartz space. Rev. Mat. Iberoam. **34**, 397–412 (2018)
2. Schwartz, L.: Théorie des distributions. Hermann, Paris (1966)
3. Albanese, A.A., Jordá, E. Mele, C.: Dynamics of composition operators on function spaces defined by local and global properties. J. Math. Anal. Appl. **514**, 126303 (2022)
4. Hasumi, M.: Note on the n-dimensional tempered ultra-distributions. Tôhoku Math. J. **13**, 94–104 (1961)
5. Zieleźny, Z.: On the space of convolution operators in $\mathcal{K}'_1$. Stud. Math. **31**, 111–124 (1968)
6. Gel'fand, I.M., Shilov, G.E.: Generalized Functions. Vol. 2: Spaces of Fundamental and Generalized Functions. Academic Press, New York (1968)
7. Debrouwere, A., Vindas, J.: Topological properties of convolutor spaces via the short-time Fourier transform. Trans. Am. Math. Soc. **374**, 829–861 (2021)
8. Gorny, A.: Contribution à l'étude des fonctions dérivables d'une variable réelle. Acta Math. **71**, 317–358 (1939)

1D Hyperbolic Systems with Nonlinear Boundary Conditions II: Criteria for Finite Time Stability

Irina Kmit

Abstract We investigate the finite time stability property of one-dimensional nonautonomous initial boundary value problems for linear decoupled hyperbolic systems with nonlinear boundary conditions. We establish sufficient and necessary conditions under which continuous or L^2-generalized solutions stabilize to zero in a finite time. Our criteria are expressed in terms of a propagation operator along characteristic curves.

1 Introduction

1.1 Problem

Established in the middle of the 50th, the Finite Time Stability (FTS) concept attracts growing attention in view of its applications in control and system engineering [4, 5, 13, 14, 17, 18], output-feedback stabilization [6–8, 19], inverse problems [15, 16]), ATM networks [1], car suspension systems [2], and robot manipulators [3]. This concept is used in two ways. Quantitatively, it describes a restrained behavior of the dynamical system over a specified time interval. Qualitatively, it characterizes asymptotically stable dynamical systems whose trajectories reach an equilibrium point in a finite time. In this paper we characterize FTS hyperbolic systems using the qualitative notion of FTS.

In [10] we gave a comprehensive FTS analysis of a class of linear initial-boundary value problems with reflection boundary conditions for decoupled nonautonomous hyperbolic systems, providing algebraic and combinatorial criteria. In the

On leave from the Institute for Applied Problems of Mechanics and Mathematics, Ukrainian National Academy of Sciences

I. Kmit (✉)
Institute of Mathematics, Humboldt University of Berlin, Berlin, Germany
e-mail: irina.kmit@hu-berlin.de

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_34

autonomous setting, we provided also a spectral criterion. Asymptotic properties of solutions to perturbed FTS problems were studied in [12]. In the present paper, we establish FTS criteria for a class of nonlinear boundary value problems. These results can be applied to solving inverse problems for hyperbolic systems with FTS boundary conditions (as we demonstrate in Sect. 3.1).

Let $n \geq 2$. Our stability results concern the decoupled nonautonomous hyperbolic system

$$\partial_t u + A(x,t)\partial_x u + B(x,t)u = 0, \quad 0 < x < 1, t > 0, \tag{1}$$

where $u = (u_1, \dots, u_n)$ is a vector of real-valued functions and the diagonal matrices $A = \mathrm{diag}(a_1, \dots, a_n)$ and $B = \mathrm{diag}(b_1, \dots, b_n)$ have real entries.

Set $\Pi = \{(x,t) : 0 \leq x \leq 1,\ t \geq 0\}$. Suppose that

$$\inf_{(x,t)\in\Pi} a_j \geq a \ \text{ for all } \ j \leq m \quad \text{and} \quad \sup_{(x,t)\in\Pi} a_j \leq -a \ \text{ for all } \ j > m \tag{2}$$

for some $a > 0$ and $0 \leq m \leq n$. The system (1) is subjected to the initial conditions

$$u(x,0) = \varphi(x), \quad 0 \leq x \leq 1, \tag{3}$$

and the homogeneous nonlinear boundary conditions

$$u^{out}(t) = h(t, u^{in}(t)), \quad t \geq 0, \tag{4}$$

where $h = h(t,\xi) = (h_1(t,\xi), \dots, h_n(t,\xi))$, with $\xi \in \mathbb{R}^n$, is a real valued function,

$$h(t,0) = 0 \quad \text{for all } t \geq 0, \tag{5}$$

and

$$\begin{aligned} u^{out}(t) &= (u_1(0,t), \dots, u_m(0,t), u_{m+1}(1,t), \dots, u_n(1,t)), \\ u^{in}(t) &= (u_1(1,t), \dots, u_m(1,t), u_{m+1}(0,t), \dots, u_n(0,t)). \end{aligned}$$

1.2 Preliminaries on Continuous and L^2-Generalized Solutions

Let

$$\begin{aligned} \varphi^{out} &= (\varphi_1(0), \dots, \varphi_m(0), \varphi_{m+1}(1), \dots, \varphi_n(1)), \\ \varphi^{in} &= (\varphi_1(1), \dots, \varphi_m(1), \varphi_{m+1}(0), \dots, \varphi_n(0)). \end{aligned} \tag{6}$$

We say that a function φ satisfies the zero order compatibility conditions between (3) and (4) if

$$\varphi^{out} = h(0, \varphi^{in}). \tag{7}$$

We consider the set $C_h(\Pi)^n$ of functions $u \in C(\Pi)^n$ such that $u^{out}(0) = h(0, u^{in}(0))$. Note that, if $u \in C_h(\Pi)^n$, then $u(x, 0)$ satisfies the zero order compatibility conditions between (3) and (4) with $\varphi = u(x, 0)$. Let $C_h([0, 1])^n$ be a closed subset of a Banach space $C([0, 1])^n$ that consists of functions $\varphi \in C([0, 1])^n$ fulfilling the condition (7). Furthermore, $C_h^1([0, 1])^n = C_h([0, 1])^n \cap C^1([0, 1])^n$.

Let us introduce solution concepts, that will be used in the paper. To this end, we first define characteristics of (1) as follows. For given $j \leq n$, $x \in [0, 1]$, and $t > 0$, the j-th characteristic of (1) passing through the point $(x, t) \in \Pi$ is the solution $\omega_j(\xi) = \omega_j(\xi, x, t) : [0, 1] \to \mathbb{R}$ to the initial value problem

$$\partial_\xi \omega_j(\xi, x, t) = \frac{1}{a_j(\xi, \omega_j(\xi, x, t))}, \quad \omega_j(x, x, t) = t.$$

Let a continuous function $u : \Pi \to \mathbb{R}^n$ be continuously differentiable in Π excepting at most a countable number of characteristic curves of (1). If u satisfies (1), (3), and (4) in Π except the aforementioned characteristic curves, then it is called a *piecewise continuously differentiable solution* to the problem (1), (3), (4).

If the initial function φ is sufficiently smooth, then using integration along characteristics, we can transform the problem (1), (3), (4) to a system of integral equations. The characteristic curve $\tau = \omega_j(\xi, x, t)$ reaches the boundary of Π in two points with distinct ordinates. Let $x_j(x, t)$ denote the abscissa of that point whose ordinate is smaller. Note that the value of $x_j(x, t)$ does not depend on x and t if $t > 1/a$, where $a > 0$ satisfies (2). More precisely, if $t > 1/a$, then

$$x_j(x, t) = x_j = \begin{cases} 0 \text{ if } 1 \leq j \leq m \\ 1 \text{ if } m < j \leq n. \end{cases}$$

Set

$$c_j(\xi, x, t) = \exp \int_x^\xi \left(\frac{b_j}{a_j} \right) (\eta, \omega_j(\eta, x, t))\, d\eta.$$

Define a linear operator $S : C(R_+)^n \to C(\Pi)^n$ by

$$[Sv]_j(x, t) = c_j(x_j(x, t), x, t) v_j(\omega_j(x_j(x, t), x, t)), \quad j \leq n,$$

and a nonlinear operator $R : C(\Pi)^n \to C(R_+)^n$ by

$$[Ru]_j(t) = h_j(t, u^{in}(t)), \quad j \leq n.$$

As it follows from the method of characteristics, any piecewise continuously differentiable solution u to the problem (1), (3), (4) satisfies the following system of functional equations:

$$u_j(x,t) = [Qu]_j(x,t) \tag{8}$$

where the affine operator $Q : D(Q) \subset C_h(\Pi)^n \to C_h(\Pi)^n$ is defined by

$$[Qu]_j(x,t) = \begin{cases} [SRu]_j\,(x,t) & \text{if } x_j(x,t) = 0 \text{ or } x_j(x,t) = 1 \\ c_j(x_j(x,t),x,t)\varphi_j(x_j(x,t)) & \text{if } x_j(x,t) \in (0,1), \end{cases} \tag{9}$$

and

$$D(Q) = \{u \in C_h(\Pi)^n \,:\, u(x,0) = \varphi(x)\}.$$

Note that the definition of Q depends on the choice of the function φ. We will write $Q = Q_\varphi$ when we want to specify this dependence explicitly.

Vice versa, if a C-map $u : \Pi \to \mathbb{R}^n$ is piecewise continuously differentiable excepting at most a countable number of characteristic curves of (1) and satisfies (8) pointwise, then it is a piecewise continuously differentiable solution to (1), (3), (4). This motivates the following definition.

Definition 1 A continuous function $u : \Pi \to \mathbb{R}^n$ satisfying (8) in Π is called a *continuous solution* to (1), (3), and (4).

For a Banach space X, the n-th Cartesian power X^n is considered to be a Banach space of vectors $u = (u_1, \dots, u_n)$ normed by $\|u\|_{X^n} = \max_{i \le n} \|u_i\|_X$. Let $\|\cdot\|_{\max} = \max_{jk} |m_{jk}|$ denote the max-matrix norm of $M = (m_{jk})$ in the space of matrices M_n.

Below we will use our result from [9, Theorem 3.1] about the existence and uniqueness of global regular solutions.

Theorem 1 *Let the condition (2) be fulfilled. Moreover, assume that*

$$\begin{array}{l} \textit{for all } j,k \le n \textit{ the functions } a_j, b_j, \textit{ and } h_j \\ \textit{are continuously differentiable in all their arguments} \end{array} \tag{10}$$

and for each $T > 0$ there exists a positive real $C(T)$ and a polynomial H such that

$$\left\{ \left\| \nabla_\xi h(t,\xi) \right\|_{max} \,:\, 0 \le t \le T,\ \xi \in \mathbb{R}^n \right\} \le C(T)\,(\log\log H(\|\xi\|))^{1/4}\,. \tag{11}$$

Then the following is true.

1. *For every $\varphi \in C_h([0,1])^n$, the problem (1), (3), (4) has a unique continuous solution in Π.*

2. *For every $\varphi \in C_h^1([0,1])^n$, the problem (1), (3), (4) has a unique piecewise continuously differentiable solution in* Π.

We now define an L^2-generalized solution to the problem (1), (3), (4) similarly to [11, Definition 2].

Definition 2 Assume that the conditions of Theorem 1 are fulfilled. Let $\varphi \in L^2(0,1)^n$. A function $u \in C\left([0,\infty), L^2(0,1)\right)^n$ is called an *L^2-generalized solution* to the problem (1), (3), (4) if, for any sequence $\varphi^l \in C_h^1([0,1])^n$ with φ^l converging to φ in $L^2(0,1)^n$, the sequence of piecewise continuously differentiable solutions $u^l(x,t)$ to the problem (1), (3), (4) with φ replaced by φ^l fulfills the convergence condition

$$\|u^l(\cdot,t) - u(\cdot,t)\|_{L^2(0,1)^n} \to 0 \quad \text{as} \quad l \to \infty, \tag{12}$$

uniformly in t varying in the range $0 \le t \le T$, for each $T > 0$.

Here the norm in $L^2(0,1)^n$ is defined as usual by $\|u\|^2_{L^2(0,1)^n} = \int_0^1 (u,u)\,dx = \int_0^1 \sum_{i=1}^n u_i^2\,dx$, where $(\cdot,\cdot)$ here and below denotes the scalar product in $\mathbb{R}^n$.

The following existence and uniqueness result is obtained in [11, Theorem 2].

Theorem 2 *Let the conditions (2), (5), and (10) be fulfilled. Moreover, assume that for each $T > 0$ there exists a positive real $C(T)$ such that*

$$\sup\left\{\|\nabla_\xi h(t,\xi)\|_{\max} \,:\, 0 \le t \le T,\ \xi \in \mathbb{R}^n\right\} \le C(T). \tag{13}$$

Then, for every $\varphi \in L^2(0,1)^n$, the problem (1), (3), (4) has a unique L^2-generalized solution.

1.3 Our Results

If the problem (1), (3), (4), (11) has an L^2-generalized solution, then it is unique just by Definition 2. If this problem has a continuous solution, it is also unique as shown in [9] (see the proof of [9, Theorem 3.1]).

Definition 3 Assume that, for every $\varphi \in L^2(0,1)^n$ (resp., $\varphi \in C_h([0,1])^n$), the problem (1), (3), (4), (11) has an L^2-generalized solution (resp., a continuous solution). We say that this problem is *Finite Time Stabilizable (FTS)* if there exists a positive real T such that, for every $\varphi \in L^2(0,1)^n$ (resp., $\varphi \in C_h([0,1])^n$), the L^2-generalized solution (resp., a continuous solution) is a constant zero function for $t > T$. The infimum of all T with the above property is called the *optimal stabilization time* and is denoted by T_{opt}.

Since the operator Q operates with functions on shifted domains and, thus, captures propagation from the boundary $\partial\Pi$ into the domain Π, the stabilization

properties heavily depend on the powers of the operator Q. We start with a useful property of the operator Q. Given $T > 0$, set $\Pi^T = \{(x,t) \in \Pi : t \le T\}$.

Theorem 3 *For every $T > 0$ there exists $k \in \mathbb{N}$ such that the following is true. If, for $w \in C_h(\Pi)^n$, the problem (1), (3), (4), (11) with $\varphi(x) = w(x,0)$ has a unique continuous solution u in Π, then $u(x,t) = [Q^k w](x,t)$ in Π^T where $Q = Q_\varphi$ for $\varphi(x) = w(x,0)$.*

Now we formulate our stabilization criterion in the nonautonomous setting.

Theorem 4 *Let the condition (5) be fulfilled. Assume that, for every $\varphi \in L^2(0,1)^n$ (resp., $\varphi \in C_h([0,1])^n$), the problem (1), (3), (4), (11) has an L^2-generalized solution (resp., a continuous solution). Then this problem is FTS if and only if*

$$\begin{aligned} &\text{there is } T > 0 \text{ and } k \in \mathbb{N} \text{ such that, for all } w \in C_h(\Pi)^n \text{ and } x \in [0,1], \\ &\left[Q^k w\right](x,T) \equiv 0 \text{ where } Q = Q_\varphi \text{ for } \varphi(x) = w(x,0). \end{aligned} \tag{14}$$

In the autonomous setting a stabilization criterion is formulated in a stronger form.

Theorem 5 *Assume that the coefficient matrices A and B do not depend on t and the boundary function h does not explicitely depend on t, that is, $h(t,\xi) \equiv h(\xi)$. Moreover, let the condition (5) be fulfilled. Assume also that, for every $\varphi \in L^2(0,1)^n$ (resp., $\varphi \in C_h([0,1])^n$), the problem (1), (3), (4), (11) has an L^2-generalized solution (resp., a continuous solution). Then this problem is FTS if and only if*

$$\begin{aligned} &\text{there is } T > 0 \text{ and } q \in \mathbb{N} \text{ such that, for all } k \in \mathbb{N},\ w \in C_h(\Pi)^n,\ \text{and } x \in [0,1], \\ &\left[Q^{kq} w\right](x,kT) = 0 \text{ where } Q = Q_\varphi \text{ for } \varphi(x) = w(x,0). \end{aligned} \tag{15}$$

Theorems 3–5 assume the existence of L^2-generalized or continuous solutions (recall that those are always unique). While some sufficient conditions for the existence of solutions to the problem (1), (3), (4), (11) are given in Theorems 1 and 2, we want to emphasize that Theorems 3–5 are not restricted to these particular conditions and are more general.

The rest of the paper is organized as follows. The FTS-criteria of Theorems 4 and 5 are proved in Sect. 2. Discussion of our stabilization criteria are provided in Sect. 3, where we also show how our Theorem 3 can be applied to solving inverse hyperbolic problems.

2 Stabilization Criteria

2.1 *Proof of Theorem 3*

Fix an arbitrary $T > 0$. Since Q is a down-shift operator along characteristic curves up to the boundary of Π in the direction of time decrease, there exists an integer $q = q(T)$ such that all iterations of the operator Q starting from the q-th iteration stabilize, namely for every $w \in C_h(\Pi)^n$ it holds in Π^T that

$$\left[Q^q w\right](x,t) = \left[Q^{q+1} w\right](x,t), \tag{16}$$

where in the definition (9) of the operator Q we set $\varphi(x) = w(x,0)$.

Fix a function $w \in C_h(\Pi)^n$ fulfilling the conditions of Theorem 3. Then the problem (1), (3), (4), (11) with $\varphi = w(x,0)$ has a unique continuous solution. Set $u = Q^q w$. Hence, $u \in C_h(\Pi)^n$, and (16) implies that in Π^T we have

$$[Qu](x,t) = [Q^{q+1} w](x,t) = [Q^q w](x,t) = u(x,t).$$

It follows that the function $u = Q^q w$ is the continuous solution in Π^T to the problem (1), (3), (4), (11) with $\varphi = w(x,0)$. The proof of Theorem 3 is complete.

2.2 *Nonautonomous Case: Proof of Theorem 4*

Sufficiency Let $T > 0$ and $k \in \mathbb{N}$ be numbers satisfying the condition (14). Fix an arbitrary $\varphi \in L^2(0,1)^n$. Suppose that the problem (1), (3), (4), (11) has a unique L^2-generalized solution u.

First note that $C_h^1([0,1])^n$ is densely embedded into $L^2(0,1)^n$. Indeed, since the boundary conditions (4) are homogeneous (see 5), $C_0^\infty([0,1])^n$ is a subset of $C_h^1([0,1])^n$. As ususal, by $C_0^\infty([0,1])$ we denote a subspace of $C^\infty([0,1])$ that consists of functions having support within $(0,1)$. Now, we fix an arbitrary sequence $\varphi^l \in C_h^1([0,1])^n$ such that φ^l converges to φ in $L^2(0,1)^n$ and let $u^l(x,t)$ be the piecewise continuously differentiable solution to the problem (1), (3), (4), (11) with φ replaced by φ^l (see Theorem 1).

By Definition 2, the sequence $u^l(x,t)$ converges as in (12). Using integration along characteristics, we see that

$$u^l(x,t) = [Qu^l](x,t) \quad \text{for all } x \in [0,1] \text{ and } t \in [0,T].$$

This means that the function $u^l(x,t)$ is a fixed point of the operator Q and, hence, of any power of Q. Combining this with the condition (14), we conclude that

$$u^l(x,T) = \left[Q^k u^l\right](x,T) = 0 \quad \text{for all } x \in [0,1] \text{ and } l \in \mathbb{N}.$$

Since the initial boundary value problem (1), (4), (11) with the zero initial data at $t = T$ has a unique piecewise continuously differentiable solution for $t \geq T$ (see Theorem 1), we conclude that $u^l \equiv 0$ for $t \geq T$. The identity $u \equiv 0$ for $t > T$ follows from the convergence (12). The FTS property is therewith proved.

If the problem (1), (3), (4), (11) has a unique continuous solution, the proof goes along the same lines as above with obvious simlifications.

Necessity Consider first the case when the problem (1), (3), (4), (11) is FTS and all L^2-generalized solutions stabilize to zero in a finite time. Fix an arbitrary $T > T_{opt}$ and an integer $q = q(T)$ fulfilling the condition (16) in Π^T. Fix an arbitrary $w \in C_h(\Pi)^n$ and put $\varphi(x) = w(x,0) \in C_h([0,1])$. Then, by assumption, the problem (1), (3), (4), (11) has a unique L^2-generalized solution. Moreover, as $\varphi \in C_h([0,1])$, then by Theorem 1, this problem has a unique continuous solution. We, therefore, fall into the conditions of Theorem 3. As shown in the proof of Theorem 3, the function $u = Q^q w \in C_h(\Pi)^n$ is a continuous solution in Π^T to the problem (1), (3), (4), (11). Since any continuous solution is an L^2-generalized solution, then using the FTS property for the L^2-generalized solutions, we conclude that $[Q^q w](x,T) = 0$ for all $x \in [0,1]$, as desired.

If the problem (1), (3), (4), (11) is FTS and all continuous solutions stabilize to zero in a finite time, the argument is similar and even simpler than in the case we considered.

The proof of Theorem 4 is complete.

2.3 *Autonomous Case: Proof of Theorem 5*

Sufficiency Since the condition (15) implies (14), this part immediately follows from the sufficiency part of Theorem 4.

Necessity Consider two cases.

Case 1: the problem (1), (3), (4), (11) is FTS and all continuous solutions stabilize to zero in a finite time. Fix $T > T_{opt}$ and $q \in \mathbb{N}$ fulfilling both the condition (14) with $k = q$ and the equality (16) in Π^{2T}. For any continuous solution u we have

$$0 = \left[Q^q u\right](x,t) = \left[(SR)^q u\right](x,t) \quad \text{for all } x \in [0,1] \tag{17}$$

and for all $t \geq T$, where the second equality can be proved as follows. We first prove that this equality is fulfilled for all $t \in [T, 2T]$. By the way of

contradiction, assume that this is not true for some continuous solution u. Then there exist $x \in [0,1]$, $t \in [T, 2T]$, and $j \le n$ such that the value $[Q^q u]_j(x,t)$ can be expressed in terms of the values of u at points lying on the initial axis. Straightforward calculations show that there exist positive integers $q_1, \dots, q_n$ as well as C^1-functions $F : \mathbb{R}^{q_1+\dots+q_n} \mapsto \mathbb{R}$ and $\widetilde{F} : \mathbb{R}^{q_1} \times \dots \times \mathbb{R}^{q_n} \mapsto \mathbb{R}$, and pairwise distinct reals $x_{sr} \in [0,1]$ such that

$$[Q^q u]_j(x,t) = \widetilde{F}(\bar{v}_1^u, \dots, \bar{v}_n^u), \tag{18}$$

where

$$\begin{aligned}&\widetilde{F}(\bar{v}_1^u, \dots, \bar{v}_n^u)\\ &\quad = F(v_1^u, v_2^u, \dots, v_{q_1}^u, v_{q_1+1}^u, \dots, v_{q_1+q_2}^u, v_{q_1+q_2+1}^u, \dots, v_{q_1+\dots+q_n}^u)\end{aligned}$$

and the vector-function $\bar{v}_s^u$ for all $s \le n$ is given by

$$\bar{v}_s^u = \left(v_{q_1+q_2+\dots+q_{s-1}+1}^u, \dots, v_{q_1+q_2+\dots+q_s}^u\right) = \left(u_s(x_{s1}, 0), \dots, u_s(x_{sq_s}, 0)\right). \tag{19}$$

Since u is a solution, we have $\varphi(x) = u(x,0)$. It follows that $\widetilde{F}$ is a composition of two homogeneous operators, namely the multiplication-shift operator S and the nonlinear boundary operator R. This implies that $\widetilde{F}(0, \dots, 0) = 0$. Note that, due to (16) in Π^{2T}, the representation (18) is unique.
Equality (16) considered in Π^{2T} implies that $u(x,t) = [Q^q u](x,t)$. Combined with (18), this gives the equality

$$\begin{aligned}u_j(x,t) &= [Q^q u]_j(x,t) = \widetilde{F}(\bar{v}_1^u, \dots, \bar{v}_n^u) = \widetilde{F}(\bar{v}_1^u, \dots, \bar{v}_n^u) - \widetilde{F}(0, \dots, 0)\\ &= \sum_{i=1}^{q_1+\dots+q_n} v_i \int_0^1 \partial_i F(\gamma v_1^u, \gamma v_2^u, \dots, \gamma v_{q_1+\dots+q_n}^u)\, d\gamma,\end{aligned} \tag{20}$$

where ∂_i here and in what follows denotes the partial derivative with respect to the i-th argument. Define

$$\begin{aligned}I = \Bigg\{(s,r) \in \mathbb{N}^2 \,:\, 1 \le s \le n,\ 1 + \sum_{j=1}^{s-1} q_j \le r \le \sum_{j=1}^{s} q_j,\\ \int_0^1 \partial_r F(\gamma v_1^u, \gamma v_2^u, \dots, \gamma v_{q_1+\dots+q_n}^u)\, d\gamma \ne 0\Bigg\},\end{aligned}$$

where the sum over the empty set equals zero. Note that the set I is not empty, for else the representation (18)–(19) is impossible and we immediately get a contradiction to our assumption. Then, for an arbitrarily fixed $(s_0, r_0) \in I$, one

can choose the initial function φ such that $\varphi_{s_0}\left(x_{s_0r_0}\right) \neq 0$ while $\varphi_s\left(x_{sr}\right) = 0$ for all other $(s, r) \in I$. On account of (19), the equality (20) now reads

$$u_j(x,t) = \varphi_{s_0}(x_{s_0r_0}) \int_0^1 \partial_{r_0} F(\gamma v_1^u, \gamma v_2^u, \ldots, \gamma v_{q_1+\cdots+q_n}^u)\, d\gamma \neq 0,$$

contradicting the FTS property of our problem. We, therefore, proved that the condition (17) is true for all $t \in [T, 2T]$.

Now we show that (17) is true for all $t \geq 2T$. To this end, observe that in the autonomous case the following formulas are true:

$$\begin{aligned} &\omega_j(\xi, x, t+T) = \omega_j(\xi, x, t) + T, \quad t \geq 0, \\ &[Sv]_j(x,t) = c_j(x_j, x, t) v_j(\omega_j(x_j, x, T) + t - T), \quad t \geq T, \end{aligned} \tag{21}$$

for all $v \in C(\mathbb{R}_+)^n$. Given $w \in C_h(\Pi)^n$, set $z(x,t) = w(x, t+T)$. It follows that

$$\left[(SR)^q z\right](x,t) = \left[(SR)^q w\right](x, t+T), \quad t \geq T. \tag{22}$$

Using the above argument for (17) for $t \in [T, 2T]$ once again, we see that $T > T_{opt} > 1/a$. On account of (21), we then have $\omega_j(x_j(x,t), x, t) = \omega_j(x_j(x,t), x, t-T) + T > T$ for all $t > 2T$, $x \in [0,1]$, and $j \leq n$. Combining this with the FTS property, we conclude that $u(\cdot, t) = [Qu](\cdot, t) = [SRu](\cdot, t) \equiv 0$ for all $t > 2T$. Summarizing, the condition (17) stays true for all $t \geq T$, as desired.

Let q be now chosen such that (17) holds for $t \geq T$ and, additionally, the equality (16) is fulfilled in Π^{3T}. Let $w \in C_h(\Pi)^n$ be arbitrarily fixed. Similarly to the proof of Theorem 3, the function $[Q^q w](x,t)$ is a continuous solution to (1), (3), (4), (11) with $\varphi(x) = w(x,0)$ in the domain Π^{3T}. By (17), we have $[Q^q w](\cdot, T) \equiv 0$ and, hence the function $z^1(x,t) = [Q^q w](x, t+T) = [(SR)^q w](x, t+T)$ belongs to $C_h(\Pi)^n$ and is a continuous solution to (1), (3), (4), (11) with $\varphi(x) = 0$ in Π^{2T}. It follows from (17) that

$$0 = \left[Q^q z^1\right](x,t) = \left[(SR)^q z^1\right](x,t) \quad \text{for } t \in [T, 2T].$$

Similarly to (22), we have

$$\left[(SR)^q z^1\right](x,t) = \left[(SR)^q Q^q w\right](x, t+T) = \left[Q^{2q} w\right](x, t+T).$$

Therefore, $\left[Q^{2q} w\right](\cdot, t) \equiv 0$ for $t \in [2T, 3T]$. In the next step we set $z^2(x,t) = \left[Q^{2q} w\right](x, t+2T)$. Due to the previous step, $z^2(\cdot, 0) \equiv 0$ and, therefore, z^2 belongs

to $C_h(\Pi)^n$ and is a continuous solution to (1), (3), (4), (11) with $\varphi(x) = 0$ in Π^{2T}. Similarly, for $t \in [T, 2T]$, it holds

$$0 = \left[Q^q z^2\right](x,t) = \left[(SR)^q z^2\right](x,t) = \left[(SR)^q Q^{2q} w\right](x, t+2T)$$
$$= \left[Q^{3q} w\right](x, t+2T)$$

and, hence $\left[Q^{3q} w\right](\cdot, t) \equiv 0$ for $t \in [3T, 4T]$. Proceeding further by induction, where on the k-th step we set $z^k(x,t) = \left[Q^{kq} w\right](x, t+kT)$, $k \geq 3$, we conclude that the desired condition (15) is true. The proof of Case 1 is therewith complete.

Case 2: the problem (1), (3), (4), (11) is FTS and all L^2-generalized solutions stabilize to zero in a finite time. Let q be as in Case 1. Using the same argument as in the proof of the necessity part of Theorem 4 in the same L^2-case, fix an arbitrary $w \in C_h(\Pi)^n$, put $\varphi(x) = w(x,0)$, and conclude that the function $u = Q^q w \in C_h(\Pi)^n$ is a continuous solution to the problem (1), (3), (4), (11) in the domain Π^{3T}. Since any continuous solution is an L^2-generalized solution, then using the FTS property for the L^2-generalized solutions and (17), we conclude that $[Q^q w](x,T) = 0$ for all $x \in [0,1]$. The proof is completed by repeating the argument used at the end of Case 1.

3 Examples

3.1 Solving Inverse Problems

Let the boundary conditions (4) be linear, namely

$$u^{out}(t) = P u^{in}(t), \quad t \geq 0, \tag{23}$$

where $P = (p_{jk})$ is an $n \times n$-matrix with constant entries. We assume that the matrix $P_{abs} = (|p_{jk}|)$ is nilpotent. Then, due to [10, Theorem 1.10], the problem (1), (3), (23) is robust FTS, with respect to perturbations of the coefficients a_j and b_j.

Fix an arbitrary $r > 0$ and consider the following abstract setting of the autonomous problem (1), (3), (23) on $L^2(0,1)^n$ (as studied, e.g., in [15, 16]):

$$\frac{d}{dt}u(t) = \mathcal{A}u(t) + f, \quad (0 \leq t \leq r) \tag{24}$$
$$u(0) = u_0, \quad u(r) = u_r, \tag{25}$$

where the operator $\mathcal{A} : D(\mathcal{A}) \subset L^2(0,1)^n \to L^2(0,1)^n$ is defined by

$$(\mathcal{A}v)(x) = -A(x)v' - B(x)v,$$
$$D(\mathcal{A}) = \left\{ v \in L^2(0,1)^n \,:\, v' \in L^2(0,1)^n,\ v^{out} = Pv^{in} \right\},$$

and $u_0, u_r \in D(\mathcal{A})$ are known functions. Here v^{out}, v^{in} are defined similarly to (6). Solving the inverse problem (24)–(25), we are looking for a couple of functions (u, f) such that $u \in C^1([0,r], L^2(0,1))^n$, $u(t) \in D(\mathcal{A})$ for all $t \in [0,r]$, and $f \in L^2(0,1)^n$.

Since the problem (24)–(25) is autonomous, then, due to [12, Theorem 2.3], the operator $\mathcal{A}$ generates a C_0-semigroup $S(t)$. Since the problem (24)–(25) is FTS, the semigroup $S(t)$ is nilpotent. Hence, there exists $T > 0$ such that $S(t) = 0$ for all $t \geq T$. Accordingly to [16, Theorem 4], for any $u_0, u_r \in D(\mathcal{A})$, there is a unique function $f \in L^2(0,1)^n$ solving the inverse problem (24)–(25). Moreover, this function admits the representation

$$f = \begin{cases} -Au_r & \text{if } r \geq T \\ -Au_r + A\displaystyle\sum_{k=1}^{n_0} S(kr)(u_0 - u_r) & \text{if } r < T, \end{cases}$$

where $n_0 = \lceil T/r \rceil - 1$. Recall that $\lceil x \rceil$ denotes the integer nearest to x from above. The unknown function $u(t)$ is then given by the formula

$$u(t) = S(t)u_0 + \int_0^t S(s) f \, ds, \quad 0 \leq t \leq r.$$

Now, using Theorem 3, we conclude that there exists $k = k(T) \in \mathbb{N}$ such that for all $x \in [0,1]$ it holds that

$$[S(t)u_0](x) = \begin{cases} \left[Q^k w\right](x,t) & \text{if } t \leq T \\ 0 & \text{if } t > T, \end{cases}$$

the formula being true for any $w \in C_h(\Pi)^n$ such that $w(x,0) = u_0(x)$.

3.2 Nonlinear Boundary Conditions and FTS Property

In the domain Π we consider the 2×2-decoupled system

$$\partial_t u_1 + \partial_x u_1 = 0, \quad \partial_t u_2 - \partial_x u_2 = 0 \tag{26}$$

with the nonlinear boundary conditions

$$u_1(0,t) = r(t)\sin(u_2(0,t)), \quad u_2(1,t) = \sin^2(s(t)u_1(1,t)) \tag{27}$$

and the initial conditions

$$u_1(x,0) = \varphi_1(x), \quad u_2(x,0) = \varphi_2(x). \tag{28}$$

Here r and s are smooth and uniformly bounded functions for $t \geq 0$. Note that the boundary conditions are of the type (13). Our aim is, using Theorem 4, to find conditions on the functions r and s such that the problem (26)–(28) is FTS.

The operator Q defined by (9) is now specified to

$$[Qu]_1(x,t) = \begin{cases} \varphi_1(x-t) & \text{if } x > t \\ r(t-x)\sin(u_2(0,t-x)) & \text{if } t-x \geq 0, \end{cases}$$

$$[Qu]_2(x,t) = \begin{cases} \varphi_2(x+t) & \text{if } t+x < 1 \\ \sin^2(s(t+x-1)u_1(1,t+x-1)) & \text{if } t+x \geq 1. \end{cases}$$

The second power of Q is then given by

$$[Q^2u]_1(x,t) = \begin{cases} \varphi_1(x-t) & \text{if } x > t \\ r(t-x)\sin(\varphi_2(t-x)) & \text{if } 0 \leq t-x < 1 \\ r(t-x)\sin\left(\sin^2(s(t-x-1)u_1(1,t-x-1))\right) & \text{if } 1 \leq t-x, \end{cases}$$

$$\left[Q^2u\right]_2(x,t) = \begin{cases} \varphi_2(x+t) & \text{if } t+x < 1 \\ \sin^2\left(s(t+x-1)\varphi_1(2-(t+x))\right) & \text{if } 1 \leq t+x < 2 \\ \sin^2\left(s(t+x-1)r(t+x-2)\sin(u_2(0,t+x-2))\right) & \text{if } 2 \leq t+x. \end{cases}$$

It follows that if there exist reals $T_1 > 0$ and $T_2 > 0$ with

$$T_2 - T_1 \geq 1 \quad \text{and} \quad \Big(r(t) = 0 \text{ and } s(t) = 0 \text{ for } T_1 \leq t \leq T_2\Big), \tag{29}$$

then the condition (14) is true with $k = 1$. If there exist reals $T_1 > 0$ and $T_2 > 0$ with

$$T_2 - T_1 \geq 2 \quad \text{and} \quad \Big(r(t) = 0 \text{ or } s(t) = 0 \text{ for } T_1 \leq t \leq T_2\Big), \tag{30}$$

then the condition (14) is true with $k = 2$. In other words, (29) and (30) are two sufficient conditions for the problem (26)–(28) to be FTS.

3.3 *Theorem 4 Does not Extend for Nonhomogeneous Boundary Conditions*

In the domain Π, we consider the 2×2-decoupled system (26) with the initial conditions (28) and the boundary conditions

$$u_1(0,t) = g(t), \quad u_2(1,t) = u_1(1,t). \tag{31}$$

Fix g to be a smooth bounded function such that

$$g(t) = \begin{cases} 0 & \text{if } 0 \le t \le 4 \\ \neq 0 & \text{if } 4 < t. \end{cases}$$

The formula (9) then reads

$$[Qu]_1(x,t) = \begin{cases} \varphi_1(x-t) \text{ if } x > t \\ g(t-x) \text{ if } t - x \ge 0, \end{cases}$$

$$[Qu]_2(x,t) = \begin{cases} \varphi_2(x+t) \text{ if } t + x < 1 \\ u_1(1, t+x-1) \text{ if } t + x \ge 1, \end{cases}$$

implying that

$$\left[Q^2 u\right]_1 (x,t) = [Qu]_1 (x,t),$$

$$\left[Q^2 u\right]_2 (x,t) = \begin{cases} \varphi_2(x+t) & \text{if } t + x < 1 \\ \varphi_1(2-(t+x)) & \text{if } 1 \le t + x < 2 \\ g(t+x-2) & \text{if } 2 \le t + x. \end{cases}$$

It follows that $\left[Q^2 u\right](x,3) \equiv 0$, while the problem (26), (28), (31) is not FTS.

Acknowledgments Irina Kmit was supported by the VolkswagenStiftung Project "From Modeling and Analysis to Approximation".

References

1. Amato, F., Ariola, M., Abdallah, C., Cosentino, C.: Application of finite-time stability concepts to the control of ATM networks. Proc. Allerton Conf. Commun. Control Comput. **40(2)**, 1071–1079 (2002)
2. Amato, F., Ambrosiano, R., Ariola, M., Cosentino, C., Tommasi, G.D., et al.: Finite-Time Stability and Control. Springer, Berlin (2014)
3. Orlov, Y.: Finite time stability and robust control synthesis of uncertain switched systems. SIAM J. Control Optim. **43**(4), 1253–1271 (2004)

4. Bastin, G., Coron, J.M.: Stability and Boundary Stabilization of 1-d Hyperbolic Systems. Progress in Nonlinear Differential Equations and Their Applications, vol. 88. Birkhäser, Basel (2016)
5. Gugat, M.: Optimal Boundary Control and Boundary Stabilization of Hyperbolic Systems. Birkhäser, Basel (2015)
6. He, W., Ge, S.Z.: Robust adaptive boundary control of a vibrating string under unknown time-varying disturbance. IEEE Trans. Control Syst. Technol. **20**(1), 48–58 (2012)
7. He, W., Ge, S.Z., Zhang, S.: Adaptive boundary control of a flexible marine installation system. Automatica **47**(12), 2728–2734 (2011)
8. Karafyllis, I., Krstic, M.: Input-to-State Stability for PDEs. Springer, Berlin (2019)
9. Kmit, I.: Classical solvability of nonlinear initial-boundary problems for first-order hyperbolic systems. Int. J. Dyn. Syst. Differ. Equ. **1**(3), 191–195 (2008)
10. Kmit, I., Lyul'ko, N.: Finite time stabilization of nonautonomous first-order hyperbolic systems. SIAM J. Control Optim. **59**(5), 3179–3202 (2021)
11. Lyul'ko, N.: 1D hyperbolic systems with nonlinear boundary conditions, I: L^2-generalized solutions. This collection: Appl. Comput. Res. Persp. (2023), https://doi.org/10.1007/978-3-031-36375-7_35
12. Kmit, I., Lyul'ko, N.: Perturbations of superstable linear hyperbolic systems. J. Math. Anal. Appl. **460**(2), 838–862 (2018)
13. Pavel, L.: Classical solutions in Sobolev spaces for a class of hyperbolic Lotka-Volterra systems. SIAM J. Control Optim. **51**(3), 2132–2151 (2013)
14. Perruquetti, A., Barbot, J.P.: Sliding Mode Control in Engineering. M. Dekker, New York (2002)
15. Prilepko, A.I., Orlovsky, D.G., Vasin I.A.: Methods for Solving Inverse Problems in Matthematical Physics. Taylor & Francis, Boca Raton (2000)
16. Tikhonov I., Son Tung, V.N.: The solvability of the inverse problem for the evolution equation with a superstable semigroup. RUDN J. MIPh. **26**(2), 103–118 (2018)
17. Udwadia, F.E.: Boundary control, quiet boundaries, super-stability and super-instability. Appl. Math. Comput. **164**(2), 327–349 (2005)
18. Udwadia, F.E.: On the longitudinal vibrations of a bar with viscous boundaries: super-stability, super-instability and loss damping. Int. J. Eng. Sci. **50**(1), 79–100 (2012)
19. Xu, G.Q.: Stabilization of string system with linear boundary feedback. Nonlinear Anal. Hybrid Syst. **1**, 383–397 (2007)

1D Hyperbolic Systems with Nonlinear Boundary Conditions I: L^2-Generalized Solutions

Natalya Lyul'ko

Abstract We consider 1D nonautonomous initial boundary value problems for general linear first-order hyperbolic systems with nonlinear boundary conditions. For initial L^2-data, we prove existence and uniqueness of L^2-generalized solutions if the nonlinearities are Lipschitz continuous.

1 Our Setting and Results

Following the general idea of [9, §29], the notion of an L^2-generalized solution of initial-boundary value problems for linear first-order hyperbolic systems with linear boundary conditions was introduced in [8]. A number of interesting properties of such solutions, such as smoothing property, asymptotic stability, and finite time stabilization were investigated in [7] and [8] for wide classes of hyperbolic problems. Studying L^2-generalized solutions, it is natural to focus on their qualitative properties as we do not need to take into account compatibility conditions, which would be necessary in the case of classical solutions.

A generalized solution concept depends usually on the regularity properties of the coefficients as well as on the initial and the boundary data, cf. [4]. For linear hyperbolic systems of the first order, a natural generalized solution concept can be introduced in the form of integrable function satisfying a system of integral equations obtained by multiplying the differential system by test functions [2]. Another way to define a generalized solution is to see it as a function satisfying an integral system of equations obtained by integration along the characteristics of a hyperbolic system (see [1], [5, 6]). The definition of an L^2-generalized solution in [8] relies on the method of continuation of smooth solutions, analogous to that used in [3, 9], and [5] for linear hyperbolic systems. In the present paper we extend our

N. Lyul'ko (✉)
Sobolev Institute of Mathematics, Russian Academy of Sciences and Novosibirsk State University, Novosibirsk, Russia
e-mail: natlyl@mail.ru

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_35

results obtained in [8] for linear hyperbolic systems with linear boundary conditions to the case of nonlinear boundary conditions.

Let $n \geq 2$. We consider the following nonautonomous hyperbolic system:

$$\partial_t u + A(x,t)\partial_x u + B(x,t)u = f(x,t), \quad 0 < x < 1,\ t > 0, \tag{1}$$

where $u = (u_1, \ldots, u_n)$ and $f = (f_1, \ldots, f_n)$ are vectors of real-valued functions, and the diagonal matrix $A = \operatorname{diag}(a_1, \ldots, a_n)$ and the $n \times n$-matrix $B = (b_{jk})$ have real entries. Set

$$\Pi = \{(x,t) : 0 \leq x \leq 1,\ t \geq 0\}.$$

Suppose that

$$\inf_{(x,t)\in\Pi} a_j \geq a \ \text{ for all } \ j \leq m \quad \text{and} \quad \sup_{(x,t)\in\Pi} a_j \leq -a \ \text{ for all } \ j > m \tag{2}$$

for some $a > 0$ and $0 \leq m \leq n$. The system (1) is endowed with the initial conditions

$$u(x,0) = \varphi(x), \quad 0 \leq x \leq 1, \tag{3}$$

and the nonlinear boundary conditions

$$u^{out}(t) = h(t, u^{in}(t)), \quad t \geq 0, \tag{4}$$

where $h = h(t,\xi) = (h_1(t,\xi), \ldots, h_n(t,\xi))$ with $\xi \in \mathbb{R}^n$ is a real valued function, and

$$\begin{aligned} u^{out}(t) &= (u_1(0,t), \ldots, u_m(0,t), u_{m+1}(1,t), \ldots, u_n(1,t)), \\ u^{in}(t) &= (u_1(1,t), \ldots, u_m(1,t), u_{m+1}(0,t), \ldots, u_n(0,t)). \end{aligned}$$

We assume that

$$\begin{aligned} &\text{for all } j,k \leq n \text{ the functions } a_j,\ f_j,\ b_{jk}, \text{ and } h_j \\ &\text{are continuously differentiable in all their arguments.} \end{aligned} \tag{5}$$

Let

$$\begin{aligned} \varphi^{out} &= (\varphi_1(0), \ldots, \varphi_m(0), \varphi_{m+1}(1), \ldots, \varphi_n(1)), \\ \varphi^{in} &= (\varphi_1(1), \ldots, \varphi_m(1), \varphi_{m+1}(0), \ldots, \varphi_n(0)). \end{aligned}$$

We say that a function φ satisfies the zero order compatibility conditions between (3) and (4) if

$$\varphi^{out} = h(0, \varphi^{in}). \tag{6}$$

For a Banach space X, the n-th Cartesian power X^n is considered to be a Banach space of vectors $u = (u_1, \dots, u_n)$ with $u_i \in X$ normed by $\|u\|_{X^n} = \max_{i \le n} \|u_i\|_X$. If $X = L^2(0,1)$ is a real valued Hilbert space, then the norm in X^n is defined in a usual way as

$$\|u\|^2_{L^2(0,1)^n} = \int_0^1 (u,u)\,dx = \int_0^1 \sum_{i=1}^n u_i^2\,dx,$$

where $(\cdot,\cdot)$ here and below denotes the scalar product in $\mathbb{R}^n$.

We also consider the set $C_h(\Pi)^n$ of functions $u \in C(\Pi)^n$ such that

$$u^{out}(0) = h(0, u^{in}(0)).$$

Note that, if $u \in C_h(\Pi)^n$, then $u(x,0)$ satisfies the zero order compatibility conditions between (3) and (4) with $\varphi = u(x,0)$. Let $C_h([0,1])^n$ be a closed subset of a Banach space $C([0,1])^n$ that consists of functions $\varphi \in C([0,1])^n$ fulfilling the condition (6). Furthermore, $C^1_h([0,1])^n = C_h([0,1])^n \cap C^1([0,1])^n$.

We now introduce two solution concepts exploited below, for piecewise continuously differentiable and L^2-generalized solutions.

As usual, for given $j \le n$, $x \in [0,1]$, and $t > 0$, we define the j-th characteristic of (1) passing through the point $(x,t) \in \Pi$ as the solution $\omega_j(\cdot, x, t) : [0,1] \to \mathbb{R}$ to the problem

$$\partial_\xi \omega_j(\xi, x, t) = \frac{1}{a_j(\xi, \omega_j(\xi, x, t))}, \quad \omega_j(x, x, t) = t.$$

Definition 1 Let a continuous function $u : \Pi \to \mathbb{R}^n$ be continuously differentiable in Π excepting at most a countable number of characteristic curves of (1). If u satisfies (1), (3), and (4) in Π except the aforementioned characteristic curves, then it is called a *piecewise continuously differentiable solution* to the problem (1), (3), (4).

The derivatives of a piecewise continuously differentiable solution restricted to a compact subset of Π have at most a finite number of discontinuities (of first order) on certain characteristic curves.

Let $\|\cdot\|_{\max} = \max_{jk} |m_{jk}|$ denote the max-matrix norm of $M = (m_{jk})$ in the space of matrices M_n. The following existence and uniqueness result is proved in [6, Theorem 3.1].

Theorem 1 *Let the conditions (2) and (5) be fulfilled. Moreover, assume that for each $T > 0$ there exists a positive real $C(T)$ such that*

$$\sup \left\{ \|\nabla_\xi h(t,\xi)\|_{\max} \,:\, 0 \le t \le T,\ \xi \in \mathbb{R}^n \right\} \le C(T). \tag{7}$$

Then, for every $\varphi \in C^1_h([0,1])^n$, the problem (1), (3), (4) has a unique piecewise continuously differentiable solution in Π.

For a function h in (4), we will suppose that

$$C_h^1([0,1])^n \text{ is densely embedded into } L^2(0,1)^n. \tag{8}$$

Note that, if the boundary conditions (4) are homogeneous, namely $h(t,0) = 0$ for all $t \geq 0$, then the condition (8) is fulfilled automatically, since in this case $C_0^\infty([0,1])^n$ is a subset of $C_h^1([0,1])^n$. By $C_0^\infty([0,1])$ we denote a subspace of $C^\infty([0,1])$ that consists of functions having support within $(0,1)$.

Analogously to [8, Definition 4.3], we introduce a notion of the L^2-generalized solution.

Definition 2 Let $\varphi \in L^2(0,1)^n$ and the conditions of Theorem 1 be fulfilled. A function $u \in C\left([0,\infty), L^2(0,1)\right)^n$ is called an *L^2-generalized solution* to the problem (1), (3), (4) if, for any sequence $\varphi^l \in C_h^1([0,1])^n$ with φ^l converging to φ in $L^2(0,1)^n$, the sequence of piecewise continuously differentiable solutions $u^l(x,t)$ to the problem (1), (3), (4) with φ replaced by φ^l fulfills the convergence condition

$$\|u(\cdot,t) - u^l(\cdot,t)\|_{L^2(0,1)^n} \to 0 \quad \text{as} \quad l \to \infty, \tag{9}$$

uniformly in t varying in the range $0 \leq t \leq T$, for each $T > 0$.

In this paper we prove that the problem (1), (3), (4) has a unique L^2-generalized solution for every L^2-initial function φ whenever the boundary function $h(t,\xi)$ is Lipschitz continuous in ξ.

Theorem 2 *Let the conditions (2), (5), (7), and (8) be fulfilled. Then, for every $\varphi \in L^2(0,1)^n$, the problem (1), (3), (4) has a unique L^2-generalized solution.*

Note that, if the condition (7) is fulfilled, then the existence of a continuous and piecewise continuously differentiable solution to the problem (1), (3), (4) follows from Theorem 1.

The uniqueness of the generalized solution follows from Definition 2.

2 Proof of Theorem 2

Let $\varphi \in L^2(0,1)^n$. Due to (8), there exists a sequence $\varphi^l \in C_h^1([0,1])^n$ with $\varphi^l \to \varphi$ in $L^2(0,1)^n$. By the definition of $C_h^1([0,1])^n$, the functions φ^l satisfy the zero order compatibility conditions (6). By Theorem 1, for every $l \geq 1$, the problem (1), (3), (4) with φ replaced by φ^l has a unique piecewise continuously differentiable solution, say u^l. According to Definition 2, we have to show that the sequence u^l converges to an L^2-generalized solution to the problem (1), (3), (4) in the sense of (9). This will be done if we show that for any $T > 0$ there exists a positive constant $K(T)$ such that for all $\varphi, \bar{\varphi} \in C_h^1[0,1]^n$ the piecewise continuously differentiable solutions u

and $\bar{u}$ to the problem (1), (3), (4) with the initial functions φ and $\bar{\varphi}$, respectively, satisfy the estimate

$$\|u(\cdot,t)-\bar{u}(\cdot,t)\|_{L^2(0,1)^n} \le K(T)\|\varphi-\bar{\varphi}\|_{L^2(0,1)^n} \quad \text{for all } t\in[0,T]. \tag{10}$$

To prove (10), we exploit a general approach described in [3]. Let $T>0$ be arbitrary fixed and $\Pi^T=\{(x,t): 0\le x\le 1, 0\le t\le T\}$. Due to the definition of a piecewise continuously differentiable solution, the function $u-\bar{u}$ satisfies the system

$$\partial_t(u-\bar{u})+A(x,t)\partial_x(u-\bar{u})+B(x,t)(u-\bar{u})=0. \tag{11}$$

almost everywhere in Π^T. Taking a scalar product of (11) with $u-\bar{u}$ and integrating the resulting equality over the domain Π^t for $t\le T$, we get the equality

$$\int\int_{\Pi^t}\left[\frac{\partial}{\partial\theta}(u-\bar{u},u-\bar{u})+\frac{\partial}{\partial x}\left(A\,(u-\bar{u}),u-\bar{u}\right)\right]dxd\theta$$
$$=\int\int_{\Pi^t}\Big((\partial_x A-2B)\,(u-\bar{u}),u-\bar{u}\Big)dxd\theta.$$

Applying Green's formula to the left hand side, we conclude that

$$I(t)+\int_0^t G(\theta)\,d\theta=\|\varphi-\bar{\varphi}\|^2_{L^2(0,1)^n}$$
$$+\int\int_{\Pi^t}\Big((\partial_x A-2B)\,(u-\bar{u}),u-\bar{u}\Big)dx\,d\theta, \tag{12}$$

where

$$I(t)=\|u(\cdot,t)-\bar{u}(\cdot,t)\|^2_{L^2(0,1)^n},$$
$$G(\theta)=\sum_{j=1}^n\Big[a_j(1,\theta)(u_j(1,\theta)-\bar{u}_j(1,\theta))^2-a_j(0,\theta)(u_j(0,\theta)-\bar{u}_j(0,\theta))^2\Big].$$

The following upper bound for the second summand in the right hand side of (12) is obvious:

$$\int\int_{\Pi^t}\Big|\Big((\partial_x A-2B)\,(u-\bar{u}),u-\bar{u}\Big)\Big|dx\,d\theta\le d(T)\int_0^t I(\theta)\,d\theta, \tag{13}$$

where $d(T)=n\max\{||(\partial_x A-2B)||_{max}\ :\ (x,t)\in\Pi^T\}$.

On account of (4), the function $G(\theta)$ admits the representation

$$\begin{aligned} G(\theta) = &\sum_{j=1}^{m} a_j(1,\theta)\left[u_j(1,\theta) - \bar{u}_j(1,\theta)\right]^2 \\ &+ \sum_{j=m+1}^{n} a_j(1,\theta)\left[h_j(\theta, u^{in}(\theta)) - h_j(\theta, \bar{u}^{in}(\theta))\right]^2 \\ &- \sum_{j=1}^{m} a_j(0,\theta)\left[h_j(\theta, u^{in}(\theta)) - h_j(\theta, \bar{u}^{in}(\theta))\right]^2 \\ &- \sum_{j=m+1}^{n} a_j(0,\theta)\left[u_j(0,\theta) - \bar{u}_j(0,\theta)\right]^2. \end{aligned} \tag{14}$$

Let

$$\alpha_j(t) = \begin{cases} a_j(1,t) \text{ if } 1 \le j \le m \\ -a_j(0,t) \text{ if } m < j \le n, \end{cases} \qquad \beta_j(t) = \begin{cases} -a_j(0,t) \text{ if } 1 \le j \le m \\ a_j(1,t) \text{ if } m < j \le n. \end{cases} \tag{15}$$

Due to (2), there exists positive reals α_*, α^*, and β_* such that for all $j \le n$ and $t \ge 0$ it holds

$$\alpha_* \le \alpha_j(t) \le \alpha^*, \quad -\beta_* \le \beta_j(t) \le -\alpha_*. \tag{16}$$

Using the notation (15), the formula (14) reads

$$G(\theta) = \sum_{j=1}^{n} \alpha_j(\theta)(u_j^{in} - \bar{u}_j^{in})^2 + \sum_{j=1}^{n} \beta_j(\theta)\left[h_j(\theta, u^{in}(\theta)) - h_j(\theta, \bar{u}^{in}(\theta))\right]^2.$$

By the mean value theorem, for every $\theta > 0$ there exists a real number $\eta(\theta)$ such that $0 < \eta(\theta) < 1$ and

$$\begin{aligned} G(\theta) = &\sum_{j=1}^{n} \alpha_j(\theta)(u_j^{in} - \bar{u}_j^{in})^2 \\ &+ \sum_{j=1}^{n} \beta_j(\theta)\Big(\nabla_\xi h_j(\theta, u^{in}(\theta) + \eta(\theta)\bar{u}^{in}(\theta)), u^{in}(\theta) - \bar{u}^{in}(\theta)\Big)^2. \end{aligned} \tag{17}$$

Note that $G(\theta)$ is a quadratic form with respect to the vector-function $u^{in} - \bar{u}^{in}$. Assume first that this quadratic form is nonnegative for $0 \le \theta \le T$, that is the boundary conditions (4) are dissipative. Then, combining (12) and (13), we arrive at the inequality

$$I(t) \le \|\varphi - \bar{\varphi}\|^2_{L^2(0,1)^n} + d(T)\int_0^t I(\theta)\, d\theta.$$

Applying Gronwall's argument, we obtain the desired inequality (10) with the constant $K(T) = e^{Td(T)/2}$. Note that the constant $K(T)$ depends on A, B, and h, but not on the initial data φ and $\bar{\varphi}$.

In the rest of the proof we consider the case that the form is not nonnegative and show that it can be made nonnegative by appropriately changing the unknown functions (and, hence, the same argument as above applies). Let $\mu_j(x,t)$ be arbitrary smooth functions satisfying the conditions

$$\inf_{\Pi^T} |\mu_j| > 0 \text{ and } \sup_{\Pi^T} |\mu_j| < \infty \quad \text{for all } j \le n.$$

Changing of each variable u_j to $v_j = \mu_j u_j$, we bring the system (1) to

$$\partial_t v_j + a_j(x,t)\partial_x v_j - \frac{\partial_t \mu_j + a_j \partial_x \mu_j}{\mu_j} v_j + \sum_{k=1}^{n} b_{jk} \frac{\mu_j}{\mu_k} v_k = \mu_j f_j, \quad j \le n, \tag{18}$$

the initial conditions (3) to

$$v_j(x,0) = \mu_j(x,0)\varphi_j(x), \quad j \le n, \tag{19}$$

and the boundary conditions (4) to

$$v^{out}(t) = \widetilde{h}(t, v^{in}(t)), \tag{20}$$

where for $\zeta \in \mathbb{R}^n$

$$\widetilde{h}_j(t,\zeta) = \mu_j(x_j,t) h_j\left(t, \frac{\zeta_1}{\mu_1(1,t)}, \dots, \frac{\zeta_m}{\mu_m(1,t)}, \frac{\zeta_{m+1}}{\mu_{m+1}(0,t)}, \dots, \frac{\zeta_n}{\mu_n(0,t)}\right) \tag{21}$$

and

$$x_j = \begin{cases} 0 \text{ if } 1 \le j \le m \\ 1 \text{ if } m < j \le n. \end{cases}$$

Changing each variable $\bar{u}_j$ to $\bar{v}_j = \mu_j \bar{u}_j$, we get the system (18)–(20) with $v = (v_1, \dots, v_n)$ replaced by $\bar{v} = (\bar{v}_1, \dots, \bar{v}_n)$. The new system (18)–(20) is again of the form (1), (3), (4). Therefore, the equality (12) with $u - \bar{u}$ replaced by $v - \bar{v}$ reads

$$I(t) + \int_0^t G(\theta)\, d\theta = \|\Phi\|^2_{L^2(0,1)^n} + \int\int_{\Pi^t} \left((\partial_x A - 2(\widetilde{B} - M))(v - \bar{v}), v - \bar{v}\right) dx d\theta,$$

where

$$\Phi(x) = (\mu_1(x,0)(\varphi_1 - \bar{\varphi}_1), \ldots, \mu_n(x,0)(\varphi_n - \bar{\varphi}_n))$$

and

$$\begin{aligned}
\widetilde{B}(x,t) &= \left(b_{jk}\tfrac{\mu_j}{\mu_k}\right)_{j,k=1}^{n},\\
M(x,t) &= \operatorname{diag}\left(\frac{\partial_t\mu_1 + a_1\partial_x\mu_1}{\mu_1}, \ldots, \frac{\partial_t\mu_n + a_n\partial_x\mu_n}{\mu_n}\right),\\
I(t) &= \|v(\cdot,t) - \bar{v}(\cdot,t)\|^2_{L^2(0,1)^n},\\
G(\theta) &= \sum_{j=1}^{n}\alpha_j(\theta)\left(v_j^{in} - \bar{v}_j^{in}\right)^2\\
&\quad + \sum_{j=1}^{n}\beta_j(\theta)\left(\nabla_\zeta\widetilde{h}_j(\theta, v^{in}(\theta) + \widetilde{\eta}(\theta)\bar{v}^{in}(\theta)), v^{in}(\theta) - \bar{v}^{in}(\theta)\right)^2,
\end{aligned} \tag{22}$$

and $0 < \widetilde{\eta}(\theta) < 1$ for all $\theta \in [0,T]$.

Therefore, our objective is reduced to show that there exist smooth functions $\mu_j(x,t)$, $j \le n$, such that the quadratic form in the formula (22) for $G(\theta)$ is nonnegative for all $0 \le \theta \le T$. To this end, introduce the vector-function $y(\theta) = (y_1(\theta), \ldots y_n(\theta)) \equiv v^{in}(\theta) + \widetilde{\eta}(\theta)\bar{v}^{in}(\theta)$. Taking into account (21), for $j \le n$ we compute

$$\begin{aligned}
&\left(\nabla_\zeta\widetilde{h}_j(\theta, y(\theta)), v^{in}(\theta) - \bar{v}^{in}(\theta)\right)^2 =\\
&= \left[\mu_j(x_j,\theta)\left\{\partial_{\xi_1}h_j\left(\theta, \frac{y_1(\theta)}{\mu_1(1,\theta)}, \ldots, \frac{y_m(\theta)}{\mu_m(1,\theta)}, \frac{y_{m+1}(\theta)}{\mu_{m+1}(0,\theta)}, \ldots, \frac{y_n(\theta)}{\mu_n(0,\theta)}\right)\right.\right.\\
&\quad \frac{v_1^{in} - \bar{v}_1^{in}}{\mu_1(1,\theta)} + \cdots + \partial_{\xi_m}h_j\\
&\quad \left(\theta, \frac{y_1(\theta)}{\mu_1(1,\theta)}, \ldots, \frac{y_m(\theta)}{\mu_m(1,\theta)}, \frac{y_{m+1}(\theta)}{\mu_{m+1}(0,\theta)}, \ldots, \frac{y_n(\theta)}{\mu_n(0,\theta)}\right)\times\\
&\quad \frac{v_m^{in} - \bar{v}_m^{in}}{\mu_m(1,\theta)} + \partial_{\xi_{m+1}}h_j\\
&\quad \left(\theta, \frac{y_1(\theta)}{\mu_1(1,\theta)}, \ldots, \frac{y_m(\theta)}{\mu_m(1,\theta)}, \frac{y_{m+1}(\theta)}{\mu_{m+1}(0,\theta)}, \ldots, \frac{y_n(\theta)}{\mu_n(0,\theta)}\right)\frac{v_{m+1}^{in} - \bar{v}_{m+1}^{in}}{\mu_{m+1}(0,\theta)}\times\\
&\quad + \cdots + \partial_{\xi_n}h_j\left(\theta, \frac{y_1(\theta)}{\mu_1(1,\theta)}, \ldots, \frac{y_m(\theta)}{\mu_m(1,\theta)}, \frac{y_{m+1}(\theta)}{\mu_{m+1}(0,\theta)}, \ldots, \frac{y_n(\theta)}{\mu_n(0,\theta)}\right)\\
&\quad \left.\left.\frac{v_n^{in} - \bar{v}_n^{in}}{\mu_n(0,\theta)}\right\}\right]^2.
\end{aligned} \tag{23}$$

Set

$$\mu = \max_{0\le t\le T}\left\{\max_{1\le j\le m}|\mu_j(0,t)|,\ \max_{m<j\le n}|\mu_j(1,t)|\right\},$$

$$\nu = \min_{0\le t\le T}\left\{\min_{1\le j\le m}|\mu_j(1,t)|,\ \min_{m<j\le n}|\mu_j(0,t)|\right\}.$$

Taking into account (16), (22), and (23), we get

$$\begin{aligned}&\sum_{j=1}^{n}\alpha_j(\theta)(v_j^{in}-\bar v_j^{in})^2+\sum_{j=1}^{n}\beta_j(\theta)\\&\quad\times\Big(\nabla_\zeta\widetilde h_j(\theta,v^{in}(\theta)+\widetilde\eta(\theta)\bar v^{in}(\theta)),v^{in}(\theta)-\bar v^{in}(\theta)\Big)^2\\&\ge\alpha^*\|v^{in}-\bar v^{in}\|^2_{\mathbb{R}^n}-\frac{n\beta^*\mu}{\nu}\max_{t\le T,\xi,j}\|\nabla_\xi h_j(t,\xi)\|^2_{\mathbb{R}^n}\|v^{in}-\bar v^{in}\|^2_{\mathbb{R}^n}.\end{aligned}$$

Finally, we choose smooth functions μ_j such that

$$\frac{\mu}{\nu}<\frac{\alpha^*}{\beta^*(nC(T))^2},$$

where the constant $C(T)$ is defined in (7). The quadratic form becomes nonnegative and the proof of the theorem is complete.

Acknowledgments Natalya Lyul'ko was supported by the state contract of the Sobolev Institute of Mathematics, Project No. FWNF-2022-0008.

References

1. Abolinya, V.E., Myshkis, A.D.: A mixed problem for an almost linear hyperbolic system on the plane. Matematicheskij Sbornik. **50**(4), 423–442 (1960)
2. Abolinya, V.E., Myshkis, A.D.: On a mixed problem for a linear hyperbolic system on a plane (in Russian). Scholar. Not. Latv. Univ. **20**(3), 87–104 (1958)
3. Godunov, S.K.: Equations of Mathematical Physics, 2nd edn (in Russian). Nauka, Moscow (1979)
4. Haller, S., Hörmann, G.: Comparison of some solution concepts for linear first-order hyperbolic differential equations with non-smooth coefficients. Publ. Inst. Math. **84**(98), 123–157 (2008)
5. Ëltysheva, N.A.: On qualitative properties of solutions to some hyperbolic systems on the plane. Matematicheskij Sbornik **135**(2), 186–209 (1988)
6. Kmit, I.: Classical solvability of nonlinear initial-boundary problems for first-order hyperbolic systems. Int. J. Dynam. Syst. Differ. Equ. **1**(3), 191–195 (2008)
7. Kmit, I., Lyul'ko, N.: Finite time stabilization of nonautonomous first-order hyperbolic systems. SIAM J. Control Optim. **59**(5), 3179–3202 (2021)
8. Kmit, I., Lyul'ko, N.: Perturbations of superstable linear hyperbolic systems. J. Math. Anal. Appl. **460**(2), 838–862 (2018)
9. Vladimirov, V.: Equations of Mathematical Physics. Monograph and Textbooks in Pure and Applied Mathematics. M. Dekker, New York (1971)

On Classification of Semigroups Associated to Levy Processes

Irina V. Melnikova and Vadim A. Bovkun

Abstract The work is devoted to the study of properties of operator semigroups with kernels that are, in the general case, random process generalized densities of transition probabilities. The semigroup technique and the technique of the generalized Fourier transform underlie the classification of the generators of these semigroups.

1 Introduction

The study of numerous phenomena and processes, taking into account random disturbances that arise in various fields of natural science, social and biosystems, leads to models that can be described in terms of stochastic differential equations. At the present stage of research, along with taking into account "continuous" random perturbations formalized with help of Wiener processes, it becomes necessary to take into account "discontinuous", in particular jump-like, random perturbations. The most suitable random processes that allow reflecting various types of random perturbations are Levy processes.

Levy processes form an important subclass of time-homogeneous Markov processes. Homogeneity and the Markov property allow describe their behavior at time $t > 0$ using the transition probability $P(0, x; t, B)$ and language of operator theory. Namely, each Levy process corresponds to a transition semigroup of operators $\{U(t),\ t \geq 0\}$, defined on the space $C_0(\mathbb{R}^n)$, continuous functions tending to zero at infinity:

$$U(t)f(x) = \int_{\mathbb{R}^n} f(y)P(0, x; t, dy) = \langle f(\cdot),\, p(0, x; t, \cdot)\rangle. \tag{1}$$

I. V. Melnikova (✉) · V. A. Bovkun
Ural Federal University, Ekaterinburg, Russia
e-mail: Irina.Melnikova@urfu.ru; Vadim.Bovkun@urfu.ru

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_36

Here $p(0, x; t, y)$ is the (generalized) density of transition probability $P(0, x; t, B)$. The indicated connection of Levy processes with semigroups of a special form and their generators turns out to be productive in both directions. On the one hand, it allows to find various probabilistic characteristics of these processes as solutions to deterministic problems, using methods for solving partial differential equations and pseudo-differential equations. On the other hand, it allows one to find exact or approximate solutions to deterministic problems containing generators of these processes with help of a probabilistic interpretation of the solution.

The paper is devoted to the study of the possibility inclusion of Cauchy problems containing generators of Levy processes into extension of Gelfand–Shilov classification and semigroup classification.

Section 2 specifies properties of semigroups corresponding to Levy processes. Examples of semigroups corresponding to basic Levy processes are given. Based on the technique of the generalized Fourier transform, it is shown how symbols of their generators, which in the general case are pseudo-differential operators, are arranged.

Section 3 is devoted to the embedding of Levy semigroups with zero Levy measure in the scheme of connections between the Gelfand-Shilov classification and the semigroup classification. It is important to note here that the Gelfand-Shilov classification based on the generalized Fourier transform takes place for Cauchy problems with differential operators, while the semigroup classification, defined in terms of the spectral properties of generators, takes place for Cauchy problems with a wider class of operators.

Section 4 constructs a partial extension of the Gelfand–Shilov classification to the case of problems with generators containing integral terms corresponding to Levy processes. It is shown that problems with Levy process generators are Petrovsky correct in the constructed extension.

2 Levy Processes, Their Generators, and Associated Semigroups

Let's start with the definition of Levy processes. Let $\mathscr{B}(\mathbb{R}^n)$ be the Borel σ-algebra of sets in $\mathbb{R}^n$.

Definition ([1]) Let the probability space $(\Omega, \mathscr{F}, \mathscr{F}_t, \mathbb{P})$ be given. The Levy process $X = \{X(t), t \geq 0\}$ is a random process taking values in the measurable space $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$ and satisfies the following conditions:

- $X(0) = 0$ a.s.;
- is homogeneous in time : $\mathbb{P}((X(s+t) - X(s)) \in B), s, t \geq 0, B \in \mathscr{B}(\mathbb{R}^n)$, does not depend on s;
- is stochastically continuous : for any $\varepsilon > 0$ $\mathbb{P}(|X(s+t) - X(s)| > \varepsilon) \to 0$ at $t \to 0$.

There is a modification for Levy processes whose trajectories are a.s. continuous on the right and have finite limits on the left (see, e.g., [2]). In this paper, we will assume that Levy processes have trajectories with the indicated property.

Levy processes form a subclass of Markov processes. The process X taking values in the measurable space $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$, is Markov if for any $f \in B_b(\mathbb{R}^n)$ and $0 \le s \le t$ the equality [1] holds:

$$\mathbb{E}[f(X(t))\big|\mathscr{F}_s] = \mathbb{E}[f(X(t))\big|X(s)].$$

The key characteristic of such processes is the transition probability $P(s, x; t, B) := \mathbb{P}(X(t) \in B | X(s) = x)$, the probability that at time $t \ge s$ the process X is at an arbitrary point y of $B \in \mathscr{B}(\mathbb{R}^n)$ if at time s it was at x. For the transition probabilities of Markov processes, we have the Kolmogorov–Chapman equality (see, e.g., [1]):

$$P(s, x; t, B) = \int_{\mathbb{R}^n} P(s, x; r, dy) P(r, y; t, B), \quad 0 \le s \le r \le t, \ x \in \mathbb{R}^n.$$

For time-homogeneous processes $P(s, x; t, B) = P(0, x; t - s, B)$.

Due to the Kolmogorov theorem the homogeneous Markov process X, up to the distribution of $X(0)$, is determined by the set of transition probabilities $P(0, x; \tau, B)$, $\tau \ge 0$. Consequently, for such processes, along with the transition probability, the key characteristic is the family of the form (1), which, by virtue of the Kolmogorov–Chapman equation, have the semigroup property: $U(t + s) = U(t)U(s)$, $t, s \ge 0$.

Among homogeneous Markov processes, a subclass of processes is singled out whose semigroups map $C_0(\mathbb{R}^n)$ into itself and are strongly continuous on this space. Such processes are called Feller processes, and the corresponding semigroups are called Feller semigroups. Adding independence of increments to these properties allows to single out Levy processes among Feller processes. This leads to the fact that, along with the homogeneity in time, Levy processes have the property of spatial homogeneity, i.e. the transition probability of the Levy process is invariant under the shift of spatial variables (see, e.g., [3]):

$$P(s, x; t, B) = P(s, 0; t, B - x), \quad 0 \le s \le t, \ x \in \mathbb{R}^n.$$

As examples, consider semigroups and generators associated with basic processes: shift, Wiener, and Poisson. For clarity, we will consider $\mathbb{R}$-valued processes.

1. The semigroup $\{U_1(t), t \ge 0\}$ associated with the shift process $\{X_1(t), t \ge 0\}$, where $X_1(t) := x + bt$, $b \in \mathbb{R}$.

The shift process is a deterministic process, it has the properties of stochastic continuity, temporal and spatial homogeneity. Therefore, the equality

$$P_1(0, x; t, (-\infty; y)) =: P_1(0, x; t, y) = P_1(0, 0; t, y - x)$$

and $X_1 - x$ is the Levy process. On the basis of these properties, for the shift process X_1 one can set the generalized transition probability density[1] in the following way:

$$p_1(0, x; t, y) = \delta_{x+bt}(y) = \delta(y - (x + bt)).$$

Then the corresponding to X_1 the shift semigroup has the form

$$U_1(t)f(x) = \langle f(\cdot), p_1(0, x; t, \cdot)\rangle = \langle f, \delta_{x+bt}\rangle = f(x + bt).$$

It is easy to verify that the semigroup is strongly continuous on $C_0(\mathbb{R})$, and its generator is the operator $A_1 = b\frac{\partial}{\partial x}$.

2. Semigroup $\{U_2(t), t \geq 0\}$ corresponding to the Wiener process $\{X_2(t), t \geq 0\}$ with transition probability density

$$p_2(0, x; t, y) = \frac{1}{a\sqrt{2\pi t}} e^{-\frac{(x-y)^2}{2a^2 t}},$$

defined as follows:

$$U_2(t)f(x) = \frac{1}{a\sqrt{2\pi t}} \int_{\mathbb{R}} f(y) e^{-\frac{(x-y)^2}{2a^2 t}} dy.$$

This semigroup is strongly continuous on $C_0(\mathbb{R})$, and its generator is $A_2 = \frac{a^2}{2}\frac{\partial^2}{\partial x^2}$.

3. Consider a Poisson process $\{X(t), t \geq 0\}$ with jumps q, intensity λ, and $X(0) = 0$. Define $X_3(t) = X(t) + x$. Such a process can be specified using the transition probability

$$P_3(0, x; t, y) := P_3(0, x; t, (-\infty; y)) = \sum_{k=0}^{c_q} \frac{(\lambda t)^k}{k!} e^{-\lambda t},$$

where $c_q = \left[\frac{y-x}{q}\right]$ for $\left[\frac{y-x}{q}\right] \neq \frac{y-x}{q}$, and $c_q = \frac{y-x}{q} - 1$, otherwise. Then the generalized transition probability density is defined as follows

$$p_3(0, x; t, y) = \sum_{k=0}^{c_q} \frac{(\lambda t)^k}{k!} e^{-\lambda t} \delta_{x+kq}(y).$$

[1] The functional $p(0, x; t, \cdot)$ such that for any $f \in C_0(\mathbb{R})$, $\int_{\mathbb{R}} f(y)P(0, x; t, dy) = \langle f(\cdot), p(0, x; t, \cdot)\rangle$ is called the generalized transition probability density of the process $\{X(t), t \geq 0\}$. In the case when the transition probability has a derivative in the sense of Radon–Nikodim, $p(0, x; t, \cdot)$ is a regular generalized function.

For the semigroup corresponding to the process we get the equality

$$U_3(t)f(x) = \langle f(\cdot), p_3(0, x; t, \cdot)\rangle = \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} e^{-\lambda t} f(x + kq).$$

The semigroup U_3 is strongly continuous on $C_0(\mathbb{R})$, and by the definition of the generator for this semigroup, we obtain

$$A_3 f(x) = \lim_{t\to 0} \frac{1}{t} [U_3(t) - I] f(x) = \lambda(f(x + q) - f(x)), \quad f \in C_0(\mathbb{R}).$$

Thus, in contrast to the differential operators A_1, A_2, the generator of the semigroup U_3 is a difference operator.

Before proceeding to the next process, we recall the notation, used below, related to the characteristic function and the Fourier transform from the measure μ on $\mathscr{B}(\mathbb{R}^n)$, from $f \in L_1(\mathbb{R}^n)$, and from the distribution $g \in \mathscr{S}'(\mathbb{R}^n)$:

$$\mathscr{F}[\mu](\alpha) = \int_{\mathbb{R}^n} e^{-i(\alpha,y)} \mu(dy) = \widehat{\mu}(\alpha),$$

$$\mathscr{F}^{-1}[\mu](\alpha) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{i(\alpha,y)} \mu(dy),$$

$$\mathscr{F}[f](\alpha) = \int_{\mathbb{R}^n} e^{-i(\alpha,y)} f(y)dy = \widehat{f}(\alpha),$$

$$\mathscr{F}^{-1}[f](\alpha) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{i(\alpha,y)} f(y)dy,$$

$$\langle \varphi, \mathscr{F} g\rangle := (2\pi)^n \langle \mathscr{F}^{-1}\varphi, g\rangle, \ \varphi \in \mathscr{S}(\mathbb{R}^n).$$

For the characteristic function of the random variable ξ, defined by the probability measure μ_ξ, we will use the following notation:

$$\Phi_\xi(\alpha) := \mathbb{E}\left[e^{i(\alpha,\xi)}\right] = \int_{\mathbb{R}^n} e^{i(\alpha,y)} \mu_\xi(dy), \quad \alpha \in \mathbb{R}^n. \tag{2}$$

4. Consider the semigroup $\{U_4(t), t \geq 0\}$, corresponding to the process $\{X_4(t), t \geq 0\}$, which is defined by the equality: $X_4(t) = x + X_\pi(t)$, where $\{X_\pi(t), t \geq 0\}$ is the compound Poisson process, defined as follows. Let $\{z_k\}$ be a sequence of independent identically distributed $\mathbb{R}$-valued random variables with a common distribution μ_z and $\{N(t), t \geq 0\}$ be the standard Poisson process with intensity λ and $q = 1$. Then by definition, the process $X_\pi(t) := z_1 + \ \ldots \ + z_{N(t)}$ is a compound Poisson process with intensity λ.

In the general case, without explicitly having either a transition probability density or a transition probability for such a process, we use the technique of

characteristic functions to describe the semigroup corresponding to X_4. For the characteristic function of the random variable $X_\pi(t)$ for every fixed $t \geq 0$ we have

$$\Phi_{X_\pi(t)}(\alpha) = \sum_{k=0}^{\infty} \mathbb{E}\left(e^{i\alpha(z_1+\cdots+z_{N(t)})}\big|N(t)=k\right)\mathbb{P}(N(t)=k)$$

$$= \sum_{k=0}^{\infty} \Phi_z^k(\alpha)\frac{(t\lambda)^k}{k!}e^{-t\lambda} = e^{t\lambda(\Phi_z(\alpha)-1)}. \tag{3}$$

By virtue of the obtained equality (3) and the properties of random variables z_k, we obtain that for any $t \geq 0$ the random variable $X_\pi(t)$ is infinitely divisible. Therefore, the process $X_\pi = X_4 - x$ (as well as the processes $X_1 - x$, $X_2 - x$, $X_3 - x$) is the Levy process.

Since for X_π we have the equality $P_\pi(0,0;t,dy) = \mu_{X_\pi(t)}(dy)$, then, due to (2) and (3), in [13] the following representation for the transition semigroup of the process X_4 is obtained

$$U_4(t)f(x) = \mathscr{F}^{-1}\left[\widehat{f}(\sigma)e^{t\lambda(\Phi_z(\sigma)-1)}\right](x), \quad f \in \mathscr{S}(\mathbb{R}),$$

which can be extended to the space $C_0(\mathbb{R})$.

In order to obtain a representation for the generator of semigroup U_4, we use the connection between the generators of Levy processes and pseudo-differential operators (ΨD-operators). A ΨD-operator (on a class of functions f) is an operator of the form

$$Kf(x) = \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n} e^{i(\alpha,x)}s(x,\alpha)\hat{f}(\alpha)d\alpha, \quad x \in \mathbb{R}^n, \tag{4}$$

where the function $s = s(x,\alpha) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is called the ΨD-operator symbol. Depending on the specifics of problems solved with the help of ΨD-operators, different classes of symbols are distinguished.

On the function class $f \in \mathscr{S}(\mathbb{R}^n)$ for operators with symbols locally bounded in x and polynomially bounded in α we have

$$Kf\,(x) = \mathscr{F}^{-1}[s(x,\cdot)\widehat{f}(\cdot)](x), \quad x \in \mathbb{R}^n. \tag{5}$$

Such operators generalize differential operators $K = \sum_{k=0}^{n} a_k(x)\frac{d^k}{dx^k}$, $x \in \mathbb{R}$, with bounded variables coefficients, since the K can be represented as:

$$Kf\,(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{i\alpha x}s(x,\alpha)\widehat{f}(\alpha)d\alpha, \quad \text{where } s(x,\alpha) = \sum_{k=0}^{n} a_k(x)(i\alpha)^k.$$

Hence it follows that the semigroup generators A_1 and A_2 are ΨD-operators with power (in α) symbols, but the generators A_3 and A_4 are not differential operators. Further we will see that A_3 and A_4 belong to the class of ΨD-operators. To do this, we need a representation of the characteristic function of the Levy process, the Levy–Khinchin formula. It is known that if $\{X(t), t \geq 0\}$ is a Levy process with values in the space $\mathbb{R}^n$, then for any $t \geq 0$ the characteristic function of the random variable $X(t)$ has the form $\Phi_{X(t)}(\alpha) = e^{t\eta(\alpha)}$, where $\eta = \eta(\alpha), \alpha \in \mathbb{R}^n$, is defined by the Levy–Khinchin formula (see, e.g., [1]):

$$\eta(\alpha) = i(b, \alpha) - \frac{1}{2}(\alpha\,, Q\alpha) + \int_{\mathbb{R}^n \setminus \{0\}} \left(e^{i(\alpha, y)} - 1 - i(\alpha, y)\chi_{|y|\leq 1}(y)\right) \nu(dy). \tag{6}$$

In this equality, $b \in \mathbb{R}^n$, Q is a positive-definite symmetric $n \times n$-matrix, ν is the Levy measure on $\mathscr{B}(\mathbb{R}^n)$. The Levy triple (b, Q, ν) is uniquely determined by the process X. Moreover, the real part of the function η satisfies the inequality

$$Re(\eta(\alpha)) \leq 0, \quad \alpha \in \mathbb{R}^n, \tag{7}$$

and for $|\eta|$ a polynomial estimate holds: $|\eta(\alpha)| \leq C(1 + |\alpha|)^2$ (see, e.g., [5]).

Consider the operator semigroup U corresponding to the process $X + x$, where X is the Levy process, and write its representation in terms of the Fourier transform:

$$U(t)f(x) = \mathbb{E}(f(X(t) + x)) = \frac{1}{(2\pi)^n}\mathbb{E}\left(\int_{\mathbb{R}^n} e^{i(\alpha,\, x + X(t))}\,\hat{f}(\alpha)d\alpha\right)$$
$$= \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n} e^{i(\alpha,x)}e^{t\eta(\alpha)}\hat{f}(\alpha)d\alpha, \quad f \in \mathscr{S}(\mathbb{R}^n).$$

Further, using the representation, by the definition of the generator for arbitrary $f \in \mathscr{S}(\mathbb{R}^n)$ we have

$$Af(x) = \lim_{t\to 0}\frac{1}{t}\left[U(t) - I\right]f(x) = \frac{1}{(2\pi)^n}\lim_{t\to 0}\int_{\mathbb{R}^n} e^{i(\alpha,x)}\frac{e^{t\eta(\alpha)} - 1}{t}\hat{f}(\alpha)d\alpha$$
$$= \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n} e^{i(\alpha,x)}\eta(\alpha)\hat{f}(\alpha)d\alpha, \quad x \in \mathbb{R}^n. \tag{8}$$

By virtue of the polynomial estimate for η, the equality (8) correctly defines the operator A on $\mathscr{S}(\mathbb{R}^n)$, and the generator is a ΨD-operator on the space $\mathscr{S}(\mathbb{R}^n)$ with polynomially bounded symbol $s(x, \alpha) = \eta(\alpha)$.

Using this fact, we find the generator of the semigroup corresponding to process $\{X_4(t), t \geq 0\}$. From representation (3) for the characteristic function of $X_\pi(t), t \geq 0$, it follows that

$$\eta(\alpha) = \int_{\mathbb{R}} \lambda(e^{i\alpha\beta} - 1)\mu_z(d\beta).$$

Then, taking into account (8), using the Fubini theorem and the Fourier transform formulas for $f \in \mathscr{S}$, we obtain A_4:

$$A_4 f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{i\alpha x} \int_{\mathbb{R}} (e^{i\alpha\beta} - 1)\lambda\mu_z(d\beta)\hat{f}(\alpha)d\alpha$$
$$= \int_{\mathbb{R}} (f(x+\beta) - f(x))\,\lambda\mu_z(d\beta),$$

which can be extended to the space $C_0(\mathbb{R})$.

Next, we present a scheme for comparing the classifications of Cauchy problems, Gelfand–Shilov and semigroup classification, and indicate the place in this scheme occupied by transition semigroups corresponding to Levy processes with zero Levy measure.

3 Generators of the Levy Semigroup in the Framework of Two Classifications

Before proceeding to the construction of a scheme illustrating the relationship between the two classifications, we briefly present each of them. We start with the Gelfand–Shilov classification constructed for Cauchy problems with differential operators.

Let the Cauchy differential problem be given:

$$\frac{\partial}{\partial t}u(t,x) = \mathsf{A}\left(i\frac{\partial}{\partial x}\right)u(t,x), \quad t \geq 0,\ x \in \mathbb{R}^n, \quad u(0,x) = f(x), \tag{9}$$

where $\mathsf{A}\left(i\frac{\partial}{\partial x}\right) = \left\{\mathsf{A}_{j,k}\left(i\frac{\partial}{\partial x}\right)\right\}_{j,k=1}^{m}$, $\mathsf{A}_{j,k}\left(i\frac{\partial}{\partial x}\right)$ are linear differential operators of order at most l. The solution to the problem (9) for any fixed $x \in \mathbb{R}^n$ and $t \geq 0$ is an m-dimensional vector $u(t,x) = (u_1(t,x), \ldots, u_m(t,x))$. The approach proposed in [6] for solving this problem is based on applying the generalized Fourier transform to the problem (9) and solving the transformed Cauchy problem. In this case, the Fourier transform of $f = (f_1, \ldots, f_m)$ with $f_j \in L_1(\mathbb{R}^n)$ is defined as follows:

$$\widetilde{f} = (\widetilde{f}_1, \ldots, \widetilde{f}_m), \qquad \widetilde{f}_j(\alpha) = \int_{\mathbb{R}^n} e^{i(\alpha,x)} f_j(x)dx, \quad \alpha \in \mathbb{R}^n,\ j = 1, \ldots, m.$$

From the problem (9), we pass to the Fourier-transformed Cauchy problem, the solution of which is:

$$\widetilde{u}(t,\alpha) = e^{t\mathsf{A}(\alpha)}\widetilde{f}(\alpha), \quad \alpha \in \mathbb{R}^n,$$

where $\mathsf{A}(\alpha)$ is the matrix multiplication operator $\left\{\mathsf{A}_{j,k}(\alpha)\right\}_{j,k=1}^{m}$ with elements that are polynomials of degree at most l. The key role in studying the properties of the $\widetilde{u}$, solution to the dual problem, is played by the operator $e^{t\mathsf{A}(\alpha)}$ and its extension to the complex plane $e^{t\mathsf{A}(\gamma)}$, $\gamma = \alpha + i\tau$, $\alpha, \tau \in \mathbb{R}^n$. Based on the estimate

$$e^{t\Lambda(\gamma)} \le \left\| e^{t\mathsf{A}(\gamma)} \right\|_{\mathbb{R}^m} \le C(1+|\gamma|)^{l(m-1)} \cdot e^{t\Lambda(\gamma)}, \quad t \ge 0,$$

the classification of the Fourier-transformed problem (and, consequently, of the original one) is built in [6] according to the behavior of $\Lambda(\alpha)$, where $\Lambda(\gamma) = \max Re\, \lambda_k(\gamma)$, $\lambda_k(\gamma)$ are characteristic roots of $\mathsf{A}(\gamma)$.

The system (9) is called

- correct in the sense of Petrovsky if there exists a constant $C > 0$ such that $\Lambda(\alpha) \le C$, in particular, parabolic if

$$\exists C_1, h, C_2 > 0 : \ \Lambda(\alpha) \le -C_1|\alpha|^h + C_2$$

and hyperbolic if for any $\gamma \in \mathbb{C}^n$,

$$\exists C_1, C_2 > 0 : \ \Lambda(\gamma) \le C_1|\gamma| + C_2;$$

- conditionally correct if

$$\exists C_1, C_2 > 0, 0 < h < 1 : \ \Lambda(\alpha) \le C_1|\alpha|^h + C_2;$$

- incorrect if an evaluation with the reduced order l_0 ($l_0 \ge 1$) is performed:

$$\exists C_1, C_2 > 0 : \ \Lambda(\alpha) \le C_1|\alpha|^{l_0} + C_2,$$

but stronger bounds do not hold.

Depending on the type to which the system belongs, spaces of test and generalized functions are determined, in which the Fourier-transformed problem is correct. Further, due to the connection between the spaces found and spaces of their inverse Fourier transforms, the space of test and generalized functions is determined in which the original problem is well-posed (see, e.g., [6, 7])

Now we give a brief summary of results on the theory of semigroups and the well-posedness of the abstract Cauchy problem

$$u'(t) = Au(t), \quad t \in [0;\tau), \ \tau \le \infty, \quad u(0) = f, \tag{10}$$

classifying its solution operators in terms of the behavior of the A-resolvent. A detailed exposition can be found in [8, 9].

Let E be a Banach space, A be a closed linear operator in E. By the solution of (10) on $[0; T]$, $T < \tau$, we mean $u \in C([0; T], \mathrm{dom}\, A) \cap C^1([0; T], E)$.

1. Strongly continuous semigroups.

Let A be densely defined in E. The operator A generates a strongly continuous semigroup (C_0-semigroup) in E if and only if any of the (equivalent) conditions is satisfied:

- problem (10) is uniformly well-posed on $\operatorname{dom} A$, i.e. for any $T > 0$ and $f \in \operatorname{dom} A$
  (a) there is a unique solution on the interval $[0; T]$;
  (b) the solution is stable with respect to changes in the initial data, uniformly in $t \in [0; T]$: $\sup_{t\in[0;T]} \|u(t)\| \le C_T \|f\|$;
- the resolvent of A is defined in some right half-plane $Re\lambda > \omega$ and

$$\exists C > 0 : \left\| \mathscr{R}^{(k)}(\lambda) \right\|_{\mathscr{L}(E)} \le \frac{Ck!}{(Re\lambda - \omega)^{k+1}}, \ Re\lambda > \omega, \quad k \in \mathbb{N}_0. \tag{11}$$

2. Integrated semigroups.

Let the operator A be densely defined in E and the set of its regular points be non-empty. The operator A generates a (non-degenerate) n times integrated exponentially bounded semigroup of operators in E if and only if any of the following conditions is satisfied:

- problem (10) is uniformly (n, ω)-well-posed for $t \ge 0$, i.e. for any $T > 0$ and $f \in \operatorname{dom} A^{n+1}$ there is a unique solution on the segment $[0; T]$, stable with respect to changes in the initial data in the graph-norm of the operator A:

$$\|u(t)\| \le Ce^{\omega t} \|f\|_n, \ t \ge 0, \quad \|f\|_n := \|f\| + \|Af\| + \ldots + \|A^n f\|;$$

- $\exists C > 0, \omega \in \mathbb{R} : \left\| \dfrac{d^k}{d\lambda^k} \left(\dfrac{\mathscr{R}(\lambda)}{\lambda^n} \right) \right\|_{\mathscr{L}(X)} \le \dfrac{Ck!}{(Re\lambda - \omega)^{k+1}}, \ Re\lambda > \omega, \ k \in \mathbb{N}_0$;
- problem (10) is well-posed in the space of exponentially bounded distributions $\mathscr{S}'_\omega(E)$, which is defined as follows: $f \in \mathscr{S}'_\omega(E)$ if and only if $fe^{-\omega t} \in \mathscr{S}'(E) := \mathscr{L}(\mathscr{S}, E)$.

3. Convoluted semigroups.

Let $K(t)$, $t \ge 0$ be an exponentially bounded function and its Laplace transform satisfies the condition: $\left|\widetilde{K}(\lambda)\right| = \mathscr{O}\left(e^{-M(\kappa|\lambda|)}\right)$ for $|\lambda| \to \infty$, where $M(\xi)$ is a positive function of the variable $\xi \ge 0$ increasing as $\xi \to \infty$ no faster than ξ^p, $p < 1$.

The operator A generates on $[0;\tau)$ a K-convoluted semigroup in E if and only if any of the following conditions is satisfied:

- $\exists C>0, \omega\in\mathbb{R}:\ \|\mathscr{R}(\lambda)\|_{\mathscr{L}(E)}\le Ce^{\beta M(\gamma|\lambda|)},\ \lambda\in\{\mathbb{C}:\ Re\lambda>\alpha M(\gamma|\lambda|)+\omega\}$;
- problem (10) is well-posed in the space of abstract Roumier ultradistributions

$$\left(\mathscr{D}_a^{\{M_k\},B}\right)'(E):=\mathscr{L}(\mathscr{D}_a^{\{M_k\},B},E),$$

where $\{M_k\}$ is a sequence with an associated function $M(\xi)$.

4. R-semigroups.

Let $R\in L(E)$. Let A be densely defined in E, commute with R on its domain, and satisfy the condition $\overline{A|_{R(\mathrm{dom}\,A)}}=A$. The operator A generates on $[0;\tau)$ a local R-semigroup in E if and only if any of the following conditions is satisfied:

- problem (10) is R-correct on $[0;\tau)$, i.e. for any $f\in R\,(\mathrm{dom}\,A)$, $T<\tau$, there is a unique solution on the segment $[0;T]$ and

$$\exists C_T>0:\ \sup_{t\in[0;T]}\|u(t)\|\le C_T\|R^{-1}f\|;$$

- for any $t\in[0;\tau)$ there is an asymptotic R-resolvent $\mathscr{R}_t(\lambda)$ of A that satisfies for some $C_t>0$ the condition

$$\left\|\frac{d^k}{d\lambda^k}\mathscr{R}_t(\lambda)\right\|_{\mathscr{L}(X)}\le\frac{C_tk!}{|\lambda|^{k+1}},\quad \frac{k}{\lambda}\in[0;t],\quad \lambda>0,\ k\in\mathbb{N}_0.$$

In [10], these classifications are compared in the case when the problem (10) is considered with a differential operator $A=\mathsf{A}\left(i\frac{\partial}{\partial x}\right)$ in the space of vector functions $f\in E=L_2^m(\mathbb{R}^n)=L_2(\mathbb{R}^n)\times\ldots\times L_2(\mathbb{R}^n)$. The results of the comparison are clearly illustrated by the scheme, see Fig. 1.

Since the generators of Levy processes in the general case are ΨD-operators, from the point of view of this scheme, we should restrict ourselves to semigroups corresponding to Levy processes with zero Levy measure whose generators are differential operators. As shown in [1], the transition semigroups corresponding to Levy processes are strongly continuous in the space $L_2(\mathbb{R}^n)$. Therefore, the problem (10) with the generator of the Levy process in the space $L_2(\mathbb{R}^n)$ is the problem with the generator of the semigroup of class C_0. In the scheme, such semigroups are marked as "transition semigroups".

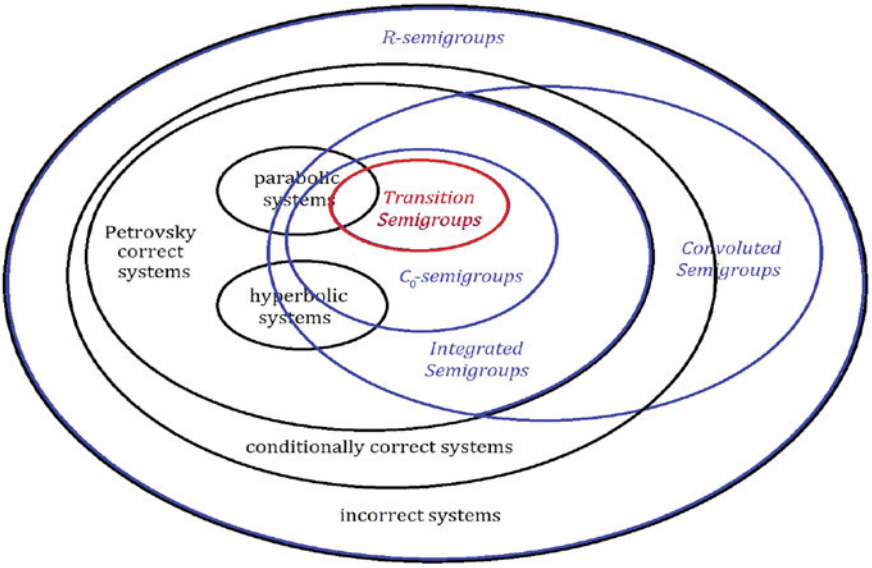


Fig. 1 Comparison scheme

4 Partial Extension of the Gelfand–Shilov Classification

As already noted, the semigroup classification takes place for problems with a wider class of operators than differential ones. In this section, we propose an extension of a part of the Gelfand–Shilov classification to problems relevant from the point of view of applications related to Levy processes. Namely, consider the Cauchy problem

$$\frac{\partial}{\partial t}u(t,x) = \left(\mathsf{A}\left(i\frac{\partial}{\partial x}\right) + \mathsf{K}\right)u(t,x), \quad t \geq 0,\ x \in \mathbb{R}^n, \quad u(0,x) = f(x). \tag{12}$$

Here $\mathsf{A}\left(i\frac{\partial}{\partial x}\right)$ is the operator defined in problem (9), $\mathsf{K} = \{\mathsf{K}_{j,k}\}_{j,k=1}^m$, where $\mathsf{K}_{j,k}$ are ΨD-operators defined (according to equality (4)) by symbols

$$s_{j,k}(\alpha) = \int_{\mathbb{R}^n\setminus\{0\}} \left(e^{i(\alpha,y)} - 1 - i(\alpha,y)\chi_{|y|\leq 1}(y)\right)\nu_{j,k}(dy),$$
$$j,k = 1,\ldots,m,\ \alpha \in \mathbb{R}^n.$$

With respect to $\nu_{j,k}$ we assume that for all $j,k=1,\dots,m$ the measures $\nu_{j,k}$ are Levy measures on $\mathscr{B}(\mathbb{R}^n)$ and, as mentioned above, $|s_{j,k}(\alpha)| \le C_{j,k}(1+|\alpha|)^2$. With this definition, the operators $\mathsf{K}_{j,k}$, $j,k=1,\dots,m$, have the form

$$\mathsf{K}_{j,k}f_k(x) = \int_{\mathbb{R}^n\setminus\{0\}} \big(f_k(x+y) - f_k(x) - (\nabla f_k(x), y)\chi_{|y|\le 1}(y)\big)\,\nu_{j,k}(dy),$$
$$x \in \mathbb{R}^n.$$

Next, from the problem (12), we pass to the Fourier-transformed problem. Due to the linearity of the Fourier transform, the solution of the transformed problem has the form

$$\widetilde{u}(t,\alpha) = e^{t(\mathsf{A}(\alpha)+\widetilde{\mathsf{K}}(\alpha))}\,\widetilde{f}(\alpha),$$

where $\widetilde{\mathsf{K}}(\alpha)$ is the operator of multiplication by a matrix whose elements are the Fourier transforms of the operators $\mathsf{K}_{j,k}$, i.e. $\widetilde{\mathsf{K}}(\alpha) = \{s_{j,k}(-\alpha)\}_{j,k=1}^m$. Note that for the functions $s_{j,k}(\alpha)$ and $s_{j,k}(-\alpha)$, as Fourier transforms with arguments that differ in signs, the same estimates take place:

$$|s_{j,k}(-\alpha)| \le C_{j,k}(1+|\alpha|)^2.$$

Hence it follows that the norm of the matrix $\widetilde{\mathsf{K}}(\alpha)$ considered as a linear operator in the space $\mathbb{R}^m$ satisfies the following estimate:

$$\|\widetilde{\mathsf{K}}(\alpha)\|^2_{\mathbb{R}^m} \le \sum_{j=1}^m\sum_{k=1}^m |s_{j,k}(-\alpha)|^2 \le C_K^2(1+|\alpha|)^4, \quad C_K>0,\ \alpha\in\mathbb{R}^n. \tag{13}$$

The estimate (13) implies an estimate for the norm of the operator of multiplication by the matrix exponent $e^{t\widetilde{\mathsf{K}}(\alpha)}$ for any $t \ge 0$:

$$\|e^{t\widetilde{\mathsf{K}}(\alpha)}\|_{\mathbb{R}^m} = \left\|\sum_{j=0}^\infty \frac{t^j}{j!}\widetilde{\mathsf{K}}^j(\alpha)\right\|_{\mathbb{R}^m} \le \sum_{j=0}^\infty \frac{t^j}{j!}\|\widetilde{\mathsf{K}}(\alpha)\|^j_{\mathbb{R}^m} \le e^{C_K(1+|\alpha|)^2 t}.$$

This estimate implies estimates for solution operators of Fourier transformed problem (12) and the possibility of extending part of the classification for the specified class of ΨD-operators. For problem (12), the following cases are possible.

- If the operator A defines a parabolic problem with parameter $h > 2$, then the estimate

$$\left\|e^{t(\mathsf{A}(\alpha)+\widetilde{\mathsf{K}}(\alpha))}\right\|_{\mathbb{R}^m} \le Ce^{t(-C_1|\alpha|^h+C_2+C_K(1+|\alpha|)^2)} \le Ce^{t(-C_4|\alpha|^{h-2}+C_5)}, \quad t\ge 0,$$

and such a problem (12) should be called parabolic.

- If the operator A defines a parabolic problem with the parameter $h = 2$, then, depending on the relationship between the constants C_1 and C_K, the problem (12) should be called parabolic (for $C_1 > C_K$) or incorrect for ($C_1 \le C_K$).
- If the operator A defines a parabolic problem with parameter $h < 2$, then

$$\left\| e^{t(\mathsf{A}(\alpha)+\widetilde{\mathsf{K}}(\alpha))} \right\|_{\mathbb{R}^m} \le C e^{t(-C_1|\alpha|^h + C_2 + C_K(1+|\alpha|)^2)} \le C e^{t(C_4|\alpha|^{2-h}+C_5)}, \quad t \ge 0,$$

  and such a problem (12) should be called conditionally well-posed (for $h > 1$) or ill-posed (for $0 \le h \le 1$).
- If A defines an incorrect system with parameter l_0, then we obtain the estimate

$$\left\| e^{t(\mathsf{A}(\alpha)+\widetilde{\mathsf{K}}(\alpha))} \right\|_{\mathbb{R}^m} \le C e^{t(C_1|\alpha|^{l_0} + C_2 + C_K(1+|\alpha|)^2)} \le C e^{t(C_4|\alpha|^r + C_5)}, \quad t \ge 0,$$

  where $r = \max\{l_0, 2\}$. Such a problem (12) should be called ill-posed.

Now, among the considered operators $\mathsf{A}\left(i\frac{\partial}{\partial x}\right) + \mathsf{K}$, we single out the operators related to the generators of Levy processes. Namely, consider the Cauchy problem with the operator $\mathsf{A}\left(i\frac{\partial}{\partial x}\right) + \mathsf{K}$, the generator of the semigroup corresponding to the Levy process , which, by virtue of the Levy-Khinchin formula, is determined by the equality

$$\left(\mathsf{A}\left(i\frac{\partial}{\partial x}\right) + \mathsf{K}\right) f(x) = (b, \nabla f(x)) + \frac{1}{2} div(Q\nabla f(x)) + \int_{\mathbb{R}^n \setminus \{0\}} \left(f(x+y) - f(x) - (\nabla f(x), y)\chi_{|y|\le 1}(y)\right) \nu(dy). \tag{14}$$

From the problem (12) with the operator (14), we pass to the Fourier-transformed problem, whose properties are determined by the behavior of the function

$$F(\alpha) := \mathsf{A}(\alpha) + \widetilde{\mathsf{K}}(\alpha) = -i(b, \alpha) - \frac{1}{2}(\alpha\,, Q\alpha) + \int_{\mathbb{R}^n \setminus \{\mathbf{0}\}} \left(e^{-i(\alpha, y)} - 1 + i(\alpha, y)\chi_{|y|\le 1}(y)\right) \nu(dy).$$

According to the formula (6), the function $F(\alpha)$ coincides with the function $\eta(-\alpha)$. Hence, by virtue of the estimate (7), we obtain

$$\left| e^{t(\mathsf{A}(\alpha)+\widetilde{\mathsf{K}}(\alpha))} \right| \le 1, \quad t \ge 0,\ \alpha \in \mathbb{R}^n.$$

Therefore, the Cauchy problem (12) for an equation with a generator (14) of a semigroup corresponding to a Levy process is well-posed in the sense of Petrovsky.

Note 1 An important fact follows from Theorems XIII.52 and XIII.53 [11]: $e^{tF(\alpha)}$ for every $t \ge 0$ is a positive-definite, in the general case, generalized function. Then

it follows from the Bochner–Schwarz [12] theorem that when constructing solutions in spaces of generalized functions for Cauchy problems with the operator (14) it suffices to consider these problems in the space of slowly growing distributions $\mathscr{S}'$. This means that, in contrast to the original Gelfand–Shilov classification for problems with differential operators, the study of the distinguished class of problems when passing to Fourier-transformed problems does not require access to spaces of generalized functions with complex arguments.

Note 2 The construction of a partial extension of the Gelfand–Shilov classification for the class of problems with operators containing generators of Levy processes leads to the idea of comparing this extension with the semigroup classification, since the latter takes place for operators that include, along with differential, pseudo-differential operators. However, the proofs of connections in the scheme [10] rely heavily on the behavior of $\Lambda(\gamma)$, the key characteristic of the matrix of differential operators. Therefore, comparing the obtained extension with the semigroup classification is the problem of a separate study, which requires qualitatively new ideas.

Acknowledgments The paper is supported by Russian Science Foundation No 23-21-00199.

References

1. Applebaum, D.: Levy processes and Stochastic Calculus. Cambridge University Press, Cambridge (2009). https://doi.org/10.1017/CBO9780511809781
2. Böttcher, B., Schilling, R., Wang, J.: Lévy Matters III. Lévy-Type Processes: Construction, Approximation and Sample Path Properties. Springer, Berlin (2013). https://doi.org/10.1007/978-3-319-02684-8
3. Sato, K.-I.: Levy Processes and Infinitely Divisible Distributions. Cambridge University Press, Cambridge (2013)
4. Hörmander, L.: The Analysis of Linear Partial Differential Operators I. Springer, Berlin (2003)
5. Jacob, N.: Pseudo-Differential Operators and Markov Processes, vol. 1. Imperial College Press, London (2001)
6. Gelfand, I.M., Shilov, G.E.: Generalized Functions, Volume 3: Theory of Differential Equations. AMS Chelsea Publishing, New York (2016)
7. Melnikova, I.V.: Stochastic Cauchy Problems in Infinite Dimensions. Regularized and Generalized Solutions. CRC Press, New York (2016). https://doi.org/10.1201/9781315372631
8. Anufrieva, U.A., Melnikova, I.V.: Peculiarities and regularization of ill-posed Cauchy problems with differential operators. J. Math. Sci. **148**(4), 481–632 (2008)
9. Melnikova, I.V., Filinkov, A.I.: The Cauchy Problem: Three Approaches. Chapman& Hall/CRC, New York (2001)
10. Melnikova, I.V., Alekseeva, U.A.: Semigroup classification and gelfand-shilov classification of systems of partial differential equations. Math. Not. **104**(6), 886–899 (2018)
11. Reed, M., Simon, B.: Methods of Modern Mathematical Physics. Volume 4: Analysis of Operators. Academic Press, Cambridge (1978)
12. Reed, M., Simon, B.: Methods of Modern Mathematical Physics. Volume 2: Fourier Analysis, Self-Adjointness. Academic Press, Cambridge (1975)
13. Melnikova, I.V., Bovkun, V.A.: Semigroups of operators related to stochastic processes in an extension of the Gelfand-Shilov classification. Trudy Instituta Matematiki i Mekhaniki URO RAN **27**(4), 74–87 (2021)

Part VIII
Harmonic Analysis and Partial Differential Equations

The Index of Toeplitz Operators on Compact Lie Groups and on Simply Connected Closed 3-Manifolds

Duván Cardona

Abstract In this note we use the notion of an operator-valued symbol in the sense of Ruzhansky and Turunen in order to compute the index of Toeplitz operators on compact Lie groups. Our approach combines the Connes index theorem and the infinite-dimensional operator-valued symbolic calculus of Ruzhansky-Turunen. We also give applications to the index of Toeplitz operators on simply connected closed 3-manifolds $\mathbb{M} \simeq \mathbb{S}^3 \simeq \mathrm{SU}(2)$ by using the Poincaré theorem (Perelman, Spaces with curvature bounded below. In: Proceedings of ICM 1994, pp. 517–525. Birkhäuser, Basel (1995). MR1403952 (97g:53055); Perelman, The entropy formula for the Ricci flow and its geometric applications (2002). arXiv.math.DG/0211159; Perelman, Ricci flow with surgery on three-manifolds (2003). arXiv.math.DG/0303109; Perelman, Finite extinction time for the solutions to the Ricci flow on certain three-manifolds (2003). arXiv.math.DG/0307245).

1 Introduction

Using the interplay between the Connes index theorem for Fredholm modules and the operator-valued symbolic calculus of Ruzhansky and Turunen, in this paper we compute the index of Toeplitz operators acting on functions in compact Lie groups. Although, the point of departure of the index theory is the Atiyah-Singer index theorem, proved in 1963 in [2] (see also, the historical references [1, 3–9]) the analysis for the index of Toeplitz operators started with the classical formula of Noether-Gohberg-Krein (see [30])

$$\mathrm{ind}(PM_fP) = -\mathrm{wn}(f) := -\frac{1}{2\pi i}\int_{\mathbb{S}^1} f^{-1}df. \tag{1}$$

D. Cardona (✉)
Ghent University, Ghent, Belgium

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_37

In the index formula (1), the function $f \in C^{\infty}(G)$ is invertible everywhere, M_f is the multiplication operator by f, and P is the projection from $L^2(\mathbb{S}^1)$ into the Hardy space $H^1(\mathbb{S}^1)$, consisting of those functions in L^2 with negative Fourier coefficients vanishing. The main feature in (1) is that the left hand side is of analytical nature, but the right hand side has topological information given by minus the winding number of f around of zero. This result was extended by V. Venugopalkrishna [45] to the unit ball in $\mathbb{C}^n$ and by L. Boutet de Monvel to arbitrary strictly pseudoconvex domains [10]. A similar formula for boundaries of strictly pseudo-convex domains in $\mathbb{C}^n$ was announced by Dynin [21]. For the spectral properties of Toeplitz operators and its index theory in several complex variables, we refer the reader to the references Douglas [19], Guillemin [22], Boutet de Monvel and Guillemin [13], Boutet de Monvel [11, 12], Murphy [27–29] as well as the monograph Upmeier[44]. The index theory in the operator-valued context can be found in Cardona [14]. We refer the reader to Hong [24] for a Lie-algebraic approach to the local index theorem on compact Lie groups and general compact homogeneous spaces.

In this paper we want to compute the index of Toeplitz operators on compact Lie groups by using the recent notion of operator valued symbol [40, 41] in the sense of Ruzhansky and Turunen. Our main theorem can be announced as follows. Here G is a compact Lie group, $\widehat{G}$ is its unitary dual, e_G is the identity element of G, and for every irreducible representation $[\xi] \in \widehat{G}$, $d_\xi := \dim[\xi : G \to \hom(\mathbb{C}^{d_\xi})]$ denotes the dimension of the representation space.

Theorem 1 (Index of Toeplitz Operators) *Let G be a compact Lie group. Let us consider a smooth and invertible function f on G. Let $\Pi : L^2(G) \to \Pi'$ be the orthogonal projection of a closed subspace $\Pi' \subset L^2(G)$. If $T_f = \Pi M_f \Pi$ is the Toeplitz operator with symbol f, then T_f extends to a Fredhlom operator on $L^2(G)$ and its analytical index is given by*

$$\begin{aligned} \operatorname{ind}(T_f) &= \int_G \sum_{[\xi]\in\hat{G}} d_\xi \operatorname{Tr}[\sigma_{I_{\Pi,f}}(x)\xi(e_G)]dx, \\ I_{\Pi,f} &:= f^{-1}\underbrace{[\Pi, f][\Pi, f^{-1}]\cdots[\Pi, f]}_{n+2-times} \end{aligned} \tag{2}$$

if n is odd, or

$$\begin{aligned} \operatorname{ind}(T_f) &= \int_G \sum_{[\xi]\in\hat{G}} d_\xi \operatorname{Tr}[\sigma_{I_{\Pi,f}}(x)\xi(e_G)]dx, \\ I_{\Pi,f} &:= f^{-1}\underbrace{[\Pi, f][\Pi, f^{-1}]\cdots[\Pi, f]}_{n+1-times}, \end{aligned} \tag{3}$$

if n is even, where

- *$\sigma_{I_{\Pi,f}}$ is the Ruzhansky-Turunen operator valued symbol associated to $I_{\Pi,f}$, and*
- *$[A, B] := AB - BA$ is the commutator operator defined by the operators A and B.*

It is important to mention that from Theorem 1, we can derive index formulae for Toeplitz operators on 3-dimensional closed manifolds. In fact, if $\mathbb{M}$ is a simply connected closed 3-manifold, the Poincaré conjecture/theorem proved by Perelman provides a diffeomorphism $\mathbb{M} \simeq \mathbb{S}^3 \simeq \mathrm{SU}(2)$, (see Ruzhansky and Turunen [40], pag. 578, and Perelman [31–34]) that allows us to construct a global (operator-valued) pseudo-differential calculus on $\mathbb{M}$. This implies that the analysis employed for Toeplitz operators on SU(2) gives a similar construction for Toeplitz operators on $\mathbb{M}$. So, if $T_\omega = \Pi M_\omega \Pi$ is a Toeplitz operator with symbol ω, such that $\omega^{-1} \in C^\infty(\mathbb{M})$, we will prove the following algebraic index formula (see Corollary 1),

$$\mathrm{ind}(T_\omega) = \int\limits_{\mathbb{M}} \sum_{\ell \in \frac{1}{2}\mathbb{N}_0} (2\ell+1)\mathrm{Tr}[\sigma_{\omega^{-1}[\Pi,\omega][\Pi,\omega^{-1}][\Pi,\omega][\Pi,\omega^{-1}][\Pi,\omega]}(x)\xi_\ell(e_{\mathbb{M}}))]dx, \tag{4}$$

where $e_{\mathbb{M}} = \Phi(e_{\mathrm{SU}(2)})$.

This paper is organized as follows. In Sect. 2 we present some basics on the matrix valued and operator valued quantizations procedure for (global) pseudo-differential operators on compact Lie groups. In Sect. 3 we study the index of Toeplitz operators on compact Lie groups. Finally, in Sect. 4 we investigate the index of Toeplitz operators on simply connected closed 3-manifolds.

2 Global Operators on Compact Lie Groups: Preliminaries

In this section we consider the pseudo-differential calculus of Ruzhansky and Turunen on compact Lie groups. Here, a Lie group is a group G that at the same time is a finitedimensional manifold of differentiability class C^2, in such a way that the two group operations of G, $x \mapsto x^{-1}$, and $(x, y) \mapsto x \cdot y$ are C^2-mappings, (see Duistermaat and Kolk [20]). Theorem 1.6.1 of [20] shows that every C^2 Lie group G, in the sense of the previous remark, can be provided with the structure of a real-analytic manifold for which it becomes a real-analytic Lie group.

2.1 Operator-Valued Quantization of Global Operators on Compact Lie Groups

In this subsection we present the Ruzhansky-Turunen operator valued quantization procedure for global operators on compact Lie groups. Throughout of this paper G is a compact Lie group endowed with its normalised Haar measure dg. Our main tool is the Fourier analysis carried by a global Fourier transform. It can be defined as follows.

Definition 1 (Operator-Valued Fourier Transform) Let G be a compact Lie group. For $f \in \mathscr{D}'(G)$, the respective right-convolution operator $r(f) : C^\infty(G) \to C^\infty(G)$ is defined by

$$r(f)g = g * f, \quad g \in C^\infty(G). \tag{5}$$

If $f \in L^2(G)$, the (right) global Fourier transform is defined by

$$\widehat{f} \equiv r(f) = \int_G f(y)\pi_R(y)^* dy, \tag{6}$$

where π_R is the right regular representation on G, defined by $\pi_R(x)f(y) = f(yx)$ and $\pi_R(x)^* = \pi_R(x^{-1})$, $x \in G$.

In terms of the Fourier transform, the Fourier inversion formula can be announced as follows.

Theorem 2 (Fourier Inversion Formula) *Let G be a compact Lie group. Let us assume that $f \in C^\infty(G)$. Then $r(f)\pi_R(x)$ is a trace class operator on $L^2(G)$, and the indentity*

$$f(x) = \mathrm{Tr}(r(f)\pi_R(x)), \quad f \in C^\infty(G), \tag{7}$$

holds true for every $x \in G$.

Definition 2 (Ruzhansky-Turunen Operator Valued Quantization) Let G be a compact Lie group. If $\rho : G \to \mathscr{B}(C^\infty(G))$ is a continuous operator, the pseudo-differential operator A associated to ρ, is defined by

$$Af(x) = \mathrm{Tr}(\rho(x)r(f)\pi_R(x)), \quad f \in C^\infty(G). \tag{8}$$

Conversely we have the following theorem due to Ruzhansky and Turunen.

Theorem 3 *If $A : C^\infty(G) \to C^\infty(G)$ is a continuous linear operator, then there exists an unique $\sigma_A : G \to \mathscr{B}(C^\infty(G))$ (called the operator-valued symbol of A) satisfying*

$$Af(x) = \mathrm{Tr}(\sigma_A(x)r(f)\pi_R(x)), \quad f \in C^\infty(G). \tag{9}$$

Proof The symbol σ_A is defined as follows. Let $K_A \in C^\infty(G)\widehat{\otimes}\mathscr{D}'(G)$ be the distributional Schwartz kernel of A and $R_A(x,y) = K(x,y^{-1}x)$ is the right-convolution kernel associated to A. If $x \in G$, and $R_A(x) \in \mathscr{D}'(G)$ is defined by $(R_A(x))(y) = R_A(x,y)$ for every $y \in G$, the (right) operator-valued symbol $\rho = \sigma_A$ associated to A is defined by $\sigma_A(x) := r(R_A(x)),\ \ x \in G$. It can be proved that this operator valued operator satisfies (9) (see Ruzhansky and Turunen [40], pag. 583) □

2.2 Matrix-Valued Quantization of Global Operators on Compact Lie Groups

In this subsection we will present the matrix-valued quantization of global operators on compact Lie groups. There are two notions of continuous operators on smooth functions on compact Lie groups we can use. Namely, the one used in the case of general manifolds (based on the idea of *local symbols* as in Hörmander [23]) and, in a much more recent context, the one of global operators on compact Lie groups as defined by M. Ruzhansky and V. Turunen [38, 39] (from *full symbols*, for which the notations and terminologies are taken from [41]).

Let us consider for every compact Lie group G its unitary dual $\widehat{G}$, that is the set of continuous, irreducible, and unitary representations on G. As is the operator-valued quantization, the main tool in the matrix-valued quantization is a suitable notion of Fourier transform. We define it as follows.

Definition 3 (Matrix-Valued Fourier transform) Let G be a compact Lie group. Let us assume that $\varphi \in C^\infty(G)$. Then, the matrix-valued Fourier transform of φ at $[\xi]$, is defined by

$$\mathscr{F}_G\varphi(\xi) := \int_G \varphi(x)\xi(x)^* dx.$$

The Peter-Weyl theorem on compact Lie groups implies the following inversion formula.

Theorem 4 (Fourier Inversion Formula) *Let G be a compact Lie group. Let us assume that $f \in L^1(G)$. Then, we have*

$$\varphi(x) = \sum_{[\xi]\in\widehat{G}} d_\xi Tr(\xi(x)\mathscr{F}_G\varphi(\xi)),$$

for all $x \in G$. In this case, the Plancherel identity on $L^2(G)$ is given by,

$$\|\varphi\|^2_{L^2(G)} = \sum_{[\xi]\in\widehat{G}} d_\xi Tr(\widehat{\varphi}(\xi)\widehat{\varphi}(\xi)^*) = \|\widehat{\varphi}\|^2_{L^2(\widehat{G})}.$$

Notice that, since $\|A\|_{\mathrm{HS}} = \sqrt{\mathrm{Tr}(AA^*)}$, the term within the sum is the Hilbert-Schmidt norm of the matrix $\widehat{\varphi}(\xi)$. The matrix-valued quantization procedure of Ruzhansky-Turunen can be introduced as follows. Any linear operator A on G mapping $C^\infty(G)$ into $\mathcal{D}'(G)$ gives rise to a *matrix-valued symbol* $\sigma_A(x,\xi) \in \mathbb{C}^{d_\xi \times d_\xi}$ given by

$$\sigma_A(x,\xi) \equiv \xi(x)^*(A\xi)(x) := \xi(x)^*[A\xi_{ij}(x)]_{i,j=1,\cdots,d_\xi}, \tag{10}$$

which can be understood from the distributional viewpoint. Then it can be shown that the operator A can be expressed in terms of such a symbol as [41]

$$Af(x) = \sum_{[\xi]\in\widehat{G}} d_\xi \mathrm{Tr}[\xi(x)\sigma_A(x,\xi)\widehat{f}(\xi)]. \tag{11}$$

We will denote by $\sigma_A(\cdot)$ and $\sigma_A(\cdot,\cdot)$ to the operator-valued symbol and the matrix-valued symbol associated to A respectively.

Lemma 1 (Matrix Symbols vs Operator-Valued Symbols) *Theorem 10.11.16 in Ruzhansky and Turunen [40] gives the identity*

$$\sigma_A(x,\xi) = \sigma_{\sigma_A(x)}(y,\xi) := \xi(y)^*(\sigma_A(x)\xi)(y),$$

for all $y \in G$. In particular, if $y = e_G$ is the identity element in G,

$$\sigma_A(x)\xi(e_G) = \sigma_A(x,\xi),\ x \in G,\ [\xi] \in \widehat{G}. \tag{12}$$

Now, we want to introduce Sobolev spaces and, for this, we give some basic tools. Let $\xi \in Rep(G) := \cup\widehat{G}$, if $x \in G$ is fixed, $\xi(x) : \mathbb{C}^{d_\xi} \to \mathbb{C}^{d_\xi}$ is an unitary operator and $d_\xi := \dim \mathbb{C}^{d_\xi} < \infty$. There exists a non-negative real number $\lambda_{[\xi]}$ depending only on the equivalence class $[\xi] \in \hat{G}$, but not on the representation ξ, such that $-\mathcal{L}_G\xi(x) = \lambda_{[\xi]}\xi(x)$; here $\mathcal{L}_G$ is the Laplacian on the group G (in this case, defined as the Casimir element on G). Let $\langle\xi\rangle$ denote the function $\langle\xi\rangle = (1+\lambda_{[\xi]})^{\frac{1}{2}}$.

Definition 4 Let G be a compact Lie group. For every $s \in \mathbb{R}$, the *Sobolev space* $H^s(G)$ on the Lie group G is defined by the condition: $f \in H^s(G)$ if only if $\langle\xi\rangle^s \widehat{f} \in L^2(\widehat{G})$.

The Sobolev space $H^s(G)$ is a Hilbert space endowed with the inner product

$$\langle f,g\rangle_{H^s(G)} = \langle J^s f, J^s g\rangle_{L^2(G)}$$

where, for every $r \in \mathbb{R}$, $J^s : H^r(G) \to H^{r-s}(G)$ is the bounded pseudo-differential operator (Bessel potential) with symbol $\sigma_{J^s}(x,\xi) := \langle\xi\rangle^s I_\xi$.

Remark In this paper the notion of Sobolev spaces $H^s(G)$ is essential, we will use this spaces in the proof of Lemma 2 and for description of global operators. An

important fact is that every global operator T of order m is a bounded operator from $H^s(G)$ into $H^{s-m}(G)$ (see Ruzhansky and Turunen [38]). □

Now we introduce, for every $m \in \mathbb{R}$, the Hörmander class $\Psi^m(G)$ of pseudo-differential operators of order m on the compact Lie group G. As a compact manifold we consider $\Psi^m(G)$ as the set of those operators which, in all local coordinate charts, give rise to pseudo-differential operators in the Hörmander class $\Psi^m(U)$ for an open set $U \subset \mathbb{R}^n$, characterized by symbols satisfying the usual estimates [23]

$$|\partial_x^\alpha \partial_\xi^\beta \sigma(x,\xi)| \leq C_{\alpha,\beta} \langle \xi \rangle^{m-|\beta|}, \tag{13}$$

for all $(x,\xi) \in T^*U \cong \mathbb{R}^{2n}$ and $\alpha, \beta \in \mathbb{N}^n$. This class contains, in particular, differential operator of degree $m > 0$ and other well-known operators in global analysis such as heat kernel operators. The class The Hörmander classes $\Psi^m(G)$ where characterized in [40, 42] by the condition: $A \in \Psi^m(G)$ if only if its matrix-valued symbol $\sigma_A(x,\xi)$ satisfies the inequalities

$$\|\partial_x^\alpha \mathbb{D}^\beta \sigma_A(x,\xi)\|_{op} \leq C_{\alpha,\beta} \langle \xi \rangle^{m-|\beta|}, \tag{14}$$

for every $\alpha, \beta \in \mathbb{N}^n$. For a rather comprehensive treatment of this global calculus we refer to [40].

2.3 *Fredholm Operators on Hilbert Spaces, Fredholm Modules on Associative Algebras and the Connes Index Theorem*

The index is defined for a broad class of operators called Fredholm operators. Now, we introduce this notion in more detail. For X, Y normed spaces $B(X,Y)$ is the set of bounded linear operators from X into Y. In particular, if $X = Y = H$, $B(H) \equiv B(H,H)$ denotes the algebra of bounded operators on H. Here, we consider $H = L^2(G)$.

Definition 5 If H_1 and H_2 are Hilbert spaces, the closed and densely defined operator $A : H_1 \to H_2$ is Fredholm if only if $\mathrm{Ker}(A)$ is finite dimensional and $A(H_1) = \mathrm{Rank}(A)$ is a closed subspace of H_2 with finite codimension. In this case, the index of A is defined by $\mathrm{Ind}(A) = \dim \mathrm{Ker}(A) - \dim \mathrm{Coker}(A)$. The index formula also can be written as

$$\mathrm{ind}(A) = \mathrm{dimKer}(A) - \mathrm{dimKer}(A^*).$$

In our analysis we use the Connes index theorem for odd Fredholm modules on associative algebras. We recall this definition as follows.

Definition 6 Let A be an associative algebra over $\mathbb{C}$. An odd Fredholm module over A is a triple (A, π, F) consisting of:

- a Hilbert space H,
- a representation π of A as bounded operators on H,
- a self-adjoint operator F such that $F^2 = I$ and $[F, \pi(a)]$ is a compact operator on H for all $a \in A$.

Additionally, if there exist p such that $[F, \pi(a)] \in L^p(H) = \{T \in B(H, H) : \sum_\nu [s_\nu(T)]^p < \infty\}$, (here, $\{s_\nu(T)\}_\nu$ denotes the sequence of singlular values of T) we say that the module (A, π, F) is p-summable. The corresponding Connes index theorem is the following (see A. Connes, [15]).

Theorem 5 (Connes Index Theorem) *Let (A, π, F) be a p-summable odd Fredholm module and P given by*

$$P = \frac{1}{2}(F + I). \tag{15}$$

Then, for every invertible element $u \in A$, the operator $PuP : PH \to PH$ is Fredholm and its analytical index is given by

$$\mathrm{ind}(PuP) = \frac{(-1)^{2k+1}}{2^{2k+1}} \mathrm{Tr}[a_0[F, a_1] \cdots [F, a_{2k+1}]], \tag{16}$$

where p is the smallest odd integer larger than n, $a_i = u^{-1}$ for i even, and $a_i = u$ for i odd.

In the next section we compute the index of a Toeplitz operator $T_f = PM_fP$. In order to use the Connes theorem, we use (the well know fact) that $(C^\infty(G), \pi, 2P - I)$ is a odd Fredholm module, where π is the representation $\pi(g) = M_g$ defined from $C^\infty(G)$ into the algebra of bounded operators on $L^2(G)$.

3 The Index of Toeplitz Operators on Compact Lie Groups

In this section, we compute the index of Toeplitz operators by using trace formulae for global operators of trace class and the index theorem of Connes mentioned above. The corresponding statement for trace class global operators is the following (for the proof, we refer the reader to Cardona [14]. The proof is based in the arguments developed by Delgado and Ruzhansky [16–18]).

Theorem 6 *Let G be a compact Lie group. Let A be a pseudo-differential operator on $\Psi^m(G)$, $m < -\dim(G)$. Then A is trace class on $L^2(G)$ and*

$$\mathrm{Tr}(A) = \int_G \sum_{[\xi]\in\hat{G}} d_\xi \mathrm{Tr}[\sigma_A(x,\xi)]dx,$$

where $\sigma_A(x,\xi)$ is the matrix-valued symbol of A.

Remark 1 The inequality $m < -\dim(G)$ in Theorem 6 is sharp. Indeed, if $B_{-m} := (1+\mathcal{L}_G)^{\frac{m}{2}}$ with $m \geq -n$, then is easy to see that B_{-m} is of order m, and its system of eigenvalues is not summable. So, the order restriction $m \geq -n$, implies that B_{-m} is not of trace class.

In order to apply the Connes theorem, we need the following well known lemma. For completeness we provide a proof.

Lemma 2 *Let G be a compact Lie group. The triple $(C^\infty(G), \pi, 2P - I)$ where $\pi : C^\infty(G) \to B(L^2(G))$ is the representation defined at g by $\pi(g) = M_g$ (multiplication operator by g,) is a p-summable odd Fredholm module for all $p > n := \dim(G)$.*

Proof We only need to prove that $F = 2P - I$ is self-adjoint, that $F^2 = I$ and the compactness of every commutator $[F, \pi(g)]$. Because $P^2 = P$ and P is orthogonal, we deduce that F is self-adjoint and $F^2 = 4P^2 - 4P + I = I$. Now, if $g \in C^\infty(G)$ the operators $M_g P$ and $P M_g$ have in local coordinates same principal symbol and the order of the commutator $T = [P, M_g]$ is -1. This implies that $|T| \in L^p(L^2(G))$ for $p > n = \dim(G)$. □

Now, with the machinery presented above, we can give a short argument for proving our main theorem.

Theorem 7 (Index of Toeplitz Operators) *Let G be a compact Lie group. Let us consider a smooth and invertible function f on G. Let $\Pi : L^2(G) \to \Pi'$ be the orthogonal projection determined by a closed subspace $\Pi' \subset L^2(G)$. If $T_f = \Pi M_f \Pi$ is the Toeplitz operator with symbol f, then T_f extends to a Fredholm operator on $L^2(G)$ and its index is given by*

$$\mathrm{ind}(T_f) = \int_G \sum_{[\xi]\in\hat{G}} d_\xi \mathrm{Tr}[\sigma_{I_{\Pi,f}}(x)\xi(e_G)]dx,$$

$$I_{\Pi,f} := f^{-1}\underbrace{[\Pi, f][\Pi, f^{-1}]\cdots[\Pi, f]}_{n+2-times} \tag{17}$$

if n is odd, or

$$\mathrm{ind}(T_f) = \int_G \sum_{[\xi]\in\hat{G}} d_\xi \mathrm{Tr}[\sigma_{I_{\Pi,f}}(x)\xi(e_G)]dx,$$

$$I_{\Pi,f} := f^{-1}\underbrace{[\Pi, f][\Pi, f^{-1}]\cdots[\Pi, f]}_{n+1-times}, \tag{18}$$

if n is even, where $\sigma_{I_{\Pi,f}}$ is the Ruzhansky-Turunen operator valued symbol associated to $I_{\Pi,f}$.

Proof Let us observe that the index of T can be computed as,

$$\mathrm{ind}(T_f) = \mathrm{ind}(\Pi f \Pi) = \frac{(-1)^{2k+1}}{2^{2k+1}} \mathrm{Tr}[a_0[F, a_1] \cdots [F, a_{2k+1}]], \tag{19}$$

where $p = 2k + 1 > n := \dim(G)$, is the smallest odd integer larger that n, $a_i = f^{-1}$ for i even, and $a_i = f$ for i odd. Here we have used that $(C^\infty(G), M_f, 2\Pi - I)$ is a p-summable odd Fredholm module as well as the Connes index Theorem. Since $F = 2P - I$, we have

$$[F, a_1] \cdots [F, a_{2k+1}] = 2^{2k+1}[\Pi, a_1] \cdots [\Pi, a_{2k+1}], \tag{20}$$

and

$$\mathrm{ind}(T_f) = \mathrm{ind}(\Pi f \Pi) = -\mathrm{Tr}[a_0[\Pi, a_1] \cdots [\Pi, a_{2k+1}]]. \tag{21}$$

Now we need to compute the trace of the operator $I_{\Pi,f} = a_0[\Pi, a_1] \cdots [\Pi, a_{2k+1}] \in S^{-p}(G)$, $p = 2k + 1 > n$, using its operator-valued symbol. This can be done with Theorem 6. Using (12), we have the following matrix identity

$$\sigma_{I_{\Pi,f}}(x)\xi(e_G) = \sigma_{I_{\Pi,f}}(x, \xi), \ x \in G, \ [\xi] \in \widehat{G}. \tag{22}$$

From Theorem 6, we have

$$\mathrm{ind}(T_f) = \mathrm{ind}(\Pi f \Pi) = -\mathrm{Tr}[I_{\Pi,f}] = -\int_G \sum_{[\xi] \in \hat{G}} d_\xi \mathrm{Tr}[\sigma_{I_{\Pi,f}}(x, \xi)]dx.$$

So, we have proved that

$$\mathrm{ind}(T_f) = \int_G \sum_{[\xi] \in \hat{G}} d_\xi \mathrm{Tr}[\sigma_{I_{\Pi,f}}(x)\xi(e_G)]dx,$$
$$I_{\Pi,f} := f^{-1} \underbrace{[\Pi, f][\Pi, f^{-1}] \cdots [\Pi, f]}_{n+2-times} \tag{23}$$

if n is odd, or

$$\mathrm{ind}(T_f) = \int_G \sum_{[\xi] \in \hat{G}} d_\xi \mathrm{Tr}[\sigma_{I_{\Pi,f}}(x)\xi(e_G)]dx,$$
$$I_{\Pi,f} := f^{-1} \underbrace{[\Pi, f][\Pi, f^{-1}] \cdots [\Pi, f]}_{n+1-times}, \tag{24}$$

if n is even. Thus, we finish the proof.

□

4 The Index of Toeplitz Operators on Closed 3-Manifolds

4.1 Ruzhansky-Turunen Construction

Let us consider the compact Lie group $\mathrm{SU}(2) \cong \mathbb{S}^3$ consisting of those orthogonal matrices A in $\mathbb{C}^{2\times 2}$, with $\det(A) = 1$. We recall that the unitary dual of SU(2) (see [40]) can be identified as

$$\widehat{\mathrm{SU}(2)} \equiv \{[\xi_l] : 2l \in \mathbb{N}, d_l := \dim \xi_l = (2l+1)\}. \tag{25}$$

Let us assume that $\mathbb{M}$ is a simply connected closed 3-manifold, (this means that every simple closed curve within the manifold can be deformed continuously to a point). By the Poincaré conjecture (c.f. [35]) proved by Perelman (see Perelman [31–34] and Morgan[26]), $\mathbb{M}$ is diffeomorphic to $\mathbb{S}^3 \simeq \mathrm{SU}(2)$. Every topological 3-manifold admits a differentiable structure and every homeomorphism between smooth 3-manifolds can be approximated by a diffeomorphism. Thus, classification results about topological 3-manifolds up to homeomorphism and about smooth 3-manifolds up to diffeomorphism are equivalent (see Morgan [26]).

Let $\Phi : \mathrm{SU}(2) \to \mathbb{M}$ be a diffeomorphism. By following Ruzhansky and Turunen, [40], pag. 578, $\mathbb{M}$ can be endowed with the natural Lie group structure induced by Φ. In fact, if x, y are coordinate points in $\mathbb{M}$ we can define

$$x \cdot y := \Phi(\Phi^{-1}(x) \times \Phi^{-1}(x)), \tag{26}$$

where $\times$ denotes the product of matrices on SU(2). We have a Frechet isorphism $C^\infty(\mathrm{SU}(2)) \simeq C^\infty(\mathbb{M})$ defined by

$$\Phi_* : C^\infty(\mathrm{SU}(2)) \to C^\infty(\mathbb{M}),\ \ \Phi_*(f) := f \circ \Phi^{-1};\ \ \Phi^* : C^\infty(\mathbb{M}) \to C^\infty(\mathrm{SU}(2)), \tag{27}$$

where $Phi_*(g) := g \circ \Phi$. Since $L^2(M) = \Phi_*(L^2(\mathrm{SU}(2)))$, and the Lie group structure on $\mathbb{M}$ provides a Peter-Weyl theorem on $\mathbb{M}$, we have $\widehat{\mathrm{SU}(2)} = \frac{1}{2}\mathbb{N}_0 \simeq \widehat{\mathbb{M}}$ in the sense that

$$\Phi_* : \widehat{\mathrm{SU}(2)} \to \widehat{M},\ \ \Phi_*[\xi_\ell] = [\Phi_*\xi_\ell],\ \ \Phi_*\xi_\ell := \xi_\ell \circ \Phi \equiv [\xi_{\ell,ij} \circ \Phi]_{i,j=-\ell}^{\ell}, \tag{28}$$

for $\ell \in \frac{1}{2}\mathbb{N}_0$, is a well defined isomorphism. We have used that $\Phi_* : C^\infty(\mathrm{SU}(2)) \to C^\infty(\mathbb{M})$ extends to a linear unitary bijection from $L^2(\mathrm{SU}(2))$ into $L^2(\mathbb{M})$, via

$$\langle g, h\rangle_{L^2(\mathbb{M})} = \langle g \circ \Phi, h \circ \Phi\rangle_{L^2(\mathrm{SU}(2))} = \langle \Phi_*(g), \Phi_*(h)\rangle_{L^2(\mathrm{SU}(2))}. \tag{29}$$

As it was pointed out in [40], this immediately implies that the whole construction of matrix-valued symbols on $\mathbb{M}$ is equivalent to that on SU(2). Because we can

endowed to $\mathbb{M}$ with a Lie group structure, Theorem 1 can be used to analise the index of Toeplitz operators on $L^2(\mathbb{M})$.

4.2 Toeplitz Operators

Theorem 7 implies the following result.

Corollary 1 *Let us consider a smooth function and invertible function ω on $\mathbb{M}$. Let $\Pi : L^2(\mathbb{M}) \to \Pi'$ be the orthogonal projection of a closed subspace $\Pi' \subset L^2(\mathbb{M})$. If $T_\omega = \Pi M_\omega \Pi$ is the Toeplitz operator with symbol ω, then T_ω extends to a extends to a Fredhlom operator on $L^2(\mathbb{M})$ and its index is given by*

$$\operatorname{ind}(T_\omega) = \int_{\mathbb{M}} \sum_{\ell \in \frac{1}{2}\mathbb{N}_0} (2\ell+1)\operatorname{Tr}[\sigma_{\omega^{-1}[\Pi,\omega][\Pi,\omega^{-1}][\Pi,\omega][\Pi,\omega^{-1}][\Pi,\omega]}(x)\xi_\ell(e_{\mathbb{M}}))]dx, \tag{30}$$

where $e_{\mathbb{M}} = \Phi(e_{\mathrm{SU}(2)})$.

Proof With the product defined in (26), $\mathbb{M}$ has the Lie group structure diffeomorphic to that on SU(2). The previous Ruzhansky-Turunen construction gives $\widehat{\mathbb{M}} \simeq \frac{1}{2}\mathbb{N}_0$, and every unitary and strongly continuous unitary representation on $\widehat{\mathbb{M}}$ has the form $\Phi_*\xi_\ell : \mathbb{M} \to \mathrm{U}(\mathbb{C}^{d_{\xi_\ell}})$, $\ell \in \frac{1}{2}\mathbb{N}_0$, $\dim(\Phi_*\xi_\ell) = d_{\xi_\ell} = 2\ell+1$. So, the proof now follows from Theorem 1. □

Now, we study the previous formula in local coordinates.

Remark By using the diffeomorphism $\varrho : \mathbb{M} \simeq \mathrm{SU}(2) \to \mathbb{S}^3$, defined by

$$\varrho(z) = x := (x_1, x_2, x_3, x_4), \quad \text{for} \quad z = \begin{bmatrix} x_1 + ix_2 & x_3 + ix_4 \\ -x_3 + ix_4 & x_1 - ix_2 \end{bmatrix}, \tag{31}$$

we have

$$\begin{aligned} &\operatorname{ind}(T_\omega) \\ &= \int_{\mathbb{M}} \sum_{\ell \in \frac{1}{2}\mathbb{N}_0} (2\ell+1)\operatorname{Tr}[\sigma_{\omega^{-1}[\Pi,\omega][\Pi,\omega^{-1}][\Pi,\omega][\Pi,\omega^{-1}][\Pi,\omega]}(z)\xi_\ell(e_{\mathbb{M}}))]dz, \\ &= \int_{\mathbb{S}^3} \sum_{\ell \in \frac{1}{2}\mathbb{N}_0} (2\ell+1)\operatorname{Tr}[\sigma_{\omega^{-1}[\Pi,\omega][\Pi,\omega^{-1}][\Pi,\omega][\Pi,\omega^{-1}][\Pi,\omega]}(x)\xi_\ell(e_{\mathbb{M}}))]d\tau(x), \end{aligned}$$

where

$$\sigma_{\omega^{-1}[\Pi,\omega][\Pi,\omega^{-1}][\Pi,\omega][\Pi,\omega^{-1}][\Pi,\omega]}(\varrho^{-1}(x)) =: \sigma_{\omega^{-1}[\Pi,\omega][\Pi,\omega^{-1}][\Pi,\omega][\Pi,\omega^{-1}][\Pi,\omega]}(x),$$

and $z = \varrho^{-1}(x)$. If we use the parametrization of $\mathbb{S}^3$ defined by $x_1 := \cos(\frac{t}{2})$, $x_2 := \nu$, $x_3 := (\sin^2(\frac{t}{2}) - \nu^2)^{\frac{1}{2}} \cos(s)$, $x_4 := (\sin^2(\frac{t}{2}) - \nu^2)^{\frac{1}{2}} \sin(s)$, where

$$(t, \nu, s) \in D := \{(t, \nu, s) \in \mathbb{R}^3 : |\nu| \leq \sin(\frac{t}{2}),\ 0 \leq t, s \leq 2\pi\},$$

then $d\tau(x) = \sin(\frac{t}{2}) d\nu dt ds$, and

$$\mathrm{ind}(T_\omega) = \int_0^{2\pi}\int_0^{2\pi}\int_{-\sin(t/2)}^{\sin(t/2)} \sum_{\ell \in \frac{1}{2}\mathbb{N}_0} (2\ell+1)\mathrm{Tr}[\sigma_{\omega^{-1}[\Pi,\omega][\Pi,\omega^{-1}][\Pi,\omega][\Pi,\omega^{-1}][\Pi,\omega]}(\nu, t, s)] \times \sin(\frac{t}{2}) d\nu dt ds.$$

Thus, we have obtained an explicit index formula for Toeplitz operators on compact 3-manifolds (with trivial fundamental group). □

References

1. Atiyah, M.F., Bott, R.: The index problem for manifolds with boundary. Differential Analysis, Bombay Colloq. pp. 175–186. Oxford University Press, London (1964)
2. Atiyah, M.F., Singer, I.M. .: The index of elliptic operators on compact manifolds. Bull. Am. Math. Soc. **69**, 422–433 (1963)
3. Atiyah, M.F., Singer, I.M.: The index of elliptic operators. I. Ann. Math. **87**, 484–530 (1968)
4. Atiyah, M.F., Segal, G.: The index of elliptic operators. II. Ann. Math. **87**, 531–545 (1968)
5. Atiyah, M.F., Singer, I.M.: The index of elliptic operators. III. Ann. Math. **87**, 546-604 (1968)
6. Atiyah, M.F., Singer, I.M.: Index theory for skew-adjoint Fredholm operators. Inst. Hautes SL'tudes Sci. Publ. Math. **37**, 5–26 (1969)
7. Atiyah, M.F., Singer, I.M.: The index of elliptic operators. IV. Ann. Math. **93**, 119–138 (1971)
8. Atiyah, M.F., Singer, I.M.: The index of elliptic operators. V. Ann. Math. **93**, 139–149 (1971)
9. Atiyah, M., Bott, R., Patodi, V.K.: On the heat equation and the index theorem. Invent. Math. **19**, 279–330 (1973)
10. Boutet de Monvel, L.: On the index of Toeplitz operators of several complex variables. Invent. Math. **50**, 249–272 (1979)
11. Boutet de Monvel, L.: Symplectic cones and Toeplitz operators (congrès en l'honneurde Trèves, Sao Carlos). Contemp. Math. **205**, 15–24 (1997)
12. Boutet de Monvel, L.: Toeplitz operators and asymptotic equivariant index. In: Modern Aspects of the Theory of Partial Differential Equations. Operator Theory: Advances and Applications. Advanced Partial Differential Eqtions, vol. 216, pp. 1–16. Birkhäuser, Basel (2011)
13. Boutet de Monvel, L., Guillemin, V.: The Spectral Theory of Toeplitz Operators. Annals of Mathematics Studies, vol. 99. Princeton University Press, Princeton (1981)
14. Cardona, D.: On the index of pseudo-differential operators on compact Lie groups. J. Pseudo. Differ. Oper. Appl. (to appear). https://doi.org/10.1007/s11868-018-0261-0. arXiv:1805.10404
15. Connes, A.: Noncommutative differential geometry. Publ. Math. IHES **6**, 257–360 (1985)
16. Delgado, J., Ruzhansky, M.: L^p-nuclearity, traces, and Grothendieck-Lidskii formula on compact Lie groups. J. Math. Pures Appl. **(9)**, 153–172 (2014)

17. Delgado, J., Ruzhansky, M.: Schatten classes on compact manifolds: Kernel conditions. J. Funct. Anal. **267**(3), 772–798 (2014)
18. Delgado, J., Ruzhansky, M.: Kernel and symbol criteria for Schatten classes and r-nuclearity on compact manifolds. C. R. Acad. Sci. Paris. Ser. I. **352**, 779–784 (2014)
19. Douglas, R.: Banach algebra techniques in the theory of Toeplitz operators. In: SBMS Regional Conference Series in Mathematics. AMS, Providence (1973)
20. Duistermaat J.J., Kolk, J.A.C.: Lie Groups. Universitext. Springer, Berlin (2000)
21. Dynin, A.S.: An Algebra of pseudo-differential operators on the Heisenberg group. Dokl. Akad. Nauk SSSR 227 (1976). $n°4$ Soviet Math. Dokl. 17, $n°2$, 508–512 (1976)
22. Guillemin, V.: Toeplitz operators in n dimensions. Integral Equ. Oper. Theory **7**, 145–205 (1984)
23. Hörmander, L.: The Analysis of the Linear Partial Differential Operators, vol. III. Springer, Berlin (1985)
24. Hong, S.: A Lie-algebraic approach to the local index theorem on compact homogeneous spaces. Adv. Math. **296**, 127–153 (2016)
25. Higson, N.: The index theorem of Connes and Moscovici. In: Surveys in Noncommutative Geometry. Clay Mathematics Proceedings, vol. 6, pp. 71–126. American Mathematical Society, Providence (2006)
26. Morgan, J.W.: Recent progress on the Poincaré conjecture and the classification of 3-manifolds. Bull. Am. Math. Soc. **42**, 57–78 (2005)
27. Murphy, G.J.: Spectral and index theory for Toeplitz operators. Proc. Royal Irish Acad. **91**, 1–6 (1991)
28. Murphy, G.J.: An index theorem for Toeplitz operators. J. Oper. Theory **29**, 97–114 (1993)
29. Murphy, G.H.: Representation and index theory for Toeplitz operators. Trans. Am. Math. Soc. **362**(8), 3911–3946 (2010)
30. Noether, F.: Über eine Klasse singulärer Integralgleichungen. Math. Ann. **82**, 42–63 (1921)
31. Perelman, G.: Spaces with curvature bounded below. In: Proceedings of ICM 1994, pp. 517–525. Birkhäuser, Basel (1995). MR1403952 (97g:53055)
32. Perelman, G.: The entropy formula for the Ricci flow and its geometric applications (2002). arXiv.math.DG/0211159
33. Perelman, G.: Ricci flow with surgery on three-manifolds (2003). arXiv.math.DG/0303109
34. Perelman, G.: Finite extinction time for the solutions to the Ricci flow on certain three-manifolds (2003). arXiv.math.DG/0307245
35. Poincaré, H.: Cinquième complément *à* l'analysis situs. Rend. Circ. Mat. Palermo **18**, 45–110 (1904). (See Oeuvres, Tome VI, Paris, 1953, p. 498.) MR1401792 (98m:01041)
36. Rodsphon, R.: Extensions, cohomologie cyclique et théorie de l'indice. Ph.D. Thesis, École Doctorale Informatique et Mathématiques de Lyon (2014)
37. Roe, J.: Elliptic Operators, 2nd edn. Addison Wesley, Boston (1998)
38. Ruzhansky, M., Turunen, V.: On the Fourier analysis of operators on the torus. In: Modern Trends in Pseudo-Differential Operators. Operator Theory: Advances and Applications, vol. 172, pp. 87–105. Birkhäuser, Basel (2007)
39. Ruzhansky, M., Turunen, V.: On pseudo-differential operators on group SU(2). In: New Developments in Pseudo-Differential Operators. Operator Theory: Advances and Applications, vol. 189, pp. 307–322. Birkhäuser, Basel (2009)
40. Ruzhansky, M., Turunen, V.:Pseudo-Differential Operators and Symmetries: Background Analysis and Advanced Topics. Birkhäuser, Basel, (2010)
41. Ruzhansky, M., Turunen, V.: Global quantization of pseudo-differential operators on compact Lie groups, SU(2), 3-sphere, and homogeneous spaces. Int. Math. Res. Not. **11**, 2439–2496 (2013). http://dx.doi.org/10.1093/imrn/rns122
42. Ruzhansky, M., Turunen, V., Wirth, J.: Hormander class of pseudo-differential operators on compact Lie groups and global hypoellipticity. J. Fourier Anal. Appl. **20**, 476–499 (2014)
43. Upmeier, H.: Fredholm indices for Toeplitz operators on bounded symmetric domains. Am. J. Math. **110**, 811–832 (1988)

44. Upmeier, H.: Toeplitz Operators and Index Theory in Several Complex Variables. Birkhäuser Verlag, Basel (1996)
45. Venugopalkrishna, V.: Fredholm operators associated with strongly pseudoconvex domains in $\mathbb{C}^n$. J. Funct. Anal. **9**, 349–373 (1972)

Part IX
Partial Differential Equations on Curved Spacetimes

Lorentzian Spectral Zeta Functions on Asymptotically Minkowski Spacetimes

Nguyen Viet Dang and Michał Wrochna

Abstract In this note, we consider perturbations of Minkowski space as well as more general spacetimes on which the wave operator $\Box_g$ is essentially self-adjoint. We review a recent result which gives the meromorphic continuation of the Lorentzian spectral zeta function density, i.e. of the trace density of complex powers $\alpha \mapsto (\Box_g - i\varepsilon)^{-\alpha}$. In even dimension $n \geq 4$, the residue at $\frac{n}{2} - 1$ is shown to be a multiple of the scalar curvature in the limit $\varepsilon \to 0^+$. This yields a spectral action for gravity in Lorentzian signature.

1 Main Result

1.1 Motivation

Suppose (M, g) is a compact Riemannian manifold of dimension n, and let $\triangle_g$ be the Laplace–Beltrami operator. A classical result in analysis, dating back to Minakshisundaram–Pleijel [23] and Seeley [26], states that for $\operatorname{Re}\alpha > \frac{n}{2}$ the trace density of $(-\triangle_g)^{-\alpha}$, defined as the on-diagonal restriction

$$(-\triangle_g)^{-\alpha}(x, x) \tag{1}$$

of the Schwartz kernel $(-\triangle_g)^{-\alpha}(x, y)$, exists for all $x \in M$. Furthermore (1) extends to a density-valued meromorphic function of the complex variable α. Its integral over M is the celebrated *spectral zeta function* of $-\triangle_g$ (or

N. V. Dang
Institut de Mathématiques de Jussieu (UMR 7586, CNRS), Sorbonne Université, Université de Paris, Paris, France
e-mail: nguyen-viet.dang@imj-prg.fr

M. Wrochna (✉)
Laboratoire AGM (UMR 8088, CNRS), CY Cergy Paris Université, Cergy-Pontoise, France
e-mail: michal.wrochna@cyu.fr

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_38

Minakshisundaram–Pleijel zeta function), which has attracted widespread attention due to its relationships with the geometry of (M, g).

In fact, the residues of (1) are given by local geometric quantities: in particular if $n \geq 4$ is even, one finds

$$\operatorname*{res}_{\alpha=\frac{n}{2}-1} (-\triangle_g)^{-\alpha}(x,x) = \frac{R_g(x)}{6(4\pi)^{\frac{n}{2}}\Gamma(\frac{n}{2}-1)}, \tag{2}$$

where $R_g(x)$ is the scalar curvature of (M, g) at $x \in M$. This identity, often attributed to Kastler [20] and Kalau–Walze [19], and announced previously by Connes, is a consequence of classical theorems in elliptic theory (the heat kernel based argument can be found in [7, Thm. 1.148]; see [16, §1.7] for an approach in the spirit of Atiyah–Bott–Patodi [1]). Its importance in physics stems from the fact that the variational equation $\delta_g R_g = 0$ for g is equivalent to the Einstein equations in Riemannian signature. Therefore, the l.h.s. of (2) yields a *spectral action* for Euclidean gravity. Relationships of this type have also been used to justify definitions of curvature in non-commutative geometry [7, 8].

However, it is the Einstein equations in *Lorentzian* signature which have a direct physical meaning. This means that (M, g) should be replaced by a Lorentzian manifold (typically not compact), but then the problem is that the corresponding Laplace–Beltrami operator $\Box_g$ (or *wave operator*), is not elliptic nor bounded from below. In consequence, it is not at all clear if $\Box_g$ has a self-adjoint extension and even less clear if the arguments from elliptic theory can be somehow replaced (for instance, it is difficult to imagine that the heat kernel could be usefully generalized to the Lorentzian setting).

Nevertheless, it was demonstrated by Vasy [33] (followed by a generalization by Nakamura–Taira [25]) that if (M, g) is well-behaved at infinity, $\Box_g$ *is* essentially self-adjoint in $L^2(M, g)$. Consequently, complex powers $(\Box_g - i\varepsilon)^{-\alpha}$ can be defined by functional calculus for any $\varepsilon > 0$. The question is then if this global, spectral theoretical object has anything to do with the local geometry of (M, g), in particular with the Lorentzian scalar curvature $R_g(x)$.

1.2 Main Theorem

In [9] we consider Vasy's framework and provide an affirmative answer in the form of an identity largely analogous to (2). Namely, we prove the following theorem.

Theorem 1 ([9, Thm. 1.1]) *Assume (M, g) is a globally hyperbolic, non-trapping Lorentzian scattering space of even dimension $n \geq 4$. For all $\varepsilon > 0$, the Schwartz kernel of $(\Box_g - i\varepsilon)^{-\alpha}$ has for $\operatorname{Re}\alpha > \frac{n}{2}$ a well-defined on-diagonal restriction*

$(\Box_g - i\varepsilon)^{-\alpha}(x,x)$, *which extends as a meromorphic function of* $\alpha \in \mathbb{C}$ *with poles at* $\{\frac{n}{2}, \frac{n}{2}-1, \frac{n}{2}-2, \ldots, 1\}$. *Furthermore,*

$$\lim_{\varepsilon\to 0^+} \operatorname*{res}_{\alpha=\frac{n}{2}-1} \left(\Box_g - i\varepsilon\right)^{-\alpha}(x,x) = \frac{R_g(x)}{i6(4\pi)^{\frac{n}{2}}\Gamma\left(\frac{n}{2}-1\right)}, \tag{3}$$

where $R_g(x)$ *is the scalar curvature at* $x \in M$.

The meromorphic continuation of $\alpha \mapsto \zeta_{g,\varepsilon}(\alpha)(x) := \left(\Box_g - i\varepsilon\right)^{-\alpha}(x,x)$ is called the *Lorentzian spectral zeta function density* of (M, g).

Let us briefly discuss the assumptions of Theorem 1. The class of *non-trapping Lorentzian scattering spaces* introduced by Vasy [33] can be thought of having asymptotically the same structure as Minkowski space at spacetime infinity $|x| \to +\infty$, with the extra requirement that there are no trapped null geodesics. It is worth emphasizing that this is a somewhat more general class than what one would typically call "asymptotically Minkowski spacetime" in that the definition refers to the bicharacteristic flow (the null geodesic flow lifted to the cotangent bundle) and to its asymptotic properties, rather than to the precise form of the metric coefficients at infinity, see [9, 33]. *Global hyperbolicity* is a standard assumption which provides a general setting for well-posedness of the Cauchy problem for $\Box_g$ and is unlikely to entail significant loss of generality (in fact, it automatically follows from the non-trapping assumption for a large class of asymptotically Minkowski spacetimes, see [14, §4.2]).

The most essential feature of these assumptions is that they allow for perturbations of Minkowski space without assuming any particular symmetries or analyticity. This means there are sufficiently many variations of the metric to derive Einstein equations from the r.h.s. of (3). Consequently, the l.h.s. gives a spectral action for gravity in Lorentzian signature.

1.3 Further Results

Let us also briefly mention several of our further results related to Theorem 1.

In [9] we show an expansion in the spirit of the Chamseddine–Connes spectral action [5, 6]. Namely, for any Schwartz function f with Fourier transform $\widehat{f}$ supported in $]0, +\infty[$ and any $N \in \mathbb{N}_{\geqslant 0}$, we have for $\varepsilon > 0$ the large $\lambda > 0$ expansion

$$f\left((\Box_g + i\varepsilon)/\lambda^2\right)(x,x) = \sum_{j=0}^{N} \lambda^{n-2j} C_j(f)\, a_j(x) + \mathcal{O}(\varepsilon, \lambda^{n-2N-1}), \tag{4}$$

where each $C_j(f)$ depends only on $j \in \mathbb{N}_{\geqslant 0}$, the space-time dimension n and f, and $a_j(x)$ are directly related to the *Hadamard coefficients*, in particular

$$a_0(x) = (4\pi)^{-\frac{n}{2}}, \quad a_1(x) = -(4\pi)^{-\frac{n}{2}} \frac{1}{6} R_g(x),$$

with $C_0(f) = i^{-1} e^{\frac{in\pi}{4}} \int_0^\infty \widehat{f}(t) t^{\frac{n}{2}-1} dt$ and $C_1(f) = i^{-1} e^{\frac{i(n-2)\pi}{4}} \int_0^\infty \widehat{f}(t) t^{\frac{n}{2}-2} dt$.

Furthermore, we show that the identities (3)–(4) remain valid in the case of *ultrastatic spacetimes* (M, g), meaning that $M = \mathbb{R} \times Y$ and $g = dt^2 - h$ for some t-independent complete Riemannian manifold (Y, h). In this setting essential self-adjointness is due to Dereziński–Siemssen [11] and the proofs are significantly simpler because the spectral theory of $-\triangle_h$ can then be used. We remark that in the related and more general case of *stationary spacetimes* the scalar curvature can be recovered in a different spectral-theoretical way through a Gutzwiller–Duistermaat–Guillemin trace formula due to Strohmaier–Zelditch [30].

Finally, in a further work [10] we define a dynamical notion of "residue" which generalizes the *Guillemin-Wodzicki residue density* [17, 34] of pseudo-differential operators. More precisely, given a Schwartz kernel, our definition refers to the *Pollicott–Ruelle resonances* for the dynamics of scaling towards the diagonal in $M \times M$. We apply this formalism to complex powers $(\Box_g - i\varepsilon)^{-\alpha}$ and we demonstrate that residues of Lorentzian spectral zeta functions $\zeta_{g,\varepsilon}(\alpha)$ are dynamical residues indeed. This provides a Lorentzian version of the fact that the residue (2) can be expressed as a Guillemin–Wodzicki residue or, in physicists' terminology, a "scaling anomaly".

2 Sketch of Proof

2.1 From Resolvent to Complex Powers

Let us now give a sketch of the proof of Theorem 1. Let $P = \Box_g$ be the wave operator, i.e., using the notation $|g| = |\det g|$, P is the differential operator

$$\begin{aligned} P &= |g(x)|^{-\frac{1}{2}} \partial_{x^j} |g(x)|^{\frac{1}{2}} g^{jk}(x) \partial_{x^k} \\ &= \partial_{x^j} g^{jk}(x) \partial_{x^k} + b^k(x) \partial_{x^k} \end{aligned} \tag{5}$$

where we sum over repeated indices, and $b^k(x) = |g(x)|^{-\frac{1}{2}} g^{jk}(x) (\partial_{x^j} |g(x)|^{\frac{1}{2}})$. We use the same notation P for the closure of $\Box_g$ acting on test functions $C_c^\infty(M) \subset L^2(M, g)$.

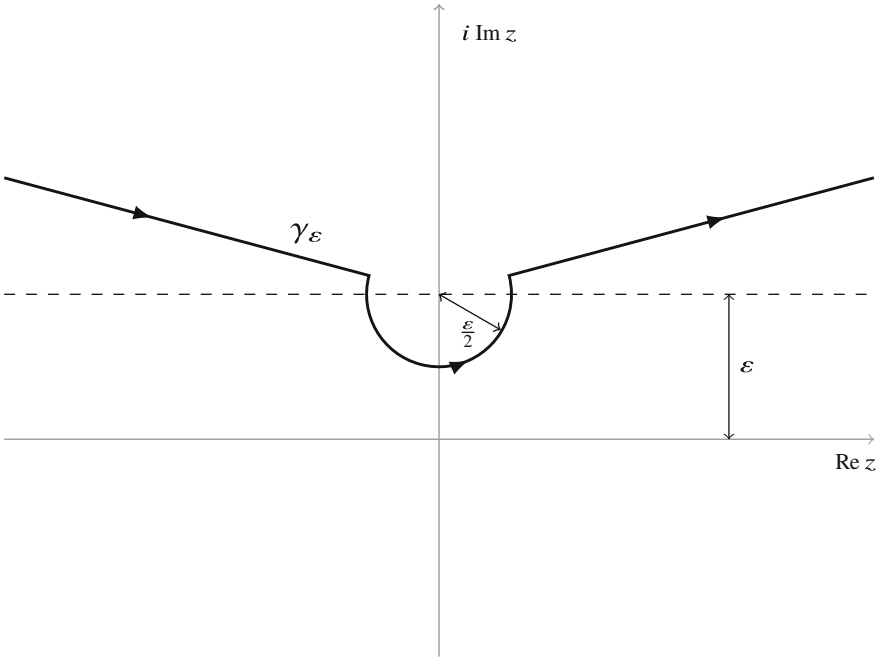


Fig. 1 The contour γ_ε used to express $(P - i\varepsilon)^{-\alpha}$ as an integral of the resolvent $(P - z)^{-1}$

For $\varepsilon > 0$ and $\mathrm{Re}\,\alpha > 0$, the power $(P - i\varepsilon)^{-\alpha}$ can be expressed as a contour integral of the form

$$(P - i\varepsilon)^{-\alpha} = \frac{1}{2\pi i} \int_{\gamma_\varepsilon} (z - i\varepsilon)^{-\alpha} (P - z)^{-1} dz, \tag{6}$$

convergent in the strong operator topology (see e.g. [9, App. B]). The contour of integration γ_ε is represented in Fig. 1 and can be written as $\gamma_\varepsilon = \tilde{\gamma}_\varepsilon + i\varepsilon$, where

$$\tilde{\gamma}_\varepsilon = e^{i(\pi-\theta)} \left]-\infty, \tfrac{\varepsilon}{2}\right] \cup \{\tfrac{\varepsilon}{2} e^{i\omega} \,|\, \pi - \theta < \omega < \theta\} \cup e^{i\theta} \left[\tfrac{\varepsilon}{2}, +\infty\right[\tag{7}$$

goes from $\mathrm{Re}\, z \ll 0$ to $\mathrm{Re}\, z \gg 0$ in the upper half-plane (for some fixed $\theta \in \left]0, \frac{\pi}{2}\right[$).

The strategy is then to construct a sufficiently explicit parametrix for the resolvent $(P - z)^{-1}$. When estimating error terms, a significant difficulty is the necessity to control what happens *uniformly in* z, with an appropriate decay rate along the infinite contour γ_ε. We remark that retarded and advanced propagators for $P - z$ are *not* expected to have this kind of decay, so in practice it is not possible to use various techniques from hyperbolic PDEs related to solving a retarded or advanced problem or a Cauchy problem for $P - z$.

2.2 Uniform Hadamard Parametrix

In contrast to the heat kernel, the *Hadamard parametrix* for the Laplace–Beltrami operator generalizes well to the Lorentzian case. Furthermore, it is known to have similar local geometric content. So, the main question is whether the Hadamard parametrix approximates the resolvent $(P - z)^{-1}$ in a reasonable sense, uniformly in z along the contour γ_ε.

Before answering this question, let us recall the construction of the Hadamard parametrix for $P - z$.

As expected from explicit formulae on Minkowski space and from the theory of Fourier integral operators, there are actually four different Hadamard parametrices with different singularities. In the case of the resolvent $(P - z)^{-1}$ with $\operatorname{Im} z > 0$, one expects that the *Feynman* Hadamard parametrix is the correct choice, see e.g. [21] for the general definition. Here we use a construction directly adapted from earlier works in the Riemannian or Lorentzian time-independent case [18, 29, 35, 36] (cf. [2] for a unified treatment of even and odd dimensions), supplemented by new estimates that are uniform in z (cf. [3, 12, 28] for uniform estimates in the Riemannian case). As expected, their proof is significantly complicated by light-cone singularities not present in the Riemannian analogue of the problem.

Step 1

Let $\eta = dx_0^2 - (dx_1^2 + \cdots + dx_{n-1}^2)$ be the Minkowski metric on $\mathbb{R}^n$, and consider the corresponding quadratic form

$$|\xi|_\eta^2 = -\xi_0^2 + \textstyle\sum_{i=1}^{n-1} \xi_i^2,$$

defined for convenience with a minus sign. For $\alpha \in \mathbb{C}$ and $\operatorname{Im} z > 0$, the distribution $\left(|\xi|_\eta^2 - z\right)^{-\alpha}$ is well-defined by pull-back from $\mathbb{R}$. More generally, for $\operatorname{Im} z \geqslant 0$, the limit $\left(|\xi|_\eta^2 - z - i0\right)^{-\alpha} = \lim_{\varepsilon \to 0^+} \left(|\xi|_\eta^2 - z - i\varepsilon\right)^{-\alpha}$ from the upper half-plane is well defined as a distribution on $\mathbb{R}^n \setminus \{0\}$. If $z \neq 0$ it can be extended to a family of homogeneous[1] distributions on $\mathbb{R}^n$, holomorphic in $\alpha \in \mathbb{C}$. We introduce special notation for its appropriately normalized inverse Fourier transform,

$$F_\alpha(z, |x|_\eta) := \frac{\Gamma(\alpha + 1)}{(2\pi)^n} \int e^{i\langle x, \xi \rangle} \left(|\xi|_\eta^2 - i0 - z\right)^{-\alpha - 1} d^n\xi. \tag{8}$$

[1] Homogeneity refers here to rescaling simultaneously the ξ variables by $\lambda > 0$ and the complex number z by λ^2.

Step 2

Next, one pull-backs the distribution $F_\alpha(z, |.|_\eta)$ to a neighborhood $\mathcal{U}$ of the diagonal $\Delta \subset M \times M$ using the exponential map. In view of the $O(1, n-1)^\uparrow_+$-invariance of $F_\alpha(z, |.|_\eta)$ there is a canonical way to define this pull-back (see [9, §5.1]), denoted in the sequel by $\mathbf{F}_\alpha(z, .)$. The Hadamard parametrix (or rather its Schwartz kernel) is constructed in normal charts using the family $\mathbf{F}_\alpha(z, .)$. Namely, for fixed $x_0 \in M$, one expresses the distribution $x \mapsto \mathbf{F}_\alpha(z, x_0, x)$ in normal coordinates centered at x_0, defined on some $U \subset T_{x_0}M$. By abuse of notation we continue to write $\mathbf{F}_\alpha(z, .)$ instead of $\mathbf{F}_\alpha(z, x_0, \exp_{x_0}(\cdot)) \in \mathcal{D}'(U)$. One then defines for large N a parametrix $H_N(z, .)$ by setting

$$H_N(z, .) = \sum_{k=0}^{N} u_k \mathbf{F}_k(z, .) \in \mathcal{D}'(U), \tag{9}$$

where $(u_k)_{k=0}^\infty$ is a sequence of functions in $C^\infty(U)$ that solves the hierarchy of transport equations

$$2ku_k + b^i(x)\eta_{ij}x^j u_k + 2x^i \partial_{x^i} u_k + 2Pu_{k-1} = 0 \tag{10}$$

with initial condition $u_0(0) = 1$ (by convention, $u_{k-1} = 0$ for $k = 0$, we sum over repeated indices, and we recall that $b^i(x)$ is defined in (5)). The transport equations imply that $H_N(z, .)$ solves

$$(P - z)\, H_N(z, .) = |g|^{-\frac{1}{2}} \delta_0 + (Pu_N)\mathbf{F}_N, \tag{11}$$

on U, where $(Pu_N)\mathbf{F}_N$ is interpreted as an error term.

Step 3

In the final step one takes into account the dependence on x_0 to obtain a parametrix on the neighborhood $\mathcal{U}$ of the diagonal. Here we make this step implicitly by sticking to the same notation $\mathbf{F}_\alpha(z, .)$ for the corresponding distribution on $\mathcal{U}$. Finally, one uses a cutoff function $\chi \in C^\infty(M^2)$ supported in $\mathcal{U}$ (with $\chi = 1$ near the diagonal) to extend the definition of $H_N(z, .)$ to M^2:

$$H_N(z, .) = \sum_{k=0}^{N} \chi u_k \mathbf{F}_k(z, .) \in \mathcal{D}'(M \times M).$$

The Hadamard parametrix extended to M^2 satisfies

$$(P - z)\, H_N(z, .) = |g|^{-\frac{1}{2}} \delta_\Delta + (Pu_N)\mathbf{F}_N(z, .)\chi + r_N(z, .), \tag{12}$$

where $|g|^{-\frac{1}{2}}\,\delta_\Delta(x_1,x_2)$ is the Schwartz kernel of the identity map and $r_N(z,.)\in\mathcal{D}'(M\times M)$ is an error term supported in a punctured neighborhood of Δ which is due to the presence of the cutoff χ.

In order to conclude a relationship between the resolvent $(P-z)^{-1}$ and the Hadamard parametrix $H_N(z,.)$, the natural next step is to apply $(P-z)^{-1}$ to both sides of (12). The objective is then to show that the composition of $(P-z)^{-1}$ with the two error terms on the r.h.s. exists, decreases in z in a suitable sense along the contour γ_ε, and is sufficiently regular (so that its on-diagonal restriction always exists and the corresponding integral on γ_ε is holomorphic in α).

It turns out that by choosing N sufficiently high we can make $(Pu_N)\mathbf{F}_N(z,.)\chi$ decaying in z and of arbitrarily high Hölder regularity. The proof is quite technical as it uses oscillatory integral representations, but most of the analysis is carried out on the level of the explicit model family $F_\alpha(z,|.|_\eta)$. In combination with regularity properties of $(P-z)^{-1}$ obtained as a corollary of Vasy's proof of essential self-adjointness, this yields an easily controllable error term.

On the other hand, the error term $r_N(z,.)$ (although it can be arranged to be supported away from the diagonal) is always *singular* regardless of the choice of N. This stands in sharp contrast with analogous constructions in the Riemannian case and is the most significant obstacle in the proof: a priori it is not even clear if the composition $(P-z)^{-1}r_N$ makes sense.

A way out is possible thanks to a remarkable property shared by the Feynman Hadamard parametrix and $(P-z)^{-1}$ when $\operatorname{Im} z>0$. Their Schwartz kernels are singular, but in a special way which allows operator composition nevertheless, and which implies that the compositions have singularities of the same type. Microlocally, they behave as the *Feynman propagator* on Minkowski space, i.e. the Fourier multiplier by $(-\xi_0^2+\xi_1^2+\cdots+\xi_{n-1}^2-i0)^{-1}$. This condition can be formulated in terms of an operatorial Sobolev wavefront set $\mathrm{WF}'^{(s)}\big((P-z)^{-1}\big)$ for large $s\in\mathbb{R}$: by definition, a pair of points $(q_1,q_2)\in(T^*M\setminus o)^{\times 2}$ does *not* belong to $\mathrm{WF}'^{(s)}\big((P-z)^{-1}\big)$ if there exists pseudo-differential operators $B_1,B_2\in\Psi^0(M)$, elliptic at respectively q_1,q_2, such that

$$B_1(P-z)^{-1}B_2^*:H^m_{\mathrm{c}}(M)\to H^{m+s}_{\mathrm{loc}}(M)$$

is bounded for all $m\in\mathbb{R}$. A uniform version particularly well adapted to our needs can be defined by requiring that the operator semi-norms are $O(\langle z\rangle^{-\frac{1}{2}})$ in z along the contour γ_ε.

For fixed z, it is relatively easy to find the wavefront set of $H_N(z)$, and one could try various existing techniques to estimate the wavefront set of $(P-z)^{-1}$. However, estimates on the *uniform* wavefront set are needed to control the contributions of the error terms after integration. The uniform wavefront set of r_N is obtained from a detailed Hölder regularity analysis of oscillatory integral representations with the help of dyadic decompositions. The uniform wavefront set of $(P-z)^{-1}$ is estimated in several steps outlined in the next paragraphs, with a central role played by *microlocal propagation estimates* including *radial estimates*. Uniform regularity

of the composition $(P - z)^{-1} r_N$ is then deduced from the two uniform wavefront sets and the property that r_N is supported away from the diagonal in $M \times M$.

2.3 Uniform Microlocal Resolvent Estimates

In the estimation of the uniform wavefront set of $(P - z)^{-1}$, the first step is to construct a parametrix $G_z = G_z^+ + G_z^-$ for $(P - z)^{-1}$, which consists of two terms $G_z^\pm$ that correspond each to solving an evolution problem of *first order in time*. This parametrix is used as reference operator with more easily computable wavefront set.

The construction relies on an approximate factorization of $P - z$. Namely, we show that after a suitable coordinate change φ and a conformal transformation by some smooth factor $c > 0$ (this step uses global hyperbolicity), $P - z$ can be written in the form

$$\begin{aligned} -c^2(\varphi^*(P - z)) &= (D_t - A(t, z))(D_t + B(t, z)) + R(t, z) \\ &= (D_t + \tilde{B}(t, z))(D_t - \tilde{A}(t, z)) + \tilde{R}(t, z), \end{aligned}$$

where $A(t, z), B(t, z), \tilde{A}(t, z), \tilde{B}(t, z) \in \Psi^1(M)$ are smooth (in t) families of pseudo-differential operators which are *elliptic with parameter* in the sense of Shubin's parameter-dependent calculus [27], with positive principal symbols, and $R(t, z), \tilde{R}(t, z)$ are smooth families of operators with arbitrarily good regularity properties, uniformly in z. The operators $G_z^\mp$ are defined through an expression which uses the retarded problem of $D_t - \tilde{A}(t, z)$, resp. advanced problem for $D_t + B(t, z)$ (these are the only two that are well-behaved for large $\operatorname{Im} z > 0$). As such, their uniform wavefront sets can be estimated by arguments closely related to Egorov's theorem.

The problem is then how to demonstrate that $G_z = G_z^+ + G_z^-$ and $(P - z)^{-1}$ have the same uniform wavefront set. Since the wavefront sets of G_z^+ and G_z^- are disjoint (they propagate singularities in the two different components $\Sigma^\mp$ of the characteristic set of P), it actually suffices to estimate the wavefront set of $(P - z)^{-1} - G_z^\pm$. The key ingredient are *microlocal propagation estimates*, which can be applied if we have some microlocal regularity of $(P - z)^{-1} - G_z^\pm$ to start with.

It turns out that there is indeed a significant property shared by $(P - z)^{-1}$ and $G_z^\pm$. Let us first explain it in the case of the resolvent $(P - z)^{-1}$. Its mapping properties are best understood in the framework of anisotropic *scattering Sobolev spaces* $H_{\mathrm{sc}}^{s,\ell}(M)$: these spaces generalize the weighted Sobolev spaces $(1+|x|^2)^{-\ell/2} H^s(\mathbb{R}^n)$ in a way that allows the weight orders ℓ to vary in phase space (more specifically, on Melrose's *scattering bundle* ${}^{\mathrm{sc}}T^*M$ [22], which in our context

provides the natural framework for microlocal analysis on the compactification of M). The key ingredient in Vasy's proof of essential self-adjointness is a *Fredholm estimate* of the form

$$\|u\|_{s,\ell} + (\operatorname{Im} z)\|u\|_{s-\frac{1}{2},\ell+\frac{1}{2}} \leqslant C(\|(P-z)u\|_{s-1,\ell+1} + \|u\|_{s,L}), \tag{13}$$

uniformly for $z \in \gamma_\varepsilon$ (with $\|u\|_{s,L}$ representing a negligible error term). Here, ℓ is chosen monotone along the bicharacteristic flow, in such a way that $\ell > -\frac{1}{2}$ at *sources at infinity* (from which bicharacteristics are assumed to originate), and $\ell < -\frac{1}{2}$ at *sinks at infinity* (to which bicharacteristics tend). One can think of this condition as imposing boundary conditions at infinity: solving $(P-z)u = f$ in the corresponding spaces is then a *Feynman problem* [13, 15, 31, 33]. The estimate (13) is responsible for the fact that if $f = (P-z)u$ is compactly supported then it *decays at a rate faster than the threshold value* $-\frac{1}{2}$ microlocally at the sources. This statement can be improved in various ways, and $(P-z)^{-1}f$ has of course even better decay properties. The key point is that within $\Sigma^{\mp}$, $G^{\pm}f$ is decaying at the same source as $(P-z)^{-1}f$. Therefore, $((P-z)^{-1} - G_z^{\pm})f$ decays at the sources microlocally in the respective component, and this property enables the use of *radial estimates* in Melrose's scattering calculus $\Psi_{\mathrm{sc}}(M)$ [22, 32, 33], to get high regularity of $A((P-z)^{-1} - G_z^{\pm})f$ if $A \in \Psi_{\mathrm{sc}}^{0,0}(M)$ is microsupported near sources in the respective component (incidentally, these are the same estimates which are used to prove (13)). Then, propagation of singularities and elementary manipulations with operatorial wavefront sets are used to deduce that $(P-z)^{-1} - G_z^{\pm}$ is (everywhere) smoothing. Crucially, in each step of this proof, the uniformity in z is under control.

This proves the desired estimate on the uniform wavefront set of $(P-z)^{-1}$, and as explained in Sect. 2.2, concludes the proof that $(P-z)^{-1}$ equals the Feynman Hadamard parametrix $H_N(z)$ modulo inessential terms.

2.4 Extraction of the Scalar Curvature

From that point on we can effectively replace the resolvent $(P-z)^{-1}$ with the Hadamard parametrix $H_N(z)$. In fact, if we integrate $(z-i\varepsilon)^{-\alpha}H_N(z)$ over the contour γ_ε instead of $(z-i\varepsilon)^{-\alpha}(P-z)^{-1}$, the result will differ from $(P-i\varepsilon)^{-\alpha}$ merely by a term whose trace density is holomorphic in α.

The integral turns out to be of the same form as the Hadamard expansion. More precisely, $(P-i\varepsilon)^{-\alpha}$ equals

$$\sum_{k=0}^{N} \chi u_k \frac{(-1)^k \Gamma(-\alpha+1)}{\Gamma(-\alpha-k+1)\Gamma(\alpha+k)} \mathbf{F}_{k+\alpha-1}(-i\varepsilon, .). \tag{14}$$

plus the irrelevant error term. The meromorphic properties of the on-diagonal restriction of (14) can be deduced from an analysis on $\mathbb{R}^n$ thanks to the identity $\mathbf{F}_\alpha(z,x,x) = F_\alpha(z,|0|_\eta)$, valid for every $x \in M$.

A toy example illustrating what happens in Euclidean signature is provided by the integral

$$\int_{\mathbb{R}^n} (\|\xi\|^2 - z)^{-\alpha} d^n\xi = \frac{1}{\Gamma(\alpha)} \int_0^\infty \left(\int_{\mathbb{R}^n} e^{-t(\|\xi\|^2 - z)} d^n\xi \right) t^{\alpha-1} dt,$$

assuming for simplicity $z < 0$ for the moment. It has the same poles as

$$\begin{aligned} &\frac{1}{\Gamma(\alpha)} \int_0^1 \left(\int_{\mathbb{R}^n} e^{-t(\|\xi\|^2 - z)} d^n\xi \right) t^{\alpha-1} dt \\ &= \frac{(2\pi)^n}{\Gamma(\alpha)(4\pi)^{\frac{n}{2}}} \sum_{k=0}^\infty \frac{z^k}{k!} \int_0^1 t^{\alpha - \frac{n}{2} + k - 1} dt = \frac{\pi^{\frac{n}{2}}}{\Gamma(\alpha)} \sum_{k=0}^\infty \frac{z^k}{k!(\alpha - \frac{n}{2} + k)}. \end{aligned}$$

In consequence, we see that the residue at $\alpha = k$, $k \in \{1, \ldots, \frac{n}{2} - 1\}$ is

$$\operatorname{res}_{\alpha=k} \int_{\mathbb{R}^n} (\|\xi\|^2 - z)^{-\alpha} d^n\xi = \frac{z^{\frac{n}{2}-k} \pi^{\frac{n}{2}}}{(\frac{n}{2} - k)! \Gamma(k)}. \tag{15}$$

In our problem, we need to deal with integrals involving the Minkowski quadratic form rather than the Euclidean one. To that end we consider the complex valued n-form

$$\omega_\alpha = \left(\sum_{i=1}^n \xi_i^2 - z \right)^{-\alpha} d\xi_1 \wedge \cdots \wedge d\xi_n \in \Omega^{n,0}$$

for z in the upper half-plane. We show that it is closed, and that Stokes' theorem can be applied to deform the signature from Euclidean to Lorentzian in integrated expressions, which eventually yields

$$\operatorname{res}_{\alpha=k} \int_{\mathbb{R}^n} \left(-\xi_1^2 + \sum_{i=2}^n \xi_i^2 - z - i0 \right)^{-\alpha} d^n\xi = i \operatorname{res}_{\alpha=k} \int_{\mathbb{R}^n} \left(\sum_{i=1}^n \xi_i^2 - z \right)^{-\alpha} d^n\xi,$$

where the r.h.s. is computed using (15).

By taking into account the Γ function factors in (14) we get the location of the poles of $(P - i\varepsilon)^{-\alpha}$, and the remaining ingredient in the computation of the residues are the on-diagonal restrictions $u_k(x,x)$ of the coefficients $u_k(x,y)$. These coefficients can be found for instance by observing that the transport equations for u_k are analogous to transport equations in the Riemannian setting, so they are given by analogous expressions in terms of the metric g and its derivatives, with

obvious sign changes to account for the switch of signature [4, 24]. The Riemannian transport equations are in turn directly related to transport equations for the more familiar heat kernel coefficients.

We are particularly interested in the residue at $\alpha = \frac{n}{2} - 1$ which comes from the coefficient $u_1(x, x)$, and this coefficient can also be found by an inspection of the first two transport equations, directly in Lorentzian signature. In normal coordinates (also denoted by x) centered around an arbitrary point $x_0 \in M$ (so x_0 is $x = 0$ in normal coordinates), we have the identity

$$P = \partial_{x^k} g^{kj}(x) \partial_{x^j} + g^{jk}(x)(\partial_{x^j} \log |g(x)|^{\frac{1}{2}}) \partial_{x^k}.$$

This can be used to express the transport equations in a more convenient form and one finds after a short computation that they imply

$$u_1(0) = -Pu_0(0) = -P(|g(0)|^{\frac{1}{4}} |g(x)|^{-\frac{1}{4}})|_{x=0}.$$

In normal coordinates, $|g(0)|^{\frac{1}{4}} = 1$ and

$$g_{ij}(x) = \eta_{ij} + \frac{1}{3} R_{ikjl} x^k x^l + \mathcal{O}(|x|^3), \quad |g(x)|^{-\frac{1}{4}} = 1 + \frac{1}{12} \mathbf{Ric}_{kl}(0) x^k x^l + \mathcal{O}(|x|^3),$$

where $\mathbf{Ric}_{kl}$ is the Ricci tensor. This implies that

$$-P\,|g(x)|^{-\frac{1}{4}} = -\frac{1}{6} g^{kl} \mathbf{Ric}_{kl}(0) + \mathcal{O}(|x|),$$

where $g^{kl}\mathbf{Ric}_{kl} = R_g(0)$ is the scalar curvature at x_0. Since x_0 was arbitrary, we conclude $u_1(x, x) = -\frac{1}{6} R_g(x)$.

Acknowledgments The authors gratefully acknowledge support from the grant ANR-20-CE40-0018.

References

1. Atiyah, M., Bott, R., Patodi, V.K.: On the heat equation and the index theorem. Invent. Math. **28**(3), 277–280 (1975)
2. Bär, C., Strohmaier, A.: Local index theory for Lorentzian manifolds. arXiv:2012.01364 (2020)
3. Bourgain, J., Shao, P., Sogge, C.D., Yao, X.: On L^p-resolvent estimates and the density of eigenvalues for compact Riemannian manifolds. Commun. Math. Phys. **333**(3), 1483–1527 (2015)
4. Bytsenko, A.A., Cognola, G., Moretti, V., Zerbini, S., Elizalde, E.: Analytic Aspects of Quantum Fields. World Scientific Publishing, Singapore (2003)
5. Chamseddine, A.H., Connes, A.: The spectral action principle. Commun. Math. Phys. **186**, 731–750 (1997)

6. Connes, A.: Gravity coupled with matter and the foundation of non-commutative geometry. Commun. Math. Phys. **182**(1), 155–176 (1996)
7. Connes, A., Marcolli, M.: Noncommutative Geometry, Quantum Fields and Motives. American Mathematical Society, Providence (2008)
8. Connes, A., Moscovici, H.: Modular curvature for noncommutative two-tori. J. Am. Math. Soc. **27**(3), 639–684 (2014)
9. Dang, N.V., Wrochna, M.: Complex powers of the wave operator and the spectral action on Lorentzian scattering spaces. arXiv:2012.00712 (2020)
10. Dang, N.V., Wrochna, M.: Dynamical residues of Lorentzian spectral zeta functions. arXiv:2108.07529 (2021)
11. Dereziński, J., Siemssen, D.: Feynman propagators on static spacetimes. Rev. Math. Phys. **30**, 1850006 (2018)
12. Ferreir, D.D.S., Kenig, C.E., Salo, M.: On L^p resolvent estimates for Laplace–Beltrami operators on compact manifolds. Forum Math. **26**(3), 815–849 (2014)
13. Gell-Redman, J., Haber, N., Vasy, A.: The Feynman propagator on perturbations of Minkowski space. Commun. Math. Phys. **342**(1), 333–384 (2016)
14. Gérard, C., Wrochna, M.: The massive Feynman propagator on asymptotically Minkowski spacetimes. Am. J. Math. **141**(6), 1501–1546 (2019)
15. Gérard, C., Wrochna, M.: The Feynman problem for the Klein–Gordon equation. arXiv:2003.14404 (2020)
16. Gilkey, P.: Invariance Theory: The Heat Equation and the Atiyah-Singer Index Theorem. CRC Press, Boca Raton (1995)
17. Guillemin, V.: A new proof of Weyl's formula on the asymptotic distribution of eigenvalues. Adv. Math. (N.Y). **55**(2), 131–160 (1985)
18. Hörmander, L.: The Analysis of Linear Partial Differential Operators III. Pseudo-Differential Operators. Classics in Mathematics. Springer, Berlin (2007)
19. Kalau, W., Walze, M.: Gravity, non-commutative geometry and the Wodzicki residue. J. Geom. Phys. **16**(4), 327–344 (1995)
20. Kastler, D.: The Dirac operator and gravitation. Commun. Math. Phys. **166**(3), 633–643 (1995)
21. Lewandowski, M.: Hadamard states for bosonic quantum field theory on globally hyperbolic spacetimes. arXiv:2008.13156 (2020)
22. Melrose, R.: Spectral and scattering theory for the Laplacian on asymptotically Euclidian spaces. In: Spectr. Scatt. Theory Proc. Taniguchi Int. Work. (1994)
23. Minakshisundaram, S., Pleijel, Å.: Some properties of the eigenfunctions of the Laplace-operator on Riemannian manifolds. Can. J. Math. **1**(3), 242–256 (1949)
24. Moretti, V.: Local ζ-function techniques vs. point-splitting procedure: a few rigorous results. Commun. Math. Phys. **201**(2), 327–363 (1999)
25. Nakamura, S., Taira, K.: Essential self-adjointness of real principal type operators. Ann. Henri Lebesgue **4**, 1035–1059 (2021)
26. Seeley, R.T.: Complex powers of an elliptic operator. Proc. Symp. Pure Math. **10**, 288–307 (1967)
27. Shubin, M.A.: Pseudodifferential Operators and Spectral Theory (2001)
28. Sogge, C.D.: Concerning the L^p norm of spectral clusters for second-order elliptic operators on compact manifolds. J. Funct. Anal. **77**(1), 123–138 (1988)
29. Sogge, C.D.: Hangzhou Lectures on Eigenfunctions of the Laplacian. Princeton University Press, Princeton (2014)
30. Strohmaier, A., Zelditch, S.: A Gutzwiller trace formula for stationary space-times. Adv. Math., 107434 (2020)
31. Taira, K.: Limiting absorption principle and equivalence of Feynman propagators on asymptotically Minkowski spacetimes. Commun. Math. Phys. **388**(1), 625–655 (2021)
32. Vasy, A.: A minicourse on microlocal analysis for wave propagation. In: Asymptot. Anal. Gen. Relativ., pp. 219–374. Cambridge University Press, Cambridge (2017)
33. Vasy, A.: Essential self-adjointness of the wave operator and the limiting absorption principle on Lorentzian scattering spaces. J. Spectr. Theory **10**(2), 439–461 (2020)

34. Wodzicki, M.: Local invariants of spectral asymmetry. Invent. Math. **75**(1), 143–177 (1984)
35. Zelditch, S.: Pluri-potential theory on Grauert tubes of real analytic Riemannian manifolds, I. In: Spectr. Geom., vol. 3, pp. 299–339. American Mathematical Society, Providence (2012)
36. Zelditch, S.: Eigenfunctions of the Laplacian of Riemannian manifolds (2017). www.math.northwestern.edu/~zelditch/Eigenfunction.pdf

Aspects of Non-associative Gauge Theory

Sergey Grigorian

Abstract A smooth loop is the direct non-associative generalization of Lie group. In this paper, we review the theory of smooth loops and smooth loop bundles. This is then used to define a non-associative analog of the Chern-Simons functional.

1 Introduction

One of highly successful areas at the intersection of differential geometry, analysis, and mathematical physics is gauge theory. As it is well-known, this is the study of connections on bundles with particular Lie groups as the structure groups. In [5], the author initiated a theory of smooth loops, which are non-associative analogs of Lie groups, and began the development of gauge theory based on loops, i.e. a non-associative gauge theory. The purpose of this note is to review the theory of smooth loops and loop bundles, and to provide a more rigorous construction of a non-associative Chern-Simons functional on 3-manifolds. In particular, the affine space of connections in a standard gauge theory is replaced by an affine space $\mathcal{T}$ of torsions, modelled on 1-forms with values in a loop algebra (the tangent space to a loop at identity). We define a 1-form on $\mathcal{T}$, show that it is a closed form, and show that it is the exterior derivative of a function on $\mathcal{T}$, which we define to be the Chern-Simons functional. Finally, we show how this functional is affected by gauge transformations.

S. Grigorian (✉)
University of Texas Rio Grande Valley, Edinburg, TX, USA
e-mail: sergey.grigorian@utrgv.edu

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_39

2 Smooth Loops

For a detailed discussion of concepts related to smooth loops, the reader is referred to [5]. The reader can also refer to [6, 7, 9, 11, 12] for a discussion of these concepts.

Definition 1 A *loop* $\mathbb{L}$ is a set with a binary operation $p \cdot q$ with identity 1, and compatible left and right quotients $p\backslash q$ and p/q, respectively.

In particular, existence of quotients is equivalent to saying that for any $q \in \mathbb{L}$, the left and right product maps L_q and R_q are invertible maps. Restricting to the smooth category, we obtain the definition of a smooth loop.

Definition 2 A *smooth loop* is a smooth manifold $\mathbb{L}$ with a loop structure such that the left and right product maps are diffeomorphisms of $\mathbb{L}$.

Definition 3 A *pseudoautomorphism* of a smooth loop $\mathbb{L}$ is a diffeomorphism $h : \mathbb{L} \longrightarrow \mathbb{L}$ for which there exists another diffeomorphism $h' : \mathbb{L} \longrightarrow \mathbb{L}$, known as the partial pseudoautomorphism corresponding to h, such that for any $p, q \in \mathbb{L}$,

$$h(pq) = h'(p)\, h(q). \tag{1}$$

In particular, $h' = R_{h(1)}^{-1} \circ h$. The element $h(1) \in \mathbb{L}$ is the *companion* of h'. As shown in [5], given h and h', we have the following properties

$$h(pq) = h'(p)\, h(q) \qquad h(q\backslash p) = h'(q)\backslash h(p) \qquad h'(p/q) = h(p)/h(q). \tag{2}$$

It is then easy to see that the sets of pseudoautomorhisms and partial pseudoautomorphisms are both groups. Denote the former by Ψ and the latter by Ψ'. We also see that the *automorphism* group of $\mathbb{L}$ is the subgroup $H \subset \Psi$ which is the stabilizer of $1 \in \mathbb{L}$. We will use $\mathbb{L}$ to denote $\mathbb{L}$ with the action of Ψ and $\mathbb{L}'$ to denote $\mathbb{L}$ with the action of Ψ', if a distinction between the G-sets is needed.

Let $r \in \mathbb{L}$, then we may define a modified product $\circ_r$ on $\mathbb{L}$ via $p \circ_r q = (p \cdot qr)/r$, so that $\mathbb{L}$ equipped with product $\circ_r$ will be denoted by $(\mathbb{L}, \circ_r)$, the corresponding quotient will be denoted by $/_r$. We have the following properties [5].

Lemma 1 *Let $h \in \Psi$. Then, for any $p, q, r \in \mathbb{L}$,*

$$h'(p \circ_r q) = h'(p) \circ_{h(r)} h'(q) \qquad h'(p/_r q) = h'(p)/_{h(r)} h'(q). \tag{3}$$

Consider the tangent space $\mathfrak{l} := T_1\mathbb{L}$ at $1 \in \mathbb{L}$. By analogy with Lie groups, for any $\xi \in \mathfrak{l}$, define the *fundamental vector field* $\rho(\xi)$ by pushing forward ξ by right translation, so that for any $q \in \mathbb{L}$, $\rho(\xi)_q = \left(R_q\right)_* \xi$.

Definition 4 ([5]) The Maurer-Cartan form θ is an $\mathfrak{l}$-valued 1-form on $\mathbb{L}$, such that $\theta(\rho(\xi)) = \xi$. Equivalently, for any vector field X, $\theta(X)|_p = \left(R_p^{-1}\right)_* X_p \in \mathfrak{l}$.

This allows us to define brackets on $\mathfrak{l}$. For each $p \in \mathbb{L}$ define the bracket $[\cdot,\cdot]^{(p)}$ given for any $\xi, \eta \in \mathfrak{l}$ by $[\xi,\eta]^{(p)} = -\theta\left([\rho(\xi), \rho(\eta)]\right)|_p$. We will denote $\mathfrak{l}$ equipped with the bracket $[\cdot,\cdot]^{(p)}$ by $\mathfrak{l}^{(p)}$. Define the *bracket function* $b : \mathbb{L} \longrightarrow \mathfrak{l} \otimes \Lambda^2 \mathfrak{l}^*$ to be the map that takes $p \mapsto [\cdot,\cdot]^{(p)} \in \mathfrak{l} \otimes \Lambda^2 \mathfrak{l}^*$, so that $b(\theta,\theta)$ is an $\mathfrak{l}$-valued 2-form on $\mathbb{L}$, i.e. $b(\theta,\theta) \in \Omega^2(\mathfrak{l})$.

Theorem 1 ([5, Theorem 3.10]) *The form θ satisfies $d\theta = \frac{1}{2} db(\theta,\theta)$.*

With respect to the action of Ψ, the bracket satisfies the following property.

Lemma 2 *If $h \in \Psi(\mathbb{L})$ and $q \in \mathbb{L}$, then, for any $\xi, \eta, \gamma \in \mathfrak{l}$, $h'_* [\xi,\eta]^{(q)} = \left[h'_*\xi, h'_*\eta\right]^{h(q)}$.*

We will assume that Ψ is a finite-dimensional Lie group, and suppose the Lie algebras of Ψ and $H_s = \mathrm{Aut}(\mathbb{L}, \circ_s)$ are $\mathfrak{p}$ and $\mathfrak{h}_s$, respectively. In particular, $\mathfrak{h}_s$ is a Lie subalgebra of $\mathfrak{p}$. Also, we will assume that Ψ acts transitively on $\mathbb{L}$. The action of Ψ on $\mathbb{L}$ induces an action of the Lie algebra $\mathfrak{p}$ on $\mathfrak{l}$, which we will denote by $\cdot$.

Definition 5 Define the map $\varphi : \mathbb{L} \longrightarrow \mathfrak{l} \otimes \mathfrak{p}^*$ such that for each $s \in \mathbb{L}$ and $\gamma \in \mathfrak{p}$,

$$\varphi_s(\gamma) = \frac{d}{dt}(\exp(t\gamma)(s))/s\Big|_{t=0} \in \mathfrak{l}. \tag{4}$$

Lemma 3 ([5]) *The map φ as in (4) is equivariant with respect to corresponding actions of $\Psi(\mathbb{L})$, in particular for $h \in \Psi$, $s \in \mathbb{L}$, $\gamma \in \mathfrak{p}$, we have*

$$\varphi_{h(s)}\left((\mathrm{Ad}_h)_* \gamma\right) = \left(h'\right)_* \varphi_s(\gamma). \tag{5}$$

Moreover, the image of φ_s is $\mathfrak{l}^{(s)}$ and the kernel is $\mathfrak{h}_s$, and hence, $\mathfrak{p} \cong \mathfrak{h}_s \oplus \mathfrak{l}^{(s)}$.

Lemma 4 ([5]) *Suppose $\xi \in \mathfrak{p}$ and $\eta, \gamma \in \mathfrak{l}$, then*

$$\xi \cdot [\eta,\gamma]^{(s)} = [\xi\cdot\eta,\gamma]^{(s)} + [\eta, \xi\cdot\gamma]^{(s)} + a_s(\eta,\gamma,\varphi_s(\xi)) \tag{6a}$$

$$\xi \cdot \varphi_s(\eta) = \eta\cdot\varphi_s(\xi) + \varphi_s\left([\xi,\eta]_{\mathfrak{p}}\right) + [\varphi_s(\xi), \varphi_s(\eta)]^{(s)}. \tag{6b}$$

Similarly as for Lie groups, we may define a Killing form $K^{(s)}$ on $\mathfrak{l}^{(s)}$. For $\xi, \eta \in \mathfrak{l}$, we have

$$K^{(s)}(\xi,\eta) = \mathrm{Tr}\left(\mathrm{ad}^{(s)}_\xi \circ \mathrm{ad}^{(s)}_\eta\right), \tag{7}$$

where $\circ$ is just composition of linear maps on $\mathfrak{l}$ and $\mathrm{ad}^{(s)}_\xi(\cdot) = [\xi,\cdot]^{(s)}$. Clearly $K^{(s)}$ is a symmetric bilinear form on $\mathfrak{l}$. In [5] it is shown that for $h \in \Psi$, and $\xi, \eta \in \mathfrak{l}$ it satisfies $K^{(h(s))}\left(h'_*\xi, h'_*\eta\right) = K^{(s)}(\xi,\eta)$.

Suppose now $K^{(s)}$ is nondegenerate and $\mathfrak{p}$-invariant, so that the action of $\mathfrak{p}$ is skew-adjoint with respect to $K^{(s)}$. Moreover suppose $\mathfrak{p}$ is semisimple itself, so that it has a nondegenerate, invariant Killing form $K_{\mathfrak{p}}$. We will use $\langle\cdot,\cdot\rangle^{(s)}$ and $\langle\cdot,\cdot\rangle_{\mathfrak{p}}$

to denote the inner products using $K^{(s)}$ and $K_{\mathfrak{p}}$, respectively. Then, given the map $\varphi_s : \mathfrak{p} \longrightarrow \mathfrak{l}^{(s)}$, we can define its adjoint with respect to these two bilinear maps.

Definition 6 Define the map $\varphi_s^t : \mathfrak{l}^{(s)} \longrightarrow \mathfrak{p}$ such that for any $\xi \in \mathfrak{l}^{(s)}$ and $\eta \in \mathfrak{p}$,

$$\left\langle \varphi_s^t(\xi), \eta \right\rangle_{\mathfrak{p}} = \langle \xi, \varphi_s(\eta) \rangle^{(s)}. \tag{8}$$

Since $\mathfrak{h}_s \cong \ker \varphi_s$, we have $\mathfrak{p} \cong \mathfrak{h}_s \oplus \operatorname{Im} \varphi_s^t$, so that $\mathfrak{h}_s^{\perp} = \operatorname{Im} \varphi_s^t$.

Lemma 5 ([5, Lemma 3.43]) *Suppose Ψ acts transitively on $\mathbb{L}$, $\mathfrak{l}$ is an irreducible representation of $\mathfrak{h}$, and suppose the base field of $\mathfrak{p}$ is $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. Then, there exists a $\lambda \in \mathbb{F}$ such that for any $s \in \mathbb{L}$, $\varphi_s \varphi_s^t = \lambda \operatorname{id}_{\mathfrak{l}}$ and $\varphi_s^t \varphi_s = \lambda \pi_{\mathfrak{h}_s^{\perp}}$.*

Thus, given our prior assumption of the transitivity of the action of Ψ, the maps φ_s and φ_s^t are isomorphisms between $\mathfrak{l}$ and $\mathfrak{h}_s^{\perp}$. If $s \in \mathbb{L}$ is fixed, and there is no ambiguity, we will use the following notation. Given $\xi \in \mathfrak{p}$, $\hat{\xi} = \varphi_s(\xi) \in \mathfrak{l}$ and given $\eta \in \mathfrak{l}$, $\check{\eta} = \frac{1}{\lambda_s} \varphi_s^t(\eta) \in \mathfrak{h}_s^{\perp}$. We can also use φ_s^t to define a new bracket $[\cdot, \cdot]_{\varphi_s}$ on $\mathfrak{l}$, such that for $\xi, \eta \in \mathfrak{l}$,

$$[\xi, \eta]_{\varphi_s} = \varphi_s \left(\left[\check{\xi}, \check{\eta} \right]_{\mathfrak{p}} \right). \tag{9}$$

Lemma 6 ([5, Lemma 3.50]) *Let $s \in \mathbb{L}$, then under the assumptions of Lemma 5, the bracket $[\cdot, \cdot]_{\varphi_s}$ satisfies the following properties. Suppose $\xi, \eta, \gamma \in \mathfrak{l}$, then*

1. $\left\langle [\xi, \eta]_{\varphi_s}, \gamma \right\rangle^{(s)} = -\left\langle \eta, [\xi, \gamma]_{\varphi_s} \right\rangle^{(s)}$.
2. *For any $h \in \Psi$, $[\xi, \eta]_{\varphi_{h(s)}} = \left(h'\right)_* \left[\left(h'\right)_*^{-1} \xi, \left(h'\right)_*^{-1} \eta \right]_{\varphi_s}$.*

3 Loop Bundles

Let M be a smooth, finite-dimensional manifold with a Ψ-principal bundle $\pi : \mathcal{P} \longrightarrow M$.

Definition 7 Let $s : \mathcal{P} \longrightarrow \mathbb{L}$ be an equivariant map. In particular, the equivalence class $\lfloor p, s_p \rfloor_{\Psi}$ defines a section of the bundle $\mathcal{Q} = \mathcal{P} \times_{\Psi} \mathbb{L}$. We will refer to s as *the defining map* (or *section*).

We will define several associated bundles related to $\mathcal{P}$. As it is well-known, sections of associated bundles are equivalent to equivariant maps. With this in mind, we also give properties of equivariant maps that correspond to sections of these bundles. Let $h \in \Psi$ and, as before, denote by h' the partial action of h.

Bundle	Equivariant map	Equivariance property
$\mathcal{P}$	$k : \mathcal{P} \longrightarrow \Psi$	$k_{ph} = h^{-1} k_p$
$\mathcal{Q}' = \mathcal{P} \times_{\Psi'} \mathbb{L}'$	$q : \mathcal{P} \longrightarrow \mathbb{L}'$	$q_{ph} = \left(h'\right)^{-1} q_p$
$\mathcal{Q} = \mathcal{P} \times_{\Psi} \mathbb{L}$	$r : \mathcal{P} \longrightarrow \mathbb{L}$	$r_{ph} = h^{-1} \left(r_p\right)$
$\mathcal{A} = \mathcal{P} \times_{\Psi'_*} \mathfrak{l}$	$\eta : \mathcal{P} \longrightarrow \mathfrak{l}$	$\eta_{ph} = \left(h'\right)^{-1}_* \eta_p$
$\mathfrak{p}_{\mathcal{P}} = \mathcal{P} \times_{(\mathrm{Ad}_\xi)_*} \mathfrak{p}$	$\xi : \mathcal{P} \longrightarrow \mathfrak{p}$	$\xi_{ph} = \left(\mathrm{Ad}_h^{-1}\right)_* \xi_p$
$\mathrm{Ad}\left(\mathcal{P}\right) = \mathcal{P} \times_{\mathrm{Ad}_\Psi} \Psi$	$u : \mathcal{P} \longrightarrow \Psi$	$u_{ph} = h^{-1} u_p h$

(10)

Given equivariant maps $q, r : \mathcal{P} \longrightarrow \mathbb{L}'$, define an equivariant product using s, given for any $p \in \mathcal{P}$ by

$$q \circ_s r|_p = q_p \circ_{s_p} r_p. \tag{11}$$

Due to Lemma 1, the corresponding map $q \circ_s r : \mathcal{P} \longrightarrow \mathbb{L}'$ is equivariant, and hence $\circ_s$ induces a fiberwise product on sections of $\mathcal{Q}$. Analogously, we define fiberwise quotients of sections of $\mathcal{Q}$. Similarly, we define an equivariant bracket $[\cdot, \cdot]^{(s)}$ and the equivariant map φ_s. Other related objects such as the Killing form $K^{(s)}$ and the adjoint φ_s^t to φ_s are then similarly also equivariant.

Suppose the principal Ψ-bundle $\mathcal{P}$ has a principal Ehresmann connection given by the decomposition $T\mathcal{P} = \mathcal{HP} \oplus \mathcal{VP}$ and the corresponding vertical $\mathfrak{p}$-valued connection 1-form ω. Given an equivariant map $f : \mathcal{P} \longrightarrow S$, define

$$d^\omega f := f_* \circ \underset{\mathcal{H}}{\mathrm{proj}} : T\mathcal{P} \longrightarrow \mathcal{HP} \longrightarrow TS. \tag{12}$$

This is then a horizontal map since it vanishes on any vertical vectors. The map $d^\omega f$ is moreover still equivariant, and hence induces a covariant derivative on sections of the associated bundle $\mathcal{P} \times_\Psi S$. If S is a vector space, then this reduces to the usual definition of the exterior covariant derivative of a vector bundle-valued function and $d^\omega f$ is a vector-bundle-valued 1-form. Note that due to our initial assumption that $K^{(s)}$ is $\mathfrak{p}$-invariant, we also see that d^ω is metric-compatible with respect to $\langle \cdot, \cdot \rangle^{(s)}$.

Following [5], let us define the torsion of the defining map s with respect to the connection ω.

Definition 8 The *torsion* $T^{(s,\omega)}$ of the defining map s with respect to ω is a horizontal $\mathfrak{l}$-valued 1-form on $\mathcal{P}$ given by $T^{(s,\omega)} = (s^*\theta) \circ \mathrm{proj}_{\mathcal{H}}$, where θ is Maurer-Cartan form of $\mathbb{L}$. Equivalently, at $p \in \mathcal{P}$, we have

$$T^{(s,\omega)}\Big|_p = \left(R_{s_p}^{-1}\right)_* d^\omega s\big|_p. \tag{13}$$

Thus, $T^{(s,\omega)}$ is the horizontal component of $\theta_s = s^*\theta$. We also easily see that it is Ψ-equivariant. Thus, $T^{(s,\omega)}$ is a *basic* (i.e. horizontal and equivariant) $\mathfrak{l}$-valued 1-form on $\mathcal{P}$, and thus defines a 1-form on M with values in the associated vector bundle $\mathcal{A} = \mathcal{P} \times_{\Psi'_*} \mathfrak{l}$. We have the following properties.

Theorem 2 *Suppose $s : \mathcal{P} \longrightarrow \mathbb{L}$, then*

$$d^{\omega}\varphi_s = \mathrm{id}_{\mathfrak{p}} \cdot T^{(s,\omega)} - \left[\varphi_s, T^{(s,\omega)}\right]^{(s)} \tag{14a}$$

$$d^{\omega}\varphi_s^t = \varphi_s^t\left(\check{T} \cdot \mathrm{id}_{\mathfrak{l}}\right) - \left[\check{T}, \varphi_s^t\right]_{\mathfrak{p}}, \tag{14b}$$

where $\mathrm{id}_{\mathfrak{p}}$ and $\mathrm{id}_{\mathfrak{l}}$ are the identity maps of $\mathfrak{p}$ and $\mathfrak{l}$, respectively, and $\cdot$ denotes the action of the Lie algebra $\mathfrak{p}$ on $\mathfrak{l}$.

Proof Equation (14a) follows from [5, Theorem 4.11]. However to obtain (14b), suppose ξ is an $\mathfrak{l}$-valued map and η is a $\mathfrak{p}$-valued map. Then,

$$\left\langle \left(d^{\omega}\varphi_s^t\right)(\xi), \eta\right\rangle^{(s)} = \left\langle \xi, \left(d^{\omega}\varphi_s\right)\eta\right\rangle^{(s)}.$$

Using (14a) and (6b), we obtain (14b).

Recall that the curvature $F^{(\omega)} \in \Omega^2(\mathcal{P}, \mathfrak{p})$ of the connection ω on $\mathcal{P}$ is given by

$$F^{(\omega)} = d\omega \circ \mathrm{proj}_{\mathcal{H}} = d\omega + \frac{1}{2}[\omega, \omega]_{\mathfrak{p}}, \tag{15}$$

where wedge product is implied. Given the defining map s, define $\hat{F}^{(s,\omega)} \in \Omega^2(\mathcal{P}, \mathfrak{l})$ to be the projection of the curvature $F^{(\omega)}$ to $\mathfrak{l}$ with respect to s, such that for any $X_p, Y_p \in T_p\mathcal{P}$,

$$\hat{F}^{(s,\omega)} = \varphi_s\left(F^{(\omega)}\right). \tag{16}$$

Theorem 3 ([5, Theorem 4.19]) *$\hat{F}^{(s,\omega)}$ and $T^{(s,\omega)}$ satisfy the following structure equation*

$$\hat{F}^{(s,\omega)} = d^{\omega}T^{(s,\omega)} - \frac{1}{2}\left[T^{(s,\omega)}, T^{(s,\omega)}\right]^{(s)}, \tag{17}$$

where a wedge product between the 1-forms $T^{(s,\omega)}$ is implied.

In the case of an octonion bundle over a 7-dimensional manifold, this relationship between the torsion and a curvature component has been shown in [2]. Using φ_s and φ_s^t, let us define an adapted covariant derivative as a map from the space of $\mathfrak{l}$-valued equivariant functions $\Omega^0_{basic}(\mathcal{P}, \mathfrak{l})$ to the space of horizontal $\mathfrak{l}$-valued equivariant (i.e. basic) 1-forms $\Omega^1_{basic}(\mathcal{P}, \mathfrak{l})$:

$$d^{\omega}_{\varphi_s} = \frac{1}{\lambda}\varphi_s \circ d^{\omega} \circ \varphi_s^t : \Omega^0_{basic}(\mathcal{P}, \mathfrak{l}) \longrightarrow \Omega^1_{basic}(\mathcal{P}, \mathfrak{l}). \tag{18}$$

We can see that this covariant derivative is metric-compatible as long as d^{ω} is.

Lemma 7 *Suppose* $\xi, \eta \in \Omega^0_{basic}(\mathcal{P}, \mathfrak{l})$, *then* d^ω *is metric-compatible if and only if*

$$d\langle \xi, \eta \rangle_s = \left\langle d^\omega_{\varphi_s} \xi, \eta \right\rangle_s + \left\langle \xi, d^\omega_{\varphi_s} \eta \right\rangle_s .$$

Proof This can be shown by explicitly expanding $d^\omega_{\varphi_s}$ and noting that since $\varphi_s \varphi_s^t = \lambda \, \mathrm{id}_{\mathfrak{l}}$, $(d^\omega \varphi_s)\, \varphi_s^t = -\varphi_s \left(d\varphi_s^t\right)$.

Theorem 4 $\hat{F}^{(s,\omega)}$ *satisfies the following Bianchi identity*

$$d^\omega_{\varphi_s} \hat{F}^{(s,\omega)} = F_{\mathfrak{h}_s} \dot{\wedge} T^{(s,\omega)}, \tag{19}$$

where $F_{\mathfrak{h}_s} = \pi_{\mathfrak{h}_s} F$ *and* $\dot{\wedge}$ *denotes the action of* $\mathfrak{p}$ *on* $\mathfrak{l}$ *combined with the wedge product of p-forms.*

Proof We can write $F = F_{\mathfrak{h}_s} + \frac{1}{\lambda} \varphi_s^t \left(\hat{F}\right)$, so applying $\varphi_s \circ d^\omega$, the left-hand side vanishes due to the standard Bianchi identity, and we are left with

$$d^\omega_{\varphi_s} \hat{F} = -\varphi_s \left(d^\omega F_{\mathfrak{h}_s}\right) = \left(d^\omega \varphi_s\right) \wedge F_{\mathfrak{h}_s} .$$

Using (14a), we obtain (19).

3.1 Gauge Theory

As discussed earlier, equivariant horizontal forms on $\mathcal{P}$ give rise to sections of corresponding associated bundles over the base manifold M. So let us now switch perspective, and work in terms of sections of bundles. Recall that the space of connections on $\mathcal{P}$ is an affine space modelled on $\Omega^1(\mathfrak{p}_\mathcal{P})$. Thus, any connection $\tilde{\omega} = \omega + A$ for some $A \in \Omega^1(\mathfrak{p}_\mathcal{P})$. Then,

$$T^{(s,\tilde{\omega})} = T^{(s,\omega)} + \varphi_s(A) \tag{20}$$

The space of possible torsions of s therefore comes from deformations by elements of $\varphi_s^t\left(\Omega^1(\mathcal{A})\right)$. So define the *torsion space* $\mathcal{T}_s \cong \Omega^1(\mathcal{A})$. Therefore, for any $\xi \in \Omega^1(\mathcal{A})$, the torsion and curvature of $\omega_\xi = \omega + \check{\xi}$ are given by

$$T^{(s,\omega_\xi)} = T^{(s,\omega)} + \xi \tag{21a}$$

$$\hat{F}^{(s,\omega_\xi)} = \hat{F}^{(s,\omega)} + d^\omega_{\varphi_s} \xi + \frac{1}{2} [\xi, \xi]_{\varphi_s} . \tag{21b}$$

Since our prior assumption of transitivity of the action of Ψ implies that φ_s is surjective, we can find a reference connection ω_0 for which $T^{(s,\omega_0)} = 0$. In

particular, ω_0 will have curvature with values in $\mathfrak{h}_s$, and in particular $\hat{F}^{(s,\omega_0)} = 0$. The torsion will unchanged if we add to ω an $\mathfrak{h}_s$-valued 1-form, hence the equivalence $\mathcal{T}_s \cong \Omega^1(\mathcal{A})$ is independent of the choice of a particular ω_0.

Suppose h is a section of the associated bundle $\mathrm{Ad}(\mathcal{P})$, then it defines a gauge-transformation and the gauge transformed connection is $h^*\omega$. In particular, for the section $s \in \Gamma(\mathcal{Q})$, we have

$$d^{h^*\omega}s = (h_*)^{-1} d^{\omega}(h(s)). \tag{22}$$

Since the torsion is determined by the covariant derivative of s, transformations of the connection and the defining section s are very closely related. Indeed, as shown in [5], the corresponding transformation of torsion is given by

$$\begin{aligned} T^{(h(s),\omega)} &= \left(R_{h(s)}\right)_*^{-1} d^{\omega}(h(s)) \\ &= h'_* \circ (R_s)_*^{-1} \circ (h_*)^{-1} d^{\omega}(h(s)) \\ &= h'_* T^{(s,h^*\omega)}, \end{aligned} \tag{23}$$

which follows from Definition 8 and properties of h (2). Recall that we assumed that Ψ acts transitively on $\mathbb{L}$, so that, for a fixed connection ω, all the possible torsions are obtained by the action of Ψ on s, with the non-trivial transformations given by cosets of Ψ/H_s, where $H_s = \mathrm{Stab}(s)$. On the other hand, as (23) shows, transformations of s correspond to gauge transformations of the connection. Since,

$$d^{h^*\omega}s = d^{\omega}s + (h)_*^{-1}\left(d^{\omega}h\right)\cdot s, \tag{24}$$

we obtain

$$T^{(s,h^*\omega)} = T^{(s,\omega)} + \varphi_s\left((h_*)^{-1}\left(d^{\omega}h\right)\right). \tag{25}$$

We will define *loop gauge transformations* to be precisely those that act non-trivially on s. Infinitesimally this corresponds to taking $h = \exp(\check{\eta})$ for $\eta \in \Omega^0(\mathcal{A})$, so that

$$T^{(s,u^*\omega)} = T^{(s,\omega)} + d^{\omega}_{\varphi_s}\eta, \tag{26}$$

hence at $T^{(s,\omega)} \in \mathcal{T}_s$, the tangent vectors to $\mathcal{T}_s$ in the directions of loop gauge transformations correspond precisely to the image of $d^{\omega}_{\varphi_s}$. Although this is beyond the scope of this note, the L_2-norm of T may be considered as a functional on gauge orbits in $\mathcal{T}_s$. Critical points then become analogues of the Coulomb gauge condition in gauge theory [1–5, 8].

The above considerations allow us to consider analogues of various functionals defined in gauge theory [5]. The key difference of course is that $\hat{F}$ does not satisfy the standard Bianchi identity.

Let us now specialize to the case of M being a smooth compact 3-dimensional manifold. Following the standard theory, as in [10], let us define a 1-form ρ on $\mathcal{T}_s$, for $\chi \in \Omega^1(\mathcal{A})$, which is also interpreted as an element of $T_\omega \mathcal{T}$, by

$$\rho(\chi)|_\omega = \int_M \left\langle \hat{F}^{(s,\omega)}, \chi \right\rangle^{(s)}. \tag{27}$$

Theorem 5 *Suppose M is a smooth compact 3-dimensional manifold, then $\rho = d\vartheta$, where ϑ is a functional on $\mathcal{T}_s \cong \Omega^1(\mathcal{A})$ given by*

$$\vartheta(\xi) = \frac{1}{2} \int_M \left\langle d^\omega_{\varphi_s} \xi + \frac{1}{3} [\xi, \xi]_{\varphi_s}, \xi \right\rangle^{(s)} dt. \tag{28}$$

The critical points of ϑ correspond to $\omega_\xi = \omega_0 + \check{\xi}$ for which $\hat{F}^{(s,\omega_\xi)} = 0$.

Proof Consider $\omega_\xi = \omega + \check{\xi}$, then using Stokes' Theorem, to first order we get

$$\begin{aligned}
\rho(\chi)|_{\omega_\xi} - \rho(\chi)|_\omega &= \int_M \left\langle d^\omega_{\varphi_s} \xi, \chi \right\rangle^{(s)} + O\left(|\xi|^2\right) \\
&= \int_M d\left\langle \xi, \chi \right\rangle^{(s)} + \int_M \left\langle \xi, d^\omega_{\varphi_s} \chi \right\rangle^{(s)} + O\left(|\xi|^2\right) \\
&= \int_M \left\langle \xi, d^\omega_{\varphi_s} \chi \right\rangle^{(s)} + O\left(|\xi|^2\right)
\end{aligned}$$

Using the same argument as in [10], we see that $d\rho = 0$. Since $\mathcal{T}_s$ is a contractible space, by Poincare lemma, $\rho = d\vartheta$ for some function ϑ on $\mathcal{T}_s$. Consider now a path $\omega(t) = \omega_0 + t\check{\xi}$ from ω_0 to $\omega = \omega_0 + \check{\xi}$, where ω_0 is such that $T^{(s,\omega_0)} = 0$. Integrating it explicitly, and noting that since ρ is closed, this is path-independent, we get,

$$\begin{aligned}
\vartheta(\xi) - \vartheta(0) &= \int_0^1 \rho_{\omega(t)} \left(\varphi_s\left(\dot{\omega}(t)\right)\right) dt \\
&= \int_0^1 \int_M \left\langle \hat{F}^{(s,\omega(t))}, \xi \right\rangle^{(s)} dt \\
&= \int_0^1 \int_M \left\langle t d^\omega_{\varphi_s} \xi + \frac{1}{2} t^2 [\xi, \xi]_{\varphi_s}, \xi \right\rangle^{(s)} dt \\
&= \frac{1}{2} \int_M \left\langle d^\omega_{\varphi_s} \xi + \frac{1}{3} [\xi, \xi]_{\varphi_s}, \xi \right\rangle^{(s)} dt.
\end{aligned}$$

Setting $\vartheta(0) = 0$, and noting that $\xi = T^{(s,\xi)}$ and $d^{\omega}_{\varphi_s}\xi = \hat{F}^{(s,\xi)} - [\xi,\xi]_{\varphi_s}$, we recover

$$\vartheta(\xi) = \frac{1}{2}\int_M \left(\left\langle T, \hat{F}\right\rangle^{(s)} - \frac{1}{6\lambda^2}\left\langle T, [T,T]_{\varphi_s}\right\rangle^{(s)}\right), \tag{29}$$

which (up to a factor of $\frac{1}{2}$), is the Loop Chern-Simons Functional defined in [5]. In particular, we see that $d\vartheta|_{\omega} = 0$ if and only if $\hat{F}^{(s,\omega)} = 0$, that is, connections for which this holds are critical points of the functional ϑ.

Unlike in the case of the standard Chern-Simons Functional, ρ does not necessarily vanish along orbits of the non-associative gauge action. As we see from (26), vectors tangent to the orbits are given by $d^{\omega}_{\varphi_s}\eta$ for some $\eta \in \Omega^0(\mathcal{A})$. Using (19), we find

$$\rho\left(d^{\omega}_{\varphi_s}\eta\right)\big|_{\omega} = \int_M \left\langle F^{\omega}_{\mathfrak{h}_s} \dot{\wedge} T^{(s,\omega)}, \eta\right\rangle^{(s)}. \tag{30}$$

Now let us consider how ϑ is affected by gauge transformations. Consider a path $t \in [0,1]$ connecting $T^{(s,\omega)}$ to $T^{(s,u^*\omega)}$. In particular, this is equivalent to a path $\xi(t) \in \Omega^1(\mathcal{A})$ such that $\xi(0) = 0$ and $\xi(1) = \varphi_s(u^*\omega - \omega)$. Then, define $\omega(t) = \omega + \check{\xi}(t)$, so that

$$\begin{aligned}\vartheta(\xi(1)) - \vartheta(0) &= \int_0^1 \rho_{\omega(t)}\left(\varphi_s\left(\dot{\omega}(t)\right)\right)dt \\ &= \int_0^1 \int_M \left\langle \hat{F}^{(s,\omega(t))}, \dot{\xi}(t)\right\rangle^{(s)} dt.\end{aligned} \tag{31}$$

As in the standard gauge theory [10], we may extend $\mathcal{P}$, and all the associated bundles, to a bundle over $\tilde{M} = M \times [0,1]$. In a local trivialization, let us define the connection $A = A_0 dt + A_i dx^i$ on $\tilde{M}$ with $A_0 = 0$ and $(A_i)_{(p,t)} = \omega_i(t)_p$. Then, we see that the curvature F_A of this connection is given by $(F_A)_{0i} = \dot{A}_i(t)$ and $(F_A)_{ij} = \left(F^{(\omega)}\right)_{ij}$. Hence

$$\hat{F}_A = \dot{\xi}_i(t)\, dt \wedge dx^i + \left(\hat{F}^{(s,\omega)}\right)_{ij} dx^i \wedge dx^j = -\dot{\xi}(t) \wedge dt + \hat{F}^{(s,\omega)}, \tag{32}$$

so that $\left\langle \hat{F}_A, \hat{F}_A\right\rangle = -2\left\langle \hat{F}^{(s,\omega(t))}, \dot{\xi}(t)\right\rangle^{(s)} \wedge dt$, and thus (31) becomes

$$\vartheta(\xi(1)) - \vartheta(0) = -2\int_{\tilde{M}} \left\langle \hat{F}_A, \hat{F}_A\right\rangle. \tag{33}$$

This shows that there is a relation between Chern-Simons and a Chern-Weil-like functionals, similar to standard gauge theory. However, the 4-form $\left\langle \hat{F}_A, \hat{F}_A \right\rangle$ on a 4-manifold is not necessarily independent of the choice of connection, so it is not a topological invariant. On the other hand, the above discussion shows that in this particular case, it is independent of the path $\omega(t)$, so it is important to understand if there is an invariant theory that is related to this non-associative gauge theory.

Acknowledgments This work was supported by the National Science Foundation grant DMS-1811754.

References

1. Donaldson, S.K.: Gauge theory: mathematical applications. In: Encyclopedia of Mathematical Physics, pp. 468–481. Academic/Elsevier Science, Oxford (2006)
2. Grigorian, S.: G_2-structures and octonion bundles. Adv. Math. **308**, 142–207 (2017). http://arxiv.org/abs/1510.04226. https://doi.org/10.1016/j.aim.2016.12.003
3. Grigorian, S.: Estimates and monotonicity for a heat flow of isometric G_2-structures. Calc. Var. Part. Differ. Equ. **58**(5), Art. 175, 37, (2019). https://doi.org/10.1007/s00526-019-1630-0
4. Grigorian, S.: Isometric flows of G_2-structures (2020). arXiv:2008.06593
5. Grigorian, S.: Smooth loops and loop bundles. Adv. Math. **393**, Paper No. 108078, 115 (2021). http://arxiv.org/abs/2008.08120. https://doi.org/10.1016/j.aim.2021.108078
6. Hofmann, K.H., Strambach, K.: Topological and analytic loops. In: Quasigroups and Loops: Theory and Applications, vol. 8. Sigma Ser. Pure Math., pp. 205–262. Heldermann, Berlin (1990)
7. Kiechle, H.: Theory of *K-loops*, vol. 1778. Lecture Notes in Mathematics. Springer, Berlin (2002). https://doi.org/10.1007/b83276
8. Loubeau, E., Sá Earp, H.N.: Harmonic flow of geometric structures (2019). arXiv:1907.06072
9. Nagy, P.T., Strambach, K.: Loops in Group Theory and Lie Theory, vol. 35. De Gruyter Expositions in Mathematics. Walter de Gruyter & Co., Berlin (2002). https://doi.org/10.1515/9783110900583
10. Sá Earp, H.N.: Instantons on G_2-manifolds (with Addendum). PhD Thesis, Imperial College London (2009)
11. Sabinin, L.V.: Smooth Quasigroups and Loops, vol. 492. Mathematics and its Applications. Kluwer Academic Publishers, Dordrecht (1999). https://doi.org/10.1007/978-94-011-4491-9
12. Smith, J.D.H.: An Introduction to Quasigroups and Their Representations. Studies in Advanced Mathematics. Chapman & Hall/CRC, Boca Raton (2007)

Remarks on Global Smoothing Effect of Solutions to Nonlinear Elastic Wave Equations with Viscoelastic Term

Yoshiyuki Kagei and Hiroshi Takeda

Abstract The aim of this paper is to give the precise statement and proof of smoothing effect and asymptotic profiles of the small global solutions to quasilinear elastic wave equations with viscoelastic terms, which were already announced in Kagei and Takeda (Nonlinear Anal 219, Paper No. 112826, 36 pp., 2022). Here the nonlinear terms in this paper include time derivative. The proof is based on the estimates for the fundamental solutions. The difference between spatial derivative and time derivative in nonlinear terms is remarked.

1 Introduction

This paper studies the Cauchy problem of the following nonlinear elastic wave equations with viscoelastic term:

$$\begin{cases} \partial_t^2 u - \mu\Delta u - (\lambda+\mu)\nabla \mathrm{div} u - \nu\Delta\partial_t u = F(u), \quad t > 0, \quad x \in \mathbb{R}^3, \\ u(0,x) = f_0(x), \quad \partial_t u(0,x) = f_1(x), \quad x \in \mathbb{R}^3, \end{cases} \tag{1}$$

where $u = {}^t(u_1, u_2, u_3)$ is the unknown function; and $f_j = {}^t(f_{j1}, f_{j2}, f_{j3})$ $(j = 0, 1)$ are initial data. Here the superscript ${}^t\cdot$ stands for the transpose of the matrix. Throughout the paper we assume that the Lamé constants satisfy

$$\mu > 0, \quad \lambda + 2\mu > 0,$$

Y. Kagei
Department of Mathematics, Tokyo Institute of Technology, Meguro-ku, Tokyo, Japan
e-mail: kagei@math.titech.ac.jp

H. Takeda (✉)
Department of Intelligent Mechanical Engineering, Faculty of Engineering, Fukuoka Institute of Technology, Higashi-ku, Fukuoka, Japan
e-mail: h-takeda@fit.ac.jp

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_40

and the viscosity parameter ν is positive.

The aim of this paper is to give the precise statement and proof of the large time behavior of the global solutions to (1), which were announced in [3] without proof. For this reason, our nonlinear term is restricted to the form

$$F(u) = \nabla u \nabla \partial_t u,$$

where ∇ is the spatial gradient. Concerning the physical background of the system (1) and related results on mathematical analysis, see [3] and references therein.

We begin with the existence of global solutions to (1) for small initial data.

Theorem 1 *Suppose that $F(u) = \nabla u \nabla \partial_t u$. Let $(f_0, f_1) \in Y := \{\dot{H}^3 \cap \dot{W}^{1,1}\}^3 \times \{H^2 \cap L^1\}^3$. If $\epsilon := \|f_0, f_1\|_Y$ is sufficiently small, then there exists a unique global solution to (1) in the class*

$$\{C([0,\infty); \dot{H}^3 \cap \dot{H}^1) \cap C^1([0,\infty); H^2)\}^3$$

satisfying the estimates

$$\begin{aligned} \|\nabla^\alpha u(t)\|_2 &\le C\epsilon(1+t)^{-\frac{1}{4}-\frac{\alpha}{2}}, \quad 1 \le \alpha \le 3, \\ \|\partial_t \nabla^\alpha u(t)\|_2 &\le C\epsilon(1+t)^{-\frac{3}{4}-\frac{\alpha}{2}}, \quad 0 \le \alpha \le 2 \end{aligned} \tag{2}$$

for $t \ge 0$, where the norm of $f \in L^p(\mathbb{R}^3)$ is defined by $\|f\|_p$ for $1 \le p \le \infty$ and for $k \ge 0$ and $1 \le p \le \infty$, $W^{k,p}(\mathbb{R}^3)$ is defined as the usual Sobolev spaces

$$W^{k,p}(\mathbb{R}^3) := \left\{ f : \mathbb{R}^3 \to \mathbb{R}; \|f\|_{W^{k,p}(\mathbb{R}^3)} := \|f\|_p + \|\nabla_x^k f\|_p < \infty \right\}$$

with the notation $W^{k,2}(\mathbb{R}^3) = H^k(\mathbb{R}^3)$.

Theorem 2 *The global solution $u(t)$ constructed in Theorem 1 satisfies*

$$u \in \{C^1((0,\infty); W^{2,6}) \cap \dot{W}^{1,\infty}(0,\infty; W^{1,\infty}) \cap C^2([0,\infty); L^2) \cap C^2((0,\infty); L^6)\}^3$$

and the estimates,

$$\|\nabla^\alpha u(t)\|_\infty \le C\epsilon(1+t)^{-\frac{3}{2}-\frac{\alpha}{2}}, \quad 0 \le \alpha \le 1 \tag{3}$$

$$\|\partial_t u(t)\|_\infty \le C\epsilon(1+t)^{-2}, \tag{4}$$

for $t \ge 0$ and

$$\|\nabla^2 \partial_t u(t)\|_p \le C\epsilon(1+t)^{-\frac{9}{4}+\frac{1}{p}} t^{-\frac{3}{4}+\frac{3}{2p}}, \quad 2 \le p \le 6, \tag{5}$$

$$\|\nabla \partial_t u(t)\|_\infty \le C\epsilon(1+t)^{-\frac{9}{4}} t^{-\frac{1}{4}}, \tag{6}$$

$$\|\partial_t^2 u(t)\|_p \le C\epsilon(1+t)^{-\frac{7}{4}+\frac{1}{p}} t^{-\frac{3}{2}(\frac{1}{2}-\frac{1}{p})}, \quad 2 \le p \le 6 \tag{7}$$

for $t \geq 0$ *with* $p = 2$ *and* $t > 0$ *with* $p \neq 2$.

Finally we mention the asymptotic profiles of the global solution u as $t \to \infty$. For this purpose, following [3], we introduce some notations. The diffusion waves $G_j^{(\beta)}(t)$ for $j = 0, 1$ depending on the parameter $\beta > 0$ are denoted by

$$G_j^{(\beta)}(t, x) := \mathcal{F}^{-1}[\mathcal{G}_j^{(\beta)}(t, \xi)],$$

where

$$\mathcal{G}_0^{(\beta)}(t, \xi) := e^{-\frac{\nu|\xi|^2}{2}t} \cos(\beta|\xi|t), \quad \mathcal{G}_1^{(\beta)}(t, \xi) := e^{-\frac{\nu|\xi|^2}{2}t} \frac{\sin(\beta|\xi|t)}{\beta|\xi|}.$$

We also denote the identity matrix by $\mathcal{I}_3 \in M(\mathbb{R}; 3)$ and define

$$\mathcal{P} := \frac{\xi}{|\xi|} \otimes \frac{\xi}{|\xi|}.$$

The 3-d valued constant vectors depending on the initial data and the nonlinear term are defined by

$$m_j = {}^t(m_{j1}, m_{j2}, m_{j3}), \quad M[u] := {}^t(M_1[u], M_2[u], M_3[u]),$$

where

$$m_{0k} := \int_{\mathbb{R}^3} \nabla f_{0k}(x)dx, \quad m_{1k} := \int_{\mathbb{R}^3} f_{1k}(x)dx$$

and

$$M_k[u] := \int_0^\infty \int_{\mathbb{R}^3} F_k(u)(\tau, y)dy\, d\tau$$

for $k = 1, 2, 3$. Using the above notation, we denote the functions G, H and $\tilde{G}$ by

$$\begin{aligned} G(t, x) :=& \nabla^{-1}\mathcal{F}^{-1}\left[\left(\mathcal{G}_0^{(\sqrt{\lambda+2\mu})}(t, \xi) - \mathcal{G}_0^{(\sqrt{\mu})}(t, \xi)\right)\mathcal{P} + \mathcal{G}_0^{(\sqrt{\mu})}(t, \xi)\right] m_0 \\ &+ \mathcal{F}^{-1}\left[\left(\mathcal{G}_1^{(\sqrt{\lambda+2\mu})}(t, \xi) - \mathcal{G}_1^{(\sqrt{\mu})}(t, \xi)\right)\mathcal{P} + \mathcal{G}_1^{(\sqrt{\mu})}(t, \xi)\right](m_1 + M[u]), \end{aligned}$$

$$\begin{aligned} &H(t, x) := \\ &\nabla^{-1}\mathcal{F}^{-1}\left[\left((\lambda + 2\mu)\mathcal{G}_1^{(\sqrt{\lambda+2\mu})}(t, \xi) - \mu\mathcal{G}_1^{(\sqrt{\mu})}(t, \xi)\right)\mathcal{P} + \mu\mathcal{G}_1^{(\sqrt{\mu})}(t, \xi)\right] m_0 \\ &+ \mathcal{F}^{-1}\left[\left(\mathcal{G}_0^{(\sqrt{\lambda+2\mu})}(t, \xi) - \mathcal{G}_0^{(\sqrt{\mu})}(t, \xi)\right)\mathcal{P} + \mathcal{G}_0^{(\sqrt{\mu})}(t, \xi)\right](m_1 + M[u]) \end{aligned}$$

and

$$\begin{aligned}
&\tilde{G}(t,x) := \\
&-\Delta\nabla^{-1}\mathcal{F}^{-1}\left[\left((\lambda+2\mu)\mathcal{G}_0^{(\sqrt{\lambda+2\mu})}(t,\xi)-\mu\mathcal{G}_0^{(\sqrt{\mu})}(t,\xi)\right)\mathcal{P}+\mu\mathcal{G}_0^{(\sqrt{\mu})}(t,\xi)\right]m_0 \\
&-\Delta\mathcal{F}^{-1}\left[\left((\lambda+2\mu)\mathcal{G}_1^{(\sqrt{\lambda+2\mu})}(t,\xi)-\mu\mathcal{G}_1^{(\sqrt{\mu})}(t,\xi)\right)\mathcal{P}+\mu\mathcal{G}_1^{(\sqrt{\mu})}(t,\xi)\right] \\
&\qquad (m_1+M[u]),
\end{aligned}$$

respectively, where $\mathcal{F}^{-1}$ represents the Fourier inverse transform. Now we formulate the approximation formulas of the global solutions by G, H and $\tilde{G}$ as $t\to\infty$.

Theorem 3 *The global solution $u(t)$ of (1) constructed in Theorem 1 satisfies the estimates*

$$\begin{aligned}
&\|\nabla^\alpha(u(t)-G(t))\|_2 = o(t^{-\frac{1}{4}-\frac{\alpha}{2}}), \quad 1\le\alpha\le 3,\\
&\|\nabla^\alpha(u(t)-G(t))\|_\infty = o(t^{-\frac{3}{2}-\frac{\alpha}{2}}), \quad 0\le\alpha\le 1,\\
&\|\nabla^\alpha(\partial_t u(t)-H(t))\|_2 = o(t^{-\frac{3}{4}-\frac{\alpha}{2}}), \quad 0\le\alpha\le 2,\\
&\|\nabla^2(\partial_t u(t)-H(t))\|_p = o(t^{-\frac{5}{2}(1-\frac{1}{p})-\frac{1}{2}}), \quad 2\le p\le 6,\\
&\|\nabla^\alpha(\partial_t u(t)-H(t))\|_\infty = o(t^{-2-\frac{\alpha}{2}}), \quad 0\le\alpha\le 1,\\
&\|\partial_t^2 u(t)-\tilde{G}(t)\|_p = o(t^{-\frac{5}{2}(1-\frac{1}{p})}), \quad 2\le p\le 6
\end{aligned}$$

as $t\to\infty$.

It is worth pointing out that as is seen in Theorem 2, we conclude that

$$u\in\{C^1((0,\infty);W^{2,6})\cap C^2((0,\infty);L^6)\}^3, \tag{8}$$

while we only have

$$u\in\{C^1((0,\infty);\bigcup_{2\le p<6}\dot{W}^{2,p})\cap W^{1,\infty}(0,\infty;W^{1,\infty})\cap C^2((0,\infty);\bigcup_{2\le p<6}L^p)\}^3$$

for $F(u)=\nabla u\nabla^2 u$, which is the reason why we separate the results corresponding to the nonlinear terms $F(u)=\nabla u\nabla^2 u$ (cf. [3]) and $F(u)=\nabla u\nabla\partial_t u$ in this paper. In other words, when we deal with the nonlinear term $F(u)=\nabla u\nabla^2 u$, it seems difficult for us to obtain the smoothing effect (8), even if we assume the extra regularity $f_1\in H^2$ as in Theorem 1–3. Indeed, the crucial point of the derivation of (8) is the estimation of $\|\nabla u\nabla^2 Du\|_{p_0}$ for some $p_0\in(2,6)$, where D is the t,x

gradient. When $F(u) = \nabla u \nabla \partial_t u$, it is easy to see that

$$\nabla F(u) = \nabla^2 u \nabla \partial_t u + \nabla u \nabla^2 \partial_t u.$$

Therefore taking $p_0 = 3$ and applying the Höloder inequality, the Sobolev inequality:

$$\|g\|_6 \le C\|\nabla g\|_2 \tag{9}$$

and the Gagliardo-Nirenberg inequality:

$$\|g\|_\infty \le C\|g\|_2^{\frac{1}{4}} \|\nabla^2 g\|_2^{\frac{3}{4}}, \tag{10}$$

we have

$$\begin{aligned} \|\nabla F(u)\|_3 &\le \|\nabla^2 u\|_6 \|\nabla \partial_t u\|_6 + \|\nabla u\|_\infty \|\nabla^2 \partial_t u\|_3 \\ &\le C\|\nabla^3 u\|_2 \|\nabla^2 \partial_t u\|_2 + C\|\nabla u\|_2^{\frac{1}{4}} \|\nabla^3 u\|_2^{\frac{3}{4}} \|\nabla^2 \partial_t u\|_3. \end{aligned} \tag{11}$$

Here we note that the estimate of $\|\nabla^2 \partial_t u\|_3$ for the case $F(u) = \nabla u \nabla \partial_t u$ is obtained independently from (8) (see Proposition 1 later). On the other hand, for the case $F(u) = \nabla u \nabla^2 u$, direct calculation gives

$$\nabla F(u) = (\nabla^2 u)^2 + \nabla u \nabla^3 u,$$

which implies we cannot expect the estimate $\|\nabla u \nabla^3 u\|_{p_0}$ for $p_0 > 2$, when u is a global solution in $\dot{H}^3 \cap \dot{H}^1$ as constructed in Theorem 1.1 in [3].

In the remainder part of this paper, we only prove the estimate (5). Indeed, it is also announced in [3] that Theorem 1 is obtained in a similar manner to the proof of Theorem 1.1 in [3] and that, once we have Theorem 2, Theorem 3 can be verified by the same way as in the proof of Theorem 1.6 in [3]. Moreover the estimate (7) is proved in the similar way to estimate (5). And we easily have the other estimates in Theorem 2, applying the same argument in [3].

2 Preliminaries

At first, we recall the decay properties of the Cauchy problem of the strong damped wave equations:

$$\begin{cases} \partial_t^2 w - \beta^2 \Delta w - \nu \Delta \partial_t w = 0, & t > 0, \quad x \in \mathbb{R}^3, \\ w(0,x) = w_0(x), \quad \partial_t w(0,x) = w_1(x), & x \in \mathbb{R}^3, \end{cases} \tag{12}$$

where $w = w(t,x) : (0,\infty)\times\mathbb{R}^3 \to \mathbb{R}$ and $\beta > 0$. For this purpose, we introduce the evolution operators $K_j^{(\beta)}(t)g$ for $j = 0, 1$, where the solutions of (12) is expressed by

$$w(t) = K_0^{(\beta)}(t)w_0 + K_1^{(\beta)}(t)w_1. \tag{13}$$

The decay properties of $K_j^{(\beta)}(t)g$ for $j = 0, 1$ are summarized as follows:

Lemma 1 ([3–6]) *Let $1 \le p \le \infty$, $1 \le q \le p$, $\ell \ge \tilde{\ell}_1 \ge 0$, $\ell \ge 2\tilde{\ell}_2 \ge 0$ and $\alpha \ge \tilde{\alpha} \ge 0$. Then it holds that*

$$\begin{aligned}\left\|\partial_t^\ell \nabla^\alpha K_0^{(\beta)}(t)g\right\|_p &\le C(1+t)^{-\frac{5}{2}(\frac{1}{q}-\frac{1}{p})+\frac{1}{2}-\frac{\ell-\tilde{\ell}_1+\alpha-\tilde{\alpha}}{2}}\|\nabla^{\tilde{\alpha}+\tilde{\ell}_1}g\|_q \\ &\quad + Ce^{-ct}(\|\nabla^{\alpha_1}g\|_p + t^{-\frac{3}{2}(\frac{1}{q}-\frac{1}{p})-\frac{\alpha-\tilde{\alpha}}{2}-(\ell-\frac{\tilde{\ell}_2}{2})+1}\|\nabla^{\tilde{\alpha}+\tilde{\ell}_2}g\|_q)\end{aligned} \tag{14}$$

for $\alpha_1 \ge \alpha$ and

$$\begin{aligned}\left\|\partial_t^\ell \nabla^\alpha K_1^{(\beta)}(t)g\right\|_p &\le C(1+t)^{-\frac{5}{2}(\frac{1}{q}-\frac{1}{p})+1-\frac{\ell-\tilde{\ell}_1+\alpha-\tilde{\alpha}}{2}}\|\nabla^{\tilde{\alpha}+\tilde{\ell}_1}g\|_q \\ &\quad + Ce^{-ct}(\|\nabla^{\alpha_1}g\|_p + t^{-\frac{3}{2}(\frac{1}{q}-\frac{1}{p})-\frac{\alpha-\tilde{\alpha}}{2}-(\ell-\frac{\tilde{\ell}_2}{2})+1}\|\nabla^{\tilde{\alpha}+\tilde{\ell}_2}g\|_q)\end{aligned} \tag{15}$$

for $\alpha_1 \ge \max\{\alpha-2, 0\}$.

Using the notation $K_j^{(\beta)}(t)g$ for $j = 0, 1$, we also have the expression of the solutions to (1) by the integral form:

Lemma 2 ([3]) *Let u be a solution of (1). Then it holds that*

$$u(t) = u_{lin}(t) + u_N(t),$$

where

$$\begin{aligned}u_{lin}(t) &:= (K_0^{(\sqrt{\lambda+2\mu})}(t) - K_0^{(\sqrt{\mu})}(t))\mathcal{F}^{-1}[\mathcal{P}\hat{f}_0] + K_0^{(\sqrt{\mu})}(t)f_0 \\ &\quad + (K_1^{(\sqrt{\lambda+2\mu})}(t) - K_1^{(\sqrt{\mu})}(t))\mathcal{F}^{-1}[\mathcal{P}\hat{f}_1] + K_1^{(\sqrt{\mu})}(t)f_1\end{aligned}$$

and

$$u_N(t) := \int_0^t (K_1^{(\sqrt{\lambda+2\mu})}(t-\tau) - K_1^{(\sqrt{\mu})}(t-\tau))\mathcal{F}^{-1}[\mathcal{P}\hat{F}(u)(\tau)]d\tau + \int_0^t K_1^{(\sqrt{\mu})}(t-\tau)F(u)(\tau)d\tau.$$

The following lemma is well-known the L^p-L^p boundedness of the Riesz transform:

Lemma 3 *Let $1 < p < \infty$. There exists $C > 0$ such that*

$$\|\mathcal{R}_a g\|_p \le C\|g\|_p, \tag{16}$$

where

$$\mathcal{R}_a g := \mathcal{F}^{-1}\left[\frac{\xi_a}{|\xi|}\hat{g}\right]$$

for $a = 1, 2, 3$.

For the proof, see e.g. [2].

3 Proof of Main Results

In this section, as mentioned above, we prove the estimate (5) in Theorem 2. To do so, we split the proof into two steps. Firstly we deal with the case $2 \le p < 6$, which is formulated as follows:

Proposition 1 *The solution $u(t)$ constructed in Theorem 1 satisifies*

$$u(t) \in \{C^1((0,\infty); \dot{W}^{2,p})\}^3$$

(5) for $2 \le p < 6$.

Proof Our proof starts with the observation of the smoothing effect of the linear solution. The estimates (14), (15) and (16) immediately lead to

$$\|\nabla^2 \partial_t u_{lin}(t)\|_{L^p(\mathbb{R}^3)} \le C\epsilon(1+t)^{-\frac{9}{4}+\frac{1}{p}} t^{-\frac{3}{4}+\frac{3}{2p}} \tag{17}$$

for $2 \le p \le 6$ and $t > 0$, where $\epsilon := \|f_0, f_1\|_Y$ and $Y := \{\dot{H}^3 \cap \dot{W}^{1,1}\}^3 \times \{H^2 \cap L^1\}^3$. For the nonlinear term, we firstly have

$$\|F(u)\|_1 \le C\|\nabla u\|_2 \|\nabla \partial_t u\|_2 \le C\epsilon(1+t)^{-2}, \tag{18}$$

$$\|F(u)\|_2 \le C\|\nabla u\|_\infty \|\nabla \partial_t u\|_2 \le C\epsilon(1+t)^{-\frac{11}{4}}, \tag{19}$$

$$\begin{aligned}\|\nabla F(u)\|_2 &\le C\|\nabla^2 u\|_4\|\nabla\partial_t u\|_4 + C\|\nabla u\|_\infty\|\nabla^2\partial_t u\|_2\\ &\le C\|\nabla u\|_\infty^{\frac12}\|\nabla^3 u\|_2^{\frac12}\|\partial_t u\|_\infty^{\frac12}\|\nabla^2\partial_t u\|_2^{\frac12} + C\|\nabla u\|_\infty\|\nabla^2\partial_t u\|_2\\ &\le C\epsilon(1+t)^{-\frac{13}{4}}\end{aligned} \tag{20}$$

by (2)–(4), (10), where we used the well-known fact (cf. [1])

$$\|\nabla g\|_{2p} \le C\|g\|_\infty^{\frac12}\|\nabla^2 g\|_p^{\frac12}, \quad 1 \le p < \infty$$

for (20). Hence we also obtain

$$\begin{aligned}\|F(u)(\tau)\|_p &\le \|F(u)(\tau)\|_2^{\frac{3}{p}-\frac12}\|F(u)(\tau)\|_6^{3(\frac12-\frac1p)}\\ &\le C\|F(u)(\tau)\|_2^{\frac{3}{p}-\frac12}\|\nabla F(u)(\tau)\|_2^{3(\frac12-\frac1p)} \le C\epsilon(1+\tau)^{-\frac72+\frac{3}{2p}}\end{aligned} \tag{21}$$

for $2 \le p \le 6$ by the Höloder inequality, (9), (19) and (20). Now we note that

$$\begin{aligned}&\left\|\nabla^2\partial_t\int_0^t K_1^{(\beta)}(t-\tau)\mathcal{R}_a\mathcal{R}_b F(u)(\tau)d\tau\right\|_p\\ &\le C\left\|\nabla^2\partial_t\int_0^t K_1^{(\beta)}(t-\tau)F(u)(\tau)d\tau\right\|_p\end{aligned} \tag{22}$$

by (16) for $2 \le p \le 6$. Therefore it follows from (22), (15), (18), (20) and (21) that

$$\begin{aligned}&\left\|\nabla^2\partial_t u_N(t)\right\|_p\\ &\le C\int_0^{\frac t2}(1+t-\tau)^{-\frac52(1-\frac1p)-\frac12}\|F(u)(\tau)\|_1 d\tau\\ &+ C\int_{\frac t2}^t(1+t-\tau)^{-\frac52(\frac12-\frac1p)}\|\nabla F(u)(\tau)\|_2 d\tau\\ &+ C\int_0^t e^{-c(t-\tau)}\{\|F(u)(\tau)\|_p + (t-\tau)^{-\frac32(\frac12-\frac1p)-\frac12}\|\nabla F(u)(\tau)\|_2\}d\tau\\ &\le C\epsilon\int_0^{\frac t2}(1+t-\tau)^{-\frac52(1-\frac1p)-\frac12}(1+\tau)^{-2}d\tau\\ &+ C\epsilon\int_{\frac t2}^t(1+t-\tau)^{-\frac52(\frac12-\frac1p)}(1+\tau)^{-\frac{13}{4}}d\tau\\ &+ C\epsilon\int_0^t e^{-c(t-\tau)}\{(1+\tau)^{-\frac72+\frac{3}{2p}} + (t-\tau)^{-\frac32(\frac12-\frac1p)-\frac12}(1+\tau)^{-\frac{13}{4}}\}d\tau\\ &\le C\epsilon(1+t)^{-\frac52(1-\frac1p)-\frac12}\end{aligned} \tag{23}$$

for $2 \le p < 6$, since we used the fact that $0 > -\frac{3}{2}(\frac{1}{2} - \frac{1}{p}) - \frac{1}{2} > -1$. Combining the estimates (17) and (23), we arrive at the desired estimate (5) for $2 \le p < 6$, which proves the proposition. □

Finally, we show the estimate (5) with $p = 6$. In the proof of Proposition 1, we already have the linear estimate (17) for $p = 6$. Then what is left is to show the estimate for the nonlinear term. For this aim, we observe that

$$\|\nabla F(u)(t)\|_3 \le C\epsilon(1+t)^{-\frac{7}{2}} + C\epsilon(1+t)^{-\frac{11}{3}}t^{-\frac{1}{2}} \le C\epsilon(1+t)^{-3}t^{-\frac{1}{2}}, \tag{24}$$

where we used (11), (5) with $p = 3$ and (21) with $p = 6$. Then we apply the estimates (22), (15), (18), (20), (21) and (24) to see that

$$\begin{aligned}
&\left\|\nabla^2 \partial_t u_N(t)\right\|_6 \\
&\le C\int_0^{\frac{t}{2}} (1+t-\tau)^{-\frac{31}{12}} \|F(u)(\tau)\|_1 d\tau \\
&+ C\int_{\frac{t}{2}}^t (1+t-\tau)^{-\frac{5}{6}} \|\nabla F(u)(\tau)\|_2 d\tau \\
&+ C\int_0^t e^{-c(t-\tau)}(\|F(u)(\tau)\|_6 + (t-\tau)^{-\frac{3}{4}}\|\nabla F(u)(\tau)\|_3) d\tau \\
&\le C\epsilon\int_0^{\frac{t}{2}} (1+t-\tau)^{-\frac{31}{12}}(1+\tau)^{-2} d\tau + C\epsilon\int_{\frac{t}{2}}^t (1+t-\tau)^{-\frac{5}{6}}(1+\tau)^{-\frac{13}{4}} d\tau \\
&+ C\epsilon\int_0^t e^{-c(t-\tau)}\{(1+\tau)^{-\frac{13}{4}} + (t-\tau)^{-\frac{3}{4}}(1+\tau)^{-3}\tau^{-\frac{1}{2}}\} d\tau \\
&\le C\epsilon(1+t)^{-\frac{7}{3}}t^{-\frac{1}{4}},
\end{aligned} \tag{25}$$

since

$$\begin{aligned}
&\int_0^t e^{-c(t-\tau)}(t-\tau)^{-\frac{3}{4}}(1+\tau)^{-3}\tau^{-\frac{1}{2}} d\tau \\
&\le Ce^{-ct}t^{-\frac{3}{4}}\int_0^{\frac{t}{2}} \tau^{-\frac{1}{2}} d\tau + C(1+t)^{-3}t^{-\frac{1}{2}}\int_{\frac{t}{2}}^t (t-\tau)^{-\frac{3}{4}} d\tau \\
&\le Ct^{-\frac{1}{4}}(e^{-ct} + (1+t)^{-3}).
\end{aligned}$$

Summing up the estimates (17) with $p = 6$ and (25), we conclude the estimate (5). We complete the proof.

Acknowledgments Y. Kagei was supported in part by JSPS Grant-in-Aid for Scientific Research (A) 20H00118. H. Takeda was partially supported by JSPS Grant-in-Aid for Scientific Research (C) 19K03596.

References

1. Cazenave, T.: Semilinear Schrödinger equations. Courant Lecture Notes in Mathematics, vol. 10. New York University, Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence (2003)
2. Grafakos, L.: Classical Fourier Analysis, 2nd edn. Graduate Texts in Mathematics, vol. 249. Springer, New York (2008)
3. Kagei, Y., Takeda, H.: Smoothing effect and large time behavior of solutions to nonlinear elastic wave equations with viscoelastic term. Nonlinear Anal. **219**, Paper No. 112826, 36 pp. (2022)
4. Kobayashi, T., Shibata, Y.: Remark on the rate of decay of solutions to linearized compressible Navier-Stokes equations. Pacific J. Math. **207**, 199–234 (2002)
5. Ponce, G.: Global existence of small solutions to a class of nonlinear evolution equations. Nonlinear Anal. **9**, 399–418 (1985)
6. Shibata, Y.: On the rate of decay of solutions to linear viscoelastic equation. Math. Methods Appl. Sci. **23**, 203–226 (2000)

Local and Global Solutions for the Semilinear Proca Equations in the de Sitter Spacetime

Makoto Nakamura

Abstract Local and global solutions for the semilinear Proca equations are considered in the de Sitter spacetime with spatially flat curvature. This paper briefly introduces some results in Nakamura (J Differ Equ 270:1218–1257, 2021) with their proofs. Based on these results, the effects of the spatial expansion on the existence of solutions of the equations are considered. Especially, it is remarked that global solutions for small data are shown based on the dissipative effect caused by the spatial expansion of exponential order.

1 Introduction

The Proca equations (see [6]) are the extension of the Maxwell equations with the massive terms taken into account, and they describe the massive vector boson with spin 1. The semilinear Proca equations are derived in the Minkowski spacetime in [5] as the Euler-Lagrange equation from a Lagrangian density. We consider the equations in the de Sitter spacetime with spatially flat curvature. The de Sitter spacetime is the solution of the Einstein equations with the cosmological constant in the vacuum under the cosmological principle. It describes the spatial expansion or contraction of exponential order.

We use the following convention. The Greek letters $\alpha, \beta, \gamma, \cdots$ run from 0 to n, and the Latin letters $j, k, \ell, \cdots$ run from 1 to n. We use the Einstein rule for the sum of indices, namely, the sum is taken for the same upper and lower repeated indices, for example, $\partial_j \phi^j := \sum_{j=1}^n \partial_j \phi^j$, $\partial_j \partial^j := \sum_{j=1}^n \partial_j \partial^j$, $T^\alpha{}_\alpha := \sum_{\alpha=0}^n T^\alpha{}_\alpha$ and $T^j{}_j := \sum_{j=1}^n T^j{}_j$ for any tensors ϕ^α and $T^\alpha{}_\beta$.

Put $x := (x^0, x^1, \cdots, x^n) \in \mathbb{R}^{1+n}$, $t := x^0$. Let $c > 0$, $H > 0$, $m > 0$, $\hbar > 0$ denote the speed of the light, the Hubble constant, the mass, the reduced Planck

M. Nakamura (✉)
Faculty of Science, Yamagata University, Yamagata, Japan

Graduate School of Information Science and Technology, Osaka University, Osaka, Japan
e-mail: nakamura@sci.kj.yamagata-u.ac.jp; makoto.nakamura.ist@osaka-u.ac.jp

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_41

constant, respectively. The de Sitter spacetime is the spacetime with a metric $\{g_{\alpha\beta}\}$ given by

$$-c^2(d\tau)^2 = g_{\alpha\beta}dx^\alpha dx^\beta := -c^2(dx^0)^2 + e^{2Hx^0}\sum_{j=1}^{n}(dx^j)^2, \tag{1}$$

where we have put the spatial curvature as 0, the variable τ denotes the proper time (see e.g., [1, 2]). When $H = 0$, the spacetime with (1) reduces to the Minkowski spacetime.

Let $g^{\alpha\beta}$ be defined such that the matrix $(g^{\alpha\beta})$ is the inverse matrix of the matrix $(g_{\alpha\beta})$. Put $\partial^\alpha := g^{\alpha\beta}\partial_\beta$ for $0 \le \alpha \le n$. Put $\nabla := (\partial_1, \cdots, \partial_n)$, $\Delta := \sum_{1\le j\le n}\partial^2/(\partial x^j)^2$. Put

$$Q := -\frac{(n-2)^2H^2}{4c^2} + \frac{m^2c^2}{\hbar^2} \tag{2}$$

which is called "curved mass." We consider the semilinear Proca equations in the de Sitter spacetime given by

$$c^{-2}\partial_t^2\phi^j - e^{-2Ht}\Delta\phi^j + Q\phi^j - \mu_0 e^{(n+2)Ht/2}J^j + e^{-n(p-1)Ht/2}f(\phi)^j = 0 \tag{3}$$

with the gauge condition

$$\operatorname{div}\phi := \partial_j\phi^j = 0 \tag{4}$$

for initial data

$$\phi^j(0,\cdot) = \phi_0^j(\cdot), \quad \partial_0\phi^j(0,\cdot) = \phi_1^j(\cdot) \tag{5}$$

for $1 \le j \le n$, where $\phi := (\phi^1, \cdots, \phi^n)$, $\phi_0 := (\phi_0^1, \cdots, \phi_0^n)$, $\phi_1 := (\phi_1^1, \cdots, \phi_1^n)$, $\mu_0 > 0$ is a constant, $J = (J^1, \cdots, J^n)$ is an electric current tensor, and we have put

$$f(\phi)^j := \lambda\left(\sum_{k=1}^{n}|\phi^k|^2\right)^{(p-1)/2}\phi^j$$

for $\lambda \in \mathbb{C}$, $p = 1 + 2\ell$ ($\ell = 0, 1, 2, \cdots$). When the current J has a potential K with $J^j := \partial^j K$, Eq. (3) with the condition (4) is rewritten as

$$c^{-2}\partial_t^2\phi^j - e^{-2Ht}\Delta\phi^j + Q\phi^j + e^{-n(p-1)Ht/2}\mathcal{H}f(\phi)^j = 0 \tag{6}$$

for $1 \le j \le n$, where $\mathcal{H}$ is the Helmholtz projection defined by

$$\mathcal{H}f(\phi)^j := f(\phi)^j - \partial^j(\partial_k\partial^k)^{-1}\partial_\ell f(\phi)^\ell \tag{7}$$

and $-\partial^j(\partial_k\partial^k)^{-1}\partial_\ell = R_jR_\ell$ for the Riesz transform $R_j := \partial_j/\sqrt{-\Delta}$ (see [7, Chapter VI] for the property of the Riesz transform). So that, the Cauchy problem (3), (4) and (5) is rewritten as

$$\begin{cases} \left(c^{-2}\partial_t^2\phi - e^{-2Ht}\Delta\phi + Q\phi + e^{-n(p-1)Ht/2}\mathcal{H}f(\phi)\right)(t,x) = 0, \\ \operatorname{div}\phi = 0, \\ \phi(0,x) = \phi_0(x), \quad \partial_0\phi(0,x) = \phi_1(x) \end{cases} \tag{8}$$

for $(t,x) \in [0,T) \times \mathbb{R}^n$. The condition (4) with $J^j = 0$ for $1 \le j \le n$ is known as the radiation gauge condition, which is the special case of the Coulomb gauge condition and the Lorenz gauge condition.

We use the Sobolev space $H^\mu(\mathbb{R}^n)$ and the homogeneous Sobolev space $\dot{H}^\mu(\mathbb{R}^n)$ of order $\mu \ge 0$. For simplicity, $\phi = (\phi^1, \cdots, \phi^n) \in H^\mu(\mathbb{R}^n)$ denotes $\phi^j \in H^\mu(\mathbb{R}^n)$ for $1 \le j \le n$ in the following.

For $0 \le \mu_0 \le \mu$, $T > 0$, $0 < R_0 \le R$ and $H \ge 0$, we put

$$X^\mu(T) := \{\phi;\ \|\phi\|_{X^\mu(T)} < \infty\},$$

$$X^{\mu_0,\mu}(T,R_0,R) := \left\{\phi \in X^\mu(T);\ \|\phi\|_{\dot{X}^{\mu_0}(T)} \le R_0,\ \|\phi\|_{X^\mu(T)} \le R\right\}$$

with $d(\phi,\psi) := \|\phi - \psi\|_{X^0(T)}$, where we have put

$$\begin{aligned} \|\phi\|_{\dot{X}^\mu(T)} := {} & \frac{1}{c}\|\partial_0\phi\|_{L^\infty((0,T),\dot{H}^\mu(\mathbb{R}^n))} + \|e^{-Ht}\nabla\phi\|_{L^\infty((0,T),\dot{H}^\mu(\mathbb{R}^n))} \\ & + \sqrt{Q}\|\phi\|_{L^\infty((0,T),\dot{H}^\mu(\mathbb{R}^n))} + \sqrt{H}\|e^{-Ht}\nabla\phi\|_{L^2((0,T),\dot{H}^\mu(\mathbb{R}^n))}, \end{aligned} \tag{9}$$

and its inhomogeneous version with $\dot{X}$, $\dot{H}$ replaced by X, H, respectively.

Definition 1 (Well-Posedness) For given $\mu \ge 0$, we consider the local and global well-posedness as follows.

(i) We say that the Cauchy problem (8) is locally well-posed if the following results hold. For any $\phi_0 \in H^{\mu+1}(\mathbb{R}^n)$ and $\phi_1 \in H^\mu(\mathbb{R}^n)$ with $\operatorname{div}\phi_0 = \operatorname{div}\phi_1 = 0$, there exist $T > 0$ and a unique solution $\phi \in C([0,T), H^{\mu+1}(\mathbb{R}^n)) \cap C^1([0,T), H^\mu(\mathbb{R}^n)) \cap X^\mu(T)$ of the Cauchy problem (8). Moreover, the solution depends on the initial data continuously in the sense that $d(\phi,\psi) \to 0$ as $\psi_0 \to \phi_0$ in $H^1(\mathbb{R}^n)$ and $\psi_1 \to \phi_1$ in $L^2(\mathbb{R}^n)$, where ψ is the solution of the problem (8) for the initial data $\psi_0(\cdot) = \psi(0,\cdot)$ and $\psi_1(\cdot) = \partial_0\psi(0,\cdot)$.

(ii) We say that the Cauchy problem (8) is globally well-posed if T can be taken as $T = \infty$ in (i).

Firstly, we consider the Cauchy problem (8) in the Minkowski spacetime (i.e., $H = 0$), which is compared with the case $H > 0$ in the following Theorem 2 to remark the effects by the spatial expansion. For $\mu_0 \geq 0$, ϕ_0 and ϕ_1 in (8), we put

$$D_{\mu_0} := \|\nabla\phi_0\|_{\dot{H}^{\mu_0}(\mathbb{R}^n)} + \sqrt{Q}\|\phi_0\|_{\dot{H}^{\mu_0}(\mathbb{R}^n)} + \frac{1}{c}\|\phi_1\|_{\dot{H}^{\mu_0}(\mathbb{R}^n)}.$$

Theorem 1 (Local Well-Posedness for $H = 0$) *Let $n \geq 2$, $H = 0$, $Q > 0$. Let $\mu_0, \mu \in \{0, 1, 2, 3, \cdots\}$ with $\mu_0 \leq \mu$ and $\mu_0 < n/2$. Let $p = 1 + 2\ell$, $\ell = 1, 2, 3, \cdots$, and*

$$p \begin{cases} \leq 1 + \frac{2}{n-2(\mu_0+1)} & \text{if } \mu_0 < \frac{n-2}{2}, \\ < \infty & \text{if } \mu_0 \geq \frac{n-2}{2}. \end{cases} \tag{10}$$

Then the Cauchy problem (8) is locally well-posed. Here, there exists a constant $C > 0$ which is independent of the data ϕ_0 and ϕ_1 such that T can be arbitrarily taken under the condition

$$0 < T \leq \frac{C}{D_{\mu_0}^{p-1}}. \tag{11}$$

Secondly, we consider the Cauchy problem (8) in the de Sitter spacetime with $H > 0$. We define $p_0(\mu_0)$ and q_{*0} by

$$p_0(\mu_0) := 1 + \frac{4}{n - 2\mu_0}, \quad \frac{1}{q_{*0}} := 1 - \frac{(p-1)(n-2\mu_0)}{4}. \tag{12}$$

Theorem 2 (Well-Posedness for $H > 0$) *Let $n \geq 2$. Let $H > 0$ satisfy $(n - 2)\hbar H/2 < mc^2$. Let $\mu_0, \mu \in \{0, 1, 2, 3, \cdots\}$ with $\mu_0 \leq \mu$ and $\mu_0 < n/2$. Let $p = 1 + 2\ell$, $\ell = 1, 2, 3, \cdots$, and p satisfy (10).*

(i) (Local well-posedness for $\mu_0 = 0$.) Let $\mu_0 = 0$. The Cauchy problem (8) is locally well-posed, where T can be arbitrarily taken under the condition

$$0 < T \leq \left(\frac{C}{D_0^{p-1}}\right)^{q_*}, \tag{13}$$

for some constant $C > 0$ which is independent of D_0. Here, q_* is arbitrarily fixed number which satisfies $1/q_{*0} \leq 1/q_* \leq 1$ and $1/q_* > 0$.

(ii) (Global well-posedness for $\mu_0 = 0$.) Let $\mu_0 = 0$ and $1 + 4/n \leq p$. If D_0 is sufficiently small, then the Cauchy problem (8) is globally well-posed.

(iii) (Local and global well-posedness for $\mu_0 > 0$.) Let $\mu_0 > 0$. The Cauchy problem (8) is locally well-posed, where T can be arbitrarily taken under the condition

$$0 < T \leq -\frac{1}{H\mu_0(p-1)} \log\left(1 - \mu_0(p-1)q_* H \left(\frac{C}{D_{\mu_0}^{p-1}}\right)^{q_*}\right) \tag{14}$$

for some constant $C > 0$ which is independent of D_{μ_0}, when

$$D_{\mu_0} > \left\{C\,(\mu_0(p-1)q_* H)^{1/q_*}\right\}^{1/(p-1)} =: D_*.$$

Here, q_* is arbitrary number which satisfies $1/q_{*0} \leq 1/q_* \leq 1$ and $1/q_* > 0$. When $D_{\mu_0} \leq D_*$, the Cauchy problem (8) is globally well-posed.

(iv) (Global well-posedness for $\mu_0 > 0$ and $p \geq p_0(\mu_0)$.) Let $\mu_0 > 0$ and $p \geq p_0(\mu_0)$. There exists a constant $C > 0$ such that if $D_{\mu_0} \leq C H^{1/(p-1)}$, then the Cauchy problem (8) is globally well-posed.

(v) (Asymptotic behaviors of global solutions.) The global solution ϕ obtained above satisfies

$$\|\phi(t) - \phi_+(t)\|_{H^{\mu-1}(\mathbb{R}^n)} \to 0 \ \text{ and } \ \|\partial_t\left(\phi(t) - \phi_+(t)\right)\|_{H^{\mu-1}(\mathbb{R}^n)} \to 0$$

as $t \to \infty$, where ϕ_+ is defined by

$$\begin{aligned}\phi_+(t) :=& K_0(t)\left(\phi_0 + c^2\int_0^\infty K_1(s)h(s)ds\right) + K_1(t)\\ &\left(\phi_1 - c^2\int_0^\infty K_0(s)h(s)ds\right)\end{aligned}$$

for $h(s) := e^{-n(p-1)Hs/2}\mathcal{H}f(\phi)(s)$, and K_0, K_1 are defined by (18), below.

In Theorem 2, we obtain the global solutions for small data which follow from the dissipative terms $\sqrt{H}\|e^{-Ht}\nabla\phi\|_{L^2\dot{H}^\mu}$ in (9) by the energy estimate when $H > 0$. The spatial expansion yields the dissipative effects to the Proca equations, and it strongly diminishes the nonlinear property of the semilinear equations. This is the first result on the semilinear Proca equations in the de Sitter spacetime. Theorem 2 plays as a toy model to study the effect by the spatial expansion for the Cauchy problem, and it will be useful for the study of the coupled equations in [3, 8, 9]. We denote the inequality $A \leq CB$ by $A \lesssim B$ for some constant $C > 0$ which is not essential.

2 Proof of Theorem 1

We give the proof for general $H \geq 0$ without the restriction $H = 0$ until (30), below, since we use the argument to prove Theorem 2. Let n, μ_0, μ and p satisfy the assumption in Theorem 1. Put

$$\frac{1}{r_*} := \frac{n-2\mu_0}{2np}, \quad \frac{1}{r_{**}} := \frac{1}{r_*} + \frac{\mu_0}{n}, \quad \theta := \frac{(n-2\mu_0)(p-1)}{2p}. \tag{15}$$

Then $1 < r_* < \infty$ and $1 < r_{**} < \infty$ by $0 \leq \mu_0 < n/2$ and $p \geq 1$. And $0 \leq \theta \leq 1$ by $0 \leq \mu_0 < n/2$ and (10). Let $p_0(\mu_0)$ and q_{*0} be defined by (12). We take q_* such that

$$\max\left\{\frac{1}{q_{*0}}, 0\right\} \leq \frac{1}{q_*} \leq 1. \tag{16}$$

Thus, we can take q_0 and q such that

$$0 \leq \frac{1}{q_0} \leq \frac{1}{2}, \quad 0 \leq \frac{1}{q} \leq \frac{1}{2} \text{ and } \frac{1}{q_*} = 1 - \frac{\theta(p-1)}{q_0} - \frac{\theta}{q} \tag{17}$$

since $q_* = 1$ for $q_0 = q = \infty$ and $q_* = q_{*0}$ for $q_0 = q = 2$ in (17).

We regard the solution of the Cauchy problem (8) as the solution of the integral equation given by

$$\phi(t) = \Psi(\phi)(t) := K_0(t)\phi_0 + K_1(t)\phi_1 - \int_0^t K(t,s)e^{-n(p-1)Hs/2}\mathcal{H}f(\phi)(s)ds, \tag{18}$$

where K_0 and K_1 denote the free propagator of the linear Proca equations, and K denotes the propagator for the inhomogeneous term. So that, the solution is obtained as the fixed point of the operator Ψ.

We show that Ψ is a contraction mapping on $X^{\mu_0,\mu}(T, R_0, R)$ for some $T > 0$, $R_0 > 0$ and $R > 0$. Since $f(\phi)$ is a polynomial of order p, we have

$$f(\phi) = \sum_{1 \leq j_1, \cdots, j_p \leq n} C(j_1, \cdots, j_p)\phi^{j_1} \cdots \phi^{j_p}$$

for $\phi = (\phi^1, \cdots, \phi^n)$ and some constants $\{C(j_1, \cdots, j_p)\}_{1 \leq j_1, \cdots, j_p \leq n}$, where we only consider the case ϕ is real-valued since the case ϕ is complex-valued follows

similarly. For any $\mu > 0$ and any multi-index μ_* with $|\mu_*| = \mu$, we have

$$\partial^{\mu_*} f(\phi) = \sum_{\substack{1 \le j_1, \cdots, j_p \le n \\ \mu_1 + \cdots + \mu_p = \mu_*}} C(j_1, \cdots, j_p, \mu_1, \cdots, \mu_p) \prod_{k=1}^{p} \partial^{\mu_k} \phi^{j_k}.$$

Put

$$\frac{1}{r_k} := \frac{1 - |\mu_k|/\mu}{r_*} + \frac{|\mu_k|}{\mu r_{**}} \tag{19}$$

for $1 \le k \le p$. Then $1 < r_k < \infty$ holds by $1 < r_*, r_{**} < \infty$ and $0 \le |\mu_k|/\mu \le 1$. We have

$$\|\partial^{\mu_*} f(\phi)\|_{L^2} \lesssim \sum_{\substack{1 \le j_1, \cdots, j_p \le n \\ \mu_1 + \cdots + \mu_p = \mu_*}} \prod_{k=1}^{p} \|\partial^{\mu_k} \phi^{j_k}\|_{L^{r_k}}$$

by the Hölder inequality. Since we have the interpolation inequality

$$\|\partial^{\mu_k} \phi^{j_k}\|_{L^{r_k}} \lesssim \|\phi^{j_k}\|_{L^{r_*}}^{1-|\mu_k|/\mu} \|\phi^{j_k}\|_{\dot{H}^{\mu, r_{**}}}^{|\mu_k|/\mu}$$

by (19), we obtain

$$\frac{1}{2} = \frac{1}{r_1} + \cdots + \frac{1}{r_p} = \frac{p-1}{r_*} + \frac{1}{r_{**}}, \quad \|\partial^{\mu_*} f(\phi)\|_{L^2} \lesssim \|\phi\|_{L^{r_*}}^{p-1} \|\phi\|_{\dot{H}^{\mu, r_{**}}}. \tag{20}$$

This inequality also holds when $\mu = \mu_* = 0$ since

$$\|f(\phi)\|_{L^2} \lesssim \|\phi\|_{L^{2p}}^{p} \text{ and } \|\phi\|_{L^{2p}} \le \|\phi\|_{L^{r_*}}^{1-1/p} \|\phi\|_{L^{r_{**}}}^{1/p}$$

hold by (20). Moreover, since r_*, r_{**} and θ satisfy

$$0 < \frac{1}{r_{**}} = \frac{1}{2} - \frac{\theta}{n} = \frac{1-\theta}{2} + \theta\left(\frac{1}{2} - \frac{1}{n}\right) < 1, \quad 0 < \frac{1}{r_*} = \frac{1}{r_{**}} - \frac{\mu_0}{n} < 1,$$
$$0 \le \theta \le 1$$

by (15), we have the interpolation inequalities

$$\|\phi\|_{L^{r_*}} \lesssim \|\phi\|_{\dot{H}^{\mu_0, r_{**}}} \lesssim \|\phi\|_{\dot{H}^{\mu_0}}^{1-\theta} \|\phi\|_{\dot{H}^{\mu_0+1}}^{\theta} \tag{21}$$

and

$$\|\phi\|_{\dot{H}^{\mu,r_{**}}} \lesssim \|\phi\|_{\dot{H}^{\mu}}^{1-\theta} \|\phi\|_{\dot{H}^{\mu+1}}^{\theta}. \tag{22}$$

By these inequalities and (20), we have

$$\|f(\phi)\|_{\dot{H}^{\mu}} \lesssim \|\phi\|_{\dot{H}^{\mu_0}}^{(1-\theta)(p-1)} \|\phi\|_{\dot{H}^{\mu_0+1}}^{\theta(p-1)} \|\phi\|_{\dot{H}^{\mu}}^{1-\theta} \|\phi\|_{\dot{H}^{\mu+1}}^{\theta}. \tag{23}$$

Put $h(\phi) := \mathcal{H} f(\phi) e^{-n(p-1)Ht/2}$. By (23), we have

$$\begin{aligned} \|h(\phi)\|_{\dot{H}^{\mu}} &\lesssim \|f(\phi)\|_{\dot{H}^{\mu}} e^{-n(p-1)Ht/2} \\ &\lesssim \|\phi\|_{\dot{H}^{\mu_0}}^{(1-\theta)(p-1)} \cdot \left\|e^{-Ht}\phi\right\|_{\dot{H}^{\mu_0+1}}^{\theta(p-1)} \cdot \left\|e^{-Ht}\phi\right\|_{\dot{H}^{\mu+1}}^{\theta} \cdot \|\phi\|_{\dot{H}^{\mu}}^{1-\theta} \cdot I, \end{aligned} \tag{24}$$

where we have used the boundedness of the Helmholtz projector $\mathcal{H}$ on $L^2(\mathbb{R}^n)$ and we have put

$$I := e^{(\theta p - n(p-1)/2)Ht}.$$

By the Hölder inequality for the time variable, we have

$$\begin{aligned} \|h(\phi)\|_{L^1\dot{H}^{\mu}} \lesssim \|\phi\|_{L^{\infty}\dot{H}^{\mu_0}}^{(1-\theta)(p-1)} \cdot \left\|e^{-Ht}\phi\right\|_{L^{q_0}\dot{H}^{\mu_0+1}}^{\theta(p-1)} \cdot \|\phi\|_{L^{\infty}\dot{H}^{\mu}}^{1-\theta} \cdot \left\|e^{-Ht}\phi\right\|_{L^{q}\dot{H}^{\mu+1}}^{\theta} \\ \cdot \|I\|_{L^{q_*}}, \end{aligned}$$

where q_0, q, q_* satisfy (16) and (17). So that, we have

$$\|h(\phi)\|_{L^1\dot{H}^{\mu}} \lesssim R_0^{p-1} R \|I\|_{L^{q_*}} \tag{25}$$

for $\phi \in X^{\mu_0,\mu}(T, R_0, R)$. Analogously, we have

$$\|h(\phi) - h(\psi)\|_{L^1L^2} \lesssim R_0^{p-1} \|I\|_{q_*} d(\phi, \psi) \tag{26}$$

for $\phi, \psi \in X^{\mu_0,\mu}(T, R_0, R)$.

We prepare the following standard energy estimates.

Lemma 1 (Energy Estimates for $H \geq 0$) *Let μ be a real number, and let $h = (h^1, \cdots, h^n) \in L^1((0,\infty), \dot{H}^{\mu}(\mathbb{R}^n))$. Let Q be the constant defined in (2). Let $\phi = (\phi^1, \cdots, \phi^n)$ be the solution of the Cauchy problem*

$$\begin{cases} -\partial_{\alpha}\partial^{\alpha}\phi^j(t,x) + Q\phi^j(t,x) + h^j(t,x) = 0, \\ \phi^j(0,x) = \phi_0^j(x), \quad \partial_t\phi^j(0,x) = \phi_1^j(x) \end{cases} \tag{27}$$

for $t \geq 0$, $x \in \mathbb{R}^n$ *and* $1 \leq j \leq n$. *If* $H \geq 0$ *and* $Q \geq 0$, *then the solution* ϕ *satisfies*

$$\begin{aligned}
&\frac{1}{c}\|\partial_0\phi^j\|_{L^\infty((0,T),\dot{H}^\mu(\mathbb{R}^n))} + \|e^{-Ht}\nabla\phi^j\|_{L^\infty((0,T),\dot{H}^\mu(\mathbb{R}^n))} \\
&+\sqrt{Q}\|\phi^j\|_{L^\infty((0,T),\dot{H}^\mu(\mathbb{R}^n))} + \sqrt{H}\|e^{-Ht}\nabla\phi^j\|_{L^2((0,T),\dot{H}^\mu(\mathbb{R}^n))} \\
&\lesssim \frac{1}{c}\|\phi_1^j\|_{\dot{H}^\mu(\mathbb{R}^n)} + \|\nabla\phi_0^j\|_{\dot{H}^\mu(\mathbb{R}^n)} + \sqrt{Q}\|\phi_0^j\|_{\dot{H}^\mu(\mathbb{R}^n)} + \|h^j\|_{L^1((0,T),\dot{H}^\mu(\mathbb{R}^n))}
\end{aligned}$$

for $0 < T \leq \infty$ *and* $1 \leq j \leq n$.

By Lemma 1 and (25), we have

$$\|\Psi(\phi)\|_{\dot{X}^\mu} \lesssim \dot{D}^\mu + \|h(\phi)\|_{L^1\dot{H}^\mu} \lesssim \dot{D}^\mu + R_0^{p-1}R\|I\|_{q_*},$$

namely,

$$\|\Psi(\phi)\|_{\dot{X}^\mu} \leq C_0\dot{D}^\mu + CR_0^{p-1}R\|I\|_{q_*}$$

for some constants $C_0 > 0$ and $C > 0$, where we have put

$$\dot{D}^\mu := \frac{1}{c}\|\phi_1\|_{\dot{H}^\mu} + \|\nabla\phi_0\|_{\dot{H}^\mu} + \sqrt{Q}\,\|\phi_0\|_{\dot{H}^\mu}.$$

So that, we obtain

$$\|\Psi(\phi)\|_{\dot{X}^{\mu_0}} \leq R_0 \ \text{ and } \ \|\Psi(\phi)\|_{X^\mu} \leq R \tag{28}$$

if

$$R_0 \geq 2C_0\dot{D}^{\mu_0}, \quad R \geq 2C_0\max\{\dot{D}^0, \dot{D}^\mu\} \ \text{ and } \ 2CR_0^{p-1}\|I\|_{q_*} \leq 1 \tag{29}$$

for $\mu \geq 0$. Analogously, by Lemma 1 and (26), we have

$$d(\Psi(\phi), \Psi(\psi)) \leq \frac{1}{2}\,d(\phi, \psi) \tag{30}$$

under (29).

When $H = 0$, we have $Q = (mc/\hbar)^2$ by (2). We take $q_0 = q = \infty$ and $q_* = 1$ which satisfy (16) and (17). Since we have $\|I\|_1 = T$, the condition $2CR_0^{p-1}\|I\|_{q_*} \leq 1$ in (29) is rewritten as

$$T \leq \frac{1}{2CR_0^{p-1}}.$$

So that, Ψ is a contraction mapping under (11) taking $R_0 = 2C_0\dot{D}^{\mu_0}$ in (29). From this, we obtain the local solution under the condition (11). The continuity of the solution follows from the continuity of the free propagators and (25). We refer to [4] for proofs of the uniqueness of the solution, and the continuous dependence of solutions on initial data.

3 Proof of Theorem 2

We are able to use the same argument in the proof of Theorem 1 until (30). We note that the continuity of the solution, the uniqueness of the solution, and the continuous dependence of the solution on the initial data follow from the same arguments in the proof of Theorem 1. When $H > 0$ and $(n-2)H < 2mc^2/\hbar$, we have $Q > 0$ by (2). We have

$$\|I\|_{q_*} = (2H)^{1/q_*-1} \|e^{-\mu_0(p-1)Ht}\|_{L_t^{q_*}((0,T))} \tag{31}$$

and

$$\begin{aligned} &\|e^{-\mu_0(p-1)Ht}\|_{L_t^{q_*}((0,T))} \\ &= \begin{cases} 1 & \text{if } q_* = \infty, \quad H\mu_0 \geq 0, \\ e^{-H\mu_0(p-1)T} \ (\geq 1) & \text{if } q_* = \infty, \quad H\mu_0 < 0, \\ T^{1/q_*} & \text{if } q_* < \infty, \quad H\mu_0 = 0, \\ \left\{ \frac{1}{H\mu_0(p-1)q_*} \left(1 - e^{-H\mu_0(p-1)q_*T}\right) \right\}^{1/q_*} & \text{if } q_* < \infty, \quad H\mu_0 \neq 0. \end{cases} \end{aligned} \tag{32}$$

(i) When $\mu_0 = 0$ and p satisfies (10), we take q_* such that

$$1 - \frac{n(p-1)}{4} \leq \frac{1}{q_*} \leq 1 \ \text{ and } \ 0 < \frac{1}{q_*}.$$

The condition $2CR_0^{p-1}\|I\|_{q_*} \leq 1$ in (29) is rewritten as

$$T \lesssim R_0^{-(p-1)q_*}$$

by (32). So that, Ψ is a contraction mapping under (13), and we obtain the local solution of the Cauchy problem.

(ii) When $\mu_0 = 0$ and p satisfies (10) and $(p_0(0) =) \ 1 + 4/n \leq p$, we are able to take $q_* = \infty$ by (12) and (16). Then we have $\|I\|_{q_*} \lesssim 1$ by (32), by which the condition (29) holds for sufficiently small $\dot{D}^0$ and $\dot{D}^{\mu}$. Namely, we obtain the global solutions for small data.

(iii) When $\mu_0 > 0$ and p satisfies (10), we take q_* such that

$$\frac{1}{q_{*0}} \le \frac{1}{q_*} \le 1, \quad \frac{1}{q_*} > 0,$$

where q_{*0} is defined by (12). Then we have

$$\|e^{-H\mu_0(p-1)t}\|_{L_t^{q_*}((0,T))} = \left\{ \frac{1}{H\mu_0(p-1)q_*} \left(1 - e^{-H\mu_0(p-1)q_*T}\right) \right\}^{1/q_*} \to 0$$

as $T \searrow 0$, and

$$\|I\|_{q_*} \le (2H)^{1/q_*-1} \left\{ \frac{1}{H\mu_0(p-1)q_*} \right\}^{1/q_*}$$

by (31). Put

$$I\!I := \frac{H\mu_0(p-1)q_*}{(2CR_0^{p-1})^{q_*}(2H)^{1-q_*}}.$$

The condition $2CR_0^{p-1}\|I\|_{q_*} \le 1$ in (29) is rewritten as

$$1 - I\!I \le e^{-H\mu_0(p-1)q_*T},$$

namely,

$$\begin{cases} T \le -\frac{1}{H\mu_0(p-1)} \log(1 - I\!I) & \text{if } I\!I < 1, \\ T < \infty & \text{if } I\!I \ge 1. \end{cases} \tag{33}$$

We note $I\!I < 1$ is rewritten as

$$R_* := \left\{ \left(\frac{\mu_0(p-1)q_*}{2} \right)^{1/q_*} \frac{H}{C} \right\}^{1/(p-1)} < R_0.$$

So that, Ψ is a contraction mapping when $R_* < R_0$ and T satisfies (33), or when $R_* \ge R_0$ and $T < \infty$. Therefore, we obtain local solutions under the condition (14), where we take a different C if necessary, and we also obtain global solutions if initial data are sufficiently small.

(iv) When $\mu_0 > 0$ and p satisfies (10) and $p_0(\mu_0) \le p$, we are able to take $q_* = \infty$. Then we have

$$\|I\|_{q_*} = \frac{1}{2H}.$$

The conditions in (29) are satisfied if the initial data are sufficiently small such that

$$2C_0\dot{D}^{\mu_0} \le R_0 \le \left(\frac{H}{C}\right)^{1/(p-1)}.$$

(v) The required result follows since $h(\phi) \in L^1((0,\infty), H^{\mu}(\mathbb{R}^n))$ holds by (25).

Acknowledgments This work was supported by JSPS KAKENHI Grant Number JP16H03940.

References

1. Carroll, S.: Spacetime and Geometry. An Introduction to General Relativity, xiv+513 pp. Addison Wesley, San Francisco (2004)
2. d'Inverno, R.: Introducing Einstein's Relativity, xii+383 pp. The Clarendon Press, Oxford University Press, New York (1992)
3. Huh, H.: The Cauchy problem for Chern-Simons-Proca-Higgs equations. Lett. Math. Phys. **91**(1), 29–44 (2010)
4. Nakamura, M.: On the Cauchy problem for the semilinear Proca equations in the de Sitter spacetime. J. Differ. Equ. **270**, 1218–1257 (2021)
5. Nobre, F.D., Plastino, A.R.: Generalized nonlinear Proca equation and its free-particle solutions. Eur. Phys. J. C **76** (2016). Article number: 343
6. Proca, A.: Sur la théorie ondulatoire des électrons positifs et négatifs. J. Phys. Radium **7**, 347–353 (1936)
7. Stein, E.M., Weiss, G.: Introduction to Fourier Analysis on Euclidean Spaces, x+297 pp. Princeton Mathematical Series, No. 32. Princeton University Press, Princeton (1971)
8. Tsutsumi, Y.: Global solutions for the Dirac-Proca equations with small initial data in $3+1$ space time dimensions. J. Math. Anal. Appl. **278**(2), 485–499 (2003)
9. Vuille, C., Ipser, J., Gallagher, J.: Einstein-Proca model, micro black holes, and naked singularities. Gen. Relat. Gravit. **34**(5), 689–696 (2002)

Numerical Simulations of Semilinear Klein–Gordon Equation in the de Sitter Spacetime with Structure-Preserving Scheme

Takuya Tsuchiya and Makoto Nakamura

Abstract We perform some simulations of the semilinear Klein–Gordon equation in the de Sitter spacetime. We reported the accurate numerical results of the equation with the structure-preserving scheme (SPS) in an earlier publication (Tsuchiya and Nakamura, J Comput Appl Math 361:396–412, 2019). To investigate the factors for the stability and accuracy of the numerical results with SPS, we perform some simulations with three discretized formulations. The first formulation is the discretized equations with SPS, the second one is with SPS that replaces the second-order difference as the standard second-order central difference, and the third one is with SPS that replaces the discretized nonlinear term as the standard discretized expression. As a result, the above two replacements in SPS are found to be effective for accurate simulations. On the other hand, the ingenuity of replacing the second-order difference in the first formulation is not effective for maintaining the stability of the simulations.

1 Introduction

Stable and accurate numerical simulations are necessary for understanding natural and social phenomena in detail. To realize this, numerical methods such as discretizations should be in a mathematically guaranteed format because the numerical errors mainly occur during the processes of discretizations. For numerical schemes of partial differential equations, there are several well-known methods such as the Crank–Nicolson and Runge–Kutta schemes. However, it is difficult to perform stable and accurate numerical simulations for nonlinear partial differential equations

T. Tsuchiya (✉)
Center for Liberal Arts and Sciences, Hachinohe Institute of Technology, Aomori, Japan
e-mail: t-tsuchiya@hi-tech.ac.jp

M. Nakamura
Department of Pure and Applied Mathematics, Graduate School of Information Science and Technology, Osaka University, Osaka, Japan
e-mail: makoto.nakamura.ist@osaka-u.ac.jp

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_42

since there are large numerical errors and vibrations in the solutions caused by nonlinearity. Thus, suitable schemes have been suggested to perform successful simulations. One of the schemes is the structure-preserving scheme (SPS) [1, 2]. This scheme conserves some structures at the continuous level, and thus enables stable and accurate numerical simulations.

In this paper, we review the discretized equations of the semilinear Klein–Gordon equation in the de Sitter spacetime with SPS and perform some simulations to investigate their stability and accuracy. For investigating the semilinear Klein–Gordon equation in the de Sitter spacetime, analytical [3–7] and numerical [8, 9] research studies have been conducted. In [9], we reported some accurate numerical results of the semilinear Klein–Gordon equation with SPS. There are some differences between the standard discretized equation and the discretized equation with SPS. In this paper, we investigate the factors for the stability and accuracy of the simulations. Here, stability means that the solution does not have vibrations in the simulations, and accuracy means the conservation of constraints in the simulations. In general, the accuracy of the simulations would be determined by examining the numerical solution of the equations. However, for the nonlinear differential equations, it is often difficult to investigate the accuracy because of the complexities. Thus, we adopt the constraints of the system as the criteria of the accuracy in this paper.

The structure of this paper is as follows. We review the canonical formulation of the semilinear Klein–Gordon equation in the de Sitter spacetime in Sect. 2 and the discretized equation with SPS in Sect. 3. In Sect. 4, we perform some simulations for investigating their stability and accuracy. We summarize this paper in Sect. 5. In this paper, indices such as $(i, j, k, \cdots)$ run from 1 to 3. We use the Einstein convention of summation of repeated up–down indices.

2 Canonical Formulation of Semilinear Klein–Gordon Equation in the de Sitter Spacetime

The semilinear Klein–Gordon equation in the de Sitter spacetime is given by

$$\partial_t^2\phi + 3H\partial_t\phi - e^{-2Ht}\delta^{ij}(\partial_i\partial_j\phi) + m^2\phi + \lambda|\phi|^{p-1}\phi = 0, \tag{1}$$

where ϕ is the field variable, H is the Hubble constant, δ^{ij} denotes the Kronecker delta, m is the mass, λ is a Boolean parameter, and p is an integer of 2 or more. In performing the simulations of Eq. (1), we often recast first-order system. In this paper, we adopt the canonical formulation as the first-order system. This is because the canonical formulation has the total Hamiltonian, and we can treat this value as a criterion for investigating the accuracy since the value is a constraint.

The Hamiltonian density of Eq. (1) is defined as

$$\mathcal{H} := \frac{1}{2}e^{-3Ht}\psi^2 + \frac{1}{2}e^{Ht}\delta^{ij}(\partial_i\phi)(\partial_j\phi) + \frac{1}{2}m^2e^{3Ht}\phi^2 + \frac{\lambda}{p+1}e^{3Ht}|\phi|^{p+1}, \tag{2}$$

where ψ is the conjugate momentum of ϕ. Then, using the canonical equations of $\mathcal{H}$, we obtain the evolution equations as

$$\partial_t\phi := \frac{\delta\mathcal{H}}{\delta\psi} = e^{-3Ht}\psi, \tag{3}$$

$$\partial_t\psi := -\frac{\delta\mathcal{H}}{\delta\phi} = e^{Ht}\delta^{ij}(\partial_j\partial_i\phi) - m^2e^{3Ht}\phi - \lambda e^{3Ht}|\phi|^{p-1}\phi. \tag{4}$$

The total Hamiltonian H_C is defined as

$$H_C := \int_{\mathbb{R}^3}\mathcal{H}d^3x, \tag{5}$$

and the time derivative of H_C with the evolution Eqs. (3) and (4) is

$$\partial_t H_C = H\int_{\mathbb{R}^3} d^3x\Big\{-\frac{3}{2}e^{-3Ht}\psi^2 + \frac{1}{2}e^{Ht}\delta^{ij}(\partial_i\phi)(\partial_j\phi) + \frac{3}{2}m^2e^{3Ht}\phi^2 + \frac{3\lambda}{p+1}e^{3Ht}|\phi|^{p+1}\Big\} + \int_{\mathbb{R}^3}\partial_j\{e^{-3Ht}\delta^{ij}\psi(\partial_i\phi)\}d^3x. \tag{6}$$

Note that H is the Hubble constant and H_C is the total Hamiltonian. If $H = 0$ and we set the boundary conditions under which the last term on the right-hand side of Eq. (6) is zero on the boundary, then $\partial_t H_C = 0$. Thus, H_C is treated as a conserved quantity. On the other hand, in the case of $H \neq 0$, H_C is not a conserved quantity in general. In the case of $H \neq 0$, we define the value as

$$\tilde{H}_C(t) := H_C(t) - \int_0^t \partial_s H_C(s)ds. \tag{7}$$

$\tilde{H}_C$ identically satisfies $\partial_t\tilde{H}_C = 0$. We call the value $\tilde{H}_C$ as the modified total Hamiltonian hereafter. In the case of $H \neq 0$, we adopt the value $\tilde{H}_C$ as a criterion for the accuracy of the simulations. To investigate the accuracy of the simulations, we monitor H_C in a flat spacetime such as $H = 0$, and $\tilde{H}_C$ in a nonflat spacetime such as $H = 10^{-3}$. If the changes in H_C in the flat spacetime or $\tilde{H}_C$ in the nonflat spacetime against the initial values are sufficiently small during the evolution, we determine that the simulations are successful. That is, the smaller the change in H_C or $\tilde{H}_C$ in the evolution, the more accurate the numerical calculations.

3 Discretizations of Semilinear Klein–Gordon Equation in the de Sitter Spacetime

The main factor for the numerical errors occurs during the processes of the discretizations of the equations. In this section, we review the discretized equations of the semilinear Klein–Gordon equation in the de Sitter spacetime.

The discretized Hamiltonian density is defined as

$$\mathcal{H}^{(\ell)}_{(\boldsymbol{k})} := \frac{1}{2}e^{-3Ht^{(\ell)}}(\psi^{(\ell)}_{(\boldsymbol{k})})^2 + \frac{1}{2}e^{Ht^{(\ell)}}\delta^{ij}(\hat{\delta}^{\langle 1\rangle}_i\phi^{(\ell)}_{(\boldsymbol{k})})(\hat{\delta}^{\langle 1\rangle}_j\phi^{(\ell)}_{(\boldsymbol{k})}) + \frac{1}{2}m^2e^{3Ht^{(\ell)}}(\phi^{(\ell)}_{(\boldsymbol{k})})^2 + \frac{\lambda}{p+1}e^{3Ht^{(\ell)}}|\phi^{(\ell)}_{(\boldsymbol{k})}|^{p+1}. \tag{8}$$

By using SPS, we can rewrite the discretized Eqs. (3) and (4) as

$$\frac{\phi^{(\ell+1)}_{(\boldsymbol{k})} - \phi^{(\ell)}_{(\boldsymbol{k})}}{\Delta t} = \frac{1}{4}(e^{-3Ht^{(\ell+1)}} + e^{-3Ht^{(\ell)}})(\psi^{(\ell+1)}_{(\boldsymbol{k})} + \psi^{(\ell)}_{(\boldsymbol{k})}), \tag{9}$$

$$\begin{aligned}\frac{\psi^{(\ell+1)}_{(\boldsymbol{k})} - \psi^{(\ell)}_{(\boldsymbol{k})}}{\Delta t} &= \frac{1}{4}(e^{Ht^{(\ell+1)}} + e^{Ht^{(\ell)}})\delta^{ij}\widehat{\delta}^{\langle 1\rangle}_i\widehat{\delta}^{\langle 1\rangle}_j(\phi^{(\ell+1)}_{(\boldsymbol{k})} + \phi^{(\ell)}_{(\boldsymbol{k})}) \\ &\quad - \frac{m^2}{4}(e^{3Ht^{(\ell+1)}} + e^{3Ht^{(\ell)}})(\phi^{(\ell+1)}_{(\boldsymbol{k})} + \phi^{(\ell)}_{(\boldsymbol{k})}) \\ &\quad - \frac{\lambda}{2(p+1)}(e^{3Ht^{(\ell+1)}} + e^{3Ht^{(\ell)}})\frac{|\phi^{(\ell+1)}_{(\boldsymbol{k})}|^{p+1} - |\phi^{(\ell)}_{(\boldsymbol{k})}|^{p+1}}{\phi^{(\ell+1)}_{(\boldsymbol{k})} - \phi^{(\ell)}_{(\boldsymbol{k})}},\end{aligned} \tag{10}$$

respectively. The upper index $^{(\ell)}$ in parentheses is the time index, and the lower index $_{(\boldsymbol{k})}$ in parentheses is the spatial grid index, where $\boldsymbol{k} = (k_1, k_2, k_3)$ and k_1, k_2, and k_3 are x, y, and z indices, respectively. $\widehat{\delta}^{\langle 1\rangle}_i$ is the discrete operator defined as

$$\widehat{\delta}^{\langle 1\rangle}_i u^{(\ell)}_{(\boldsymbol{k})} := \begin{cases} \dfrac{u^{(\ell)}_{(k_1+1,k_2,k_3)} - u^{(\ell)}_{(k_1-1,k_2,k_3)}}{2\Delta x}, & (i=1) \\ \dfrac{u^{(\ell)}_{(k_1,k_2+1,k_3)} - u^{(\ell)}_{(k_1,k_2-1,k_3)}}{2\Delta y}, & (i=2) \\ \dfrac{u^{(\ell)}_{(k_1,k_2,k_3+1)} - u^{(\ell)}_{(k_1,k_2,k_3-1)}}{2\Delta z}. & (i=3) \end{cases} \tag{11}$$

There are two features in Eq. (10). First, the second-order difference is expressed as $\widehat{\delta}^{\langle 1\rangle}_i\widehat{\delta}^{\langle 1\rangle}_j$. In general, the discrete operator of the second-order difference is usually

defined as

$$\widehat{\delta}_{ij}^{\langle 2\rangle} u_{(\boldsymbol{k})}^{(\ell)} := \begin{cases} \dfrac{u_{(k_1+1,k_2,k_3)}^{(\ell)} - 2u_{(\boldsymbol{k})}^{(\ell)} + u_{(k_1-1,k_2,k_3)}^{(\ell)}}{(\Delta x)^2}, & (i=j=1) \\ \dfrac{u_{(k_1,k_2+1,k_3)}^{(\ell)} - 2u_{(\boldsymbol{k})}^{(\ell)} + u_{(k_1,k_2-1,k_3)}^{(\ell)}}{(\Delta y)^2}, & (i=j=2) \\ \dfrac{u_{(k_1,k_2,k_3+1)}^{(\ell)} - 2u_{(\boldsymbol{k})}^{(\ell)} + u_{(k_1,k_2,k_3-1)}^{(\ell)}}{(\Delta z)^2}, & (i=j=3) \\ \widehat{\delta}_i^{\langle 1\rangle} \widehat{\delta}_j^{\langle 1\rangle} u_{(\boldsymbol{k})}^{(\ell)}. & (i \neq j) \end{cases} \tag{12}$$

In the case of $i = j$, $\widehat{\delta}_{ij}^{\langle 2\rangle} u_{(\boldsymbol{k})}^{(\ell)} \neq \widehat{\delta}_i^{\langle 1\rangle} \widehat{\delta}_j^{\langle 1\rangle} u_{(\boldsymbol{k})}^{(\ell)}$. Second, the expression of the nonlinear term, which is the last term on the right-hand side in Eq. (10), is not usual. In general, the discretized expression expected from Eq. (4) is $-\lambda e^{3Ht^{(\ell)}} |\phi_{(\boldsymbol{k})}^{(\ell)}|^{p-1} \phi_{(\boldsymbol{k})}^{(\ell)}$. These differences in the simulations are shown in Sec. 4.

The discretized total Hamiltonian $H_C^{(\ell)}$ is defined as

$$H_C^{(\ell)} := \sum_{\substack{1 \le k_1 \le n_1 \\ 1 \le k_2 \le n_2 \\ 1 \le k_3 \le n_3}} \mathcal{H}_{(\boldsymbol{k})}^{(\ell)} \Delta x \Delta y \Delta z, \tag{13}$$

where n_1, n_2, and n_3 are the grid numbers for x, y, and z, respectively. The difference quotient for $H_C^{(\ell)}$ using Eqs. (9) and (10) is calculated as

$$\begin{aligned} &\frac{H_C^{(\ell+1)} - H_C^{(\ell)}}{\Delta t} \\ &= H \sum_{\substack{1 \le k_1 \le n_1 \\ 1 \le k_2 \le n_2 \\ 1 \le k_3 \le n_3}} \Bigg[-\frac{3}{4} \{ e^{-3Ht^{(\ell+1)}} (\psi_{(\boldsymbol{k})}^{(\ell+1)})^2 + e^{-3Ht^{(\ell)}} (\psi_{(\boldsymbol{k})}^{(\ell)})^2 \} \\ &\quad + \frac{1}{4} \delta^{ij} \{ e^{Ht^{(\ell+1)}} (\hat{\delta}_i^{\langle 1\rangle} \phi_{(\boldsymbol{k})}^{(\ell+1)}) (\hat{\delta}_j^{\langle 1\rangle} \phi_{(\boldsymbol{k})}^{(\ell+1)}) + e^{Ht^{(\ell)}} (\hat{\delta}_i^{\langle 1\rangle} \phi_{(\boldsymbol{k})}^{(\ell)}) (\hat{\delta}_j^{\langle 1\rangle} \phi_{(\boldsymbol{k})}^{(\ell)}) \} \\ &\quad + \frac{3}{4} m^2 \{ e^{3Ht^{(\ell+1)}} (\phi_{(\boldsymbol{k})}^{(\ell+1)})^2 + e^{3Ht^{(\ell)}} (\phi_{(\boldsymbol{k})}^{(\ell)})^2 \} \\ &\quad + \frac{3\lambda}{2(p+1)} (e^{3Ht^{(\ell+1)}} |\phi_{(\boldsymbol{k})}^{(\ell+1)}|^{p+1} + e^{3Ht^{(\ell)}} |\phi_{(\boldsymbol{k})}^{(\ell)}|^{p+1}) \Bigg] \\ &\quad + [\text{Boundary Terms}] + O(\Delta t), \end{aligned} \tag{14}$$

where we use the relation such that

$$e^{at^{(\ell+1)}} = e^{at^{(\ell)}} + a e^{at^{(\ell)}} \Delta t + O((\Delta t)^2). \quad (\forall a \in \mathbb{R}) \tag{15}$$

The boundary terms in Eq. (14) are eliminated under the periodic boundary condition. In addition, if $H = 0$, then $H_C^{(\ell+1)}$ is consistent with $H_C^{(0)}$ in the order of Δt. Then we define the discretized modified total Hamiltonian $\tilde{H}_C^{(\ell)}$ as

$$\begin{aligned}\tilde{H}_C^{(\ell)} := H_C^{(\ell)} - H \sum_{0\le m\le \ell-1} \sum_{\substack{1\le k_1\le n_1\\ 1\le k_2\le n_2\\ 1\le k_3\le n_3}} \Bigg[& -\frac{3}{4}\{e^{-3Ht^{(m+1)}}(\psi_{(\boldsymbol{k})}^{(m+1)})^2 + e^{-3Ht^{(m)}}(\psi_{(\boldsymbol{k})}^{(m)})^2\} \\ & + \frac{1}{4}\delta^{ij}\{e^{Ht^{(m+1)}}(\hat{\delta}_i^{\langle 1\rangle}\phi_{(\boldsymbol{k})}^{(m+1)})(\hat{\delta}_j^{\langle 1\rangle}\phi_{(\boldsymbol{k})}^{(m+1)}) + e^{Ht^{(m)}}(\hat{\delta}_i^{\langle 1\rangle}\phi_{(\boldsymbol{k})}^{(m)})(\hat{\delta}_j^{\langle 1\rangle}\phi_{(\boldsymbol{k})}^{(m)})\} \\ & + \frac{3}{4}m^2\{e^{3Ht^{(m+1)}}(\phi_{(\boldsymbol{k})}^{(m+1)})^2 + e^{3Ht^{(m)}}(\phi_{(\boldsymbol{k})}^{(m)})^2\} \\ & + \frac{3\lambda}{2(p+1)}(e^{3Ht^{(m+1)}}|\phi_{(\boldsymbol{k})}^{(m+1)}|^{p+1} + e^{3Ht^{(m)}}|\phi_{(\boldsymbol{k})}^{(m)}|^{p+1})\Bigg]\Delta t\Delta x\Delta y\Delta z. \quad (16)\end{aligned}$$

We adopt this value as a criterion of the accuracy of the simulations in the case of $H \neq 0$.

4 Numerical Simulations

In this section, we perform some simulations with SPS to investigate their stability and accuracy. We perform simulations with three formulations of the discretized semilinear Klein–Gordon equation in the de Sitter spacetime. The first formulation is that for Eqs. (9), (10), and (13). We call this formulation Form I. As shown in Eq. (14), Form I is SPS. The details are shown in [9]. The second formulation is that for Eqs. (9), (13), and the following Eq. (17).

$$\begin{aligned}\frac{\psi_{(\boldsymbol{k})}^{(\ell+1)} - \psi_{(\boldsymbol{k})}^{(\ell)}}{\Delta t} = & \frac{1}{4}(e^{Ht^{(\ell+1)}} + e^{Ht^{(\ell)}})\delta^{ij}\widehat{\delta}_{ij}^{\langle 2\rangle}(\phi_{(\boldsymbol{k})}^{(\ell+1)} + \phi_{(\boldsymbol{k})}^{(\ell)}) \\ & - \frac{m^2}{4}(e^{3Ht^{(\ell+1)}} + e^{3Ht^{(\ell)}})(\phi_{(\boldsymbol{k})}^{(\ell+1)} + \phi_{(\boldsymbol{k})}^{(\ell)}) \\ & - \frac{\lambda}{2(p+1)}(e^{3Ht^{(\ell+1)}} + e^{3Ht^{(\ell)}})\frac{|\phi_{(\boldsymbol{k})}^{(\ell+1)}|^{p+1} - |\phi_{(\boldsymbol{k})}^{(\ell)}|^{p+1}}{\phi_{(\boldsymbol{k})}^{(\ell+1)} - \phi_{(\boldsymbol{k})}^{(\ell)}}\end{aligned} \quad (17)$$

We call this formulation Form II. The difference between Eqs. (10) and (17) is the second-order difference term. The third formulation is that for Eqs. (9), (13), and

the following Eq. (18).

$$\begin{aligned}\frac{\psi_{(\boldsymbol{k})}^{(\ell+1)}-\psi_{(\boldsymbol{k})}^{(\ell)}}{\Delta t} &= \frac{1}{4}(e^{Ht^{(\ell+1)}}+e^{Ht^{(\ell)}})\delta^{ij}\widehat{\delta}_i^{\langle 1\rangle}\widehat{\delta}_j^{\langle 1\rangle}(\phi_{(\boldsymbol{k})}^{(\ell+1)}+\phi_{(\boldsymbol{k})}^{(\ell)}) \\ &\quad -\frac{m^2}{4}(e^{3Ht^{(\ell+1)}}+e^{3Ht^{(\ell)}})(\phi_{(\boldsymbol{k})}^{(\ell+1)}+\phi_{(\boldsymbol{k})}^{(\ell)}) \\ &\quad -\frac{\lambda}{8}(e^{3Ht^{(\ell+1)}}+e^{3Ht^{(\ell)}})|\phi_{(\boldsymbol{k})}^{(\ell+1)}+\phi_{(\boldsymbol{k})}^{(\ell)}|^{p-1}(\phi_{(\boldsymbol{k})}^{(\ell+1)}+\phi_{(\boldsymbol{k})}^{(\ell)})\end{aligned} \tag{18}$$

We call this formulation Form III. The difference between Eqs. (10) and (18) is the expression of the discretized nonlinear term, which is the last term on the right-hand side of each of these equations.

The simulation settings are as follows.

- Initial conditions: $\phi_0 = A\cos(2\pi x)$, $\psi_0 = 2\pi A\sin(2\pi x)$, and $A = 4$
- Numerical domains: $0 \le x \le 1$, $0 \le t \le 1000$
- Boundary condition: periodic
- Grids: $\Delta x = 1/200$ and $\Delta t = 1/1000$
- Mass: $m = 1$
- Boolean parameter of the nonlinear term: $\lambda = 1$
- Number of exponents in the nonlinear term: $p = 2, 3, 4, 5$, and 6
- Hubble constant: $H = 0$ and 10^{-3}

Forms I, II, and III are expressed in three dimensions. On the other hand, the initial conditions are one-dimensional. Even if the spatial dimension of the initial conditions is one-dimensional, the differences exist in the second-order difference term and the discretized nonlinear term. Thus, the numerical simulations are expected to show differences in the one-dimensional initial conditions.

4.1 Flat Spacetime

We perform some simulations of the three formulations in the flat spacetime, which is in the case of $H = 0$. In Fig. 1, we show the relative errors of the total Hamiltonian H_C against the initial values $H_C(0)$ for each value of the exponent p in the nonlinear term. The left panel is drawn with Form I, the center panel with Form II, and the right panel with Form III. The values of $|(H_C - H_C(0))/H_C(0)|$ indicate the numerical errors because H_C is a constraint. In the right panel, we see that the value of $p = 2$ with Form III is smaller than those of the other exponents in the panel. This result indicates that the numerical errors caused by the nonlinear term are small in the case of $p = 2$. We see that the values of the center and right panels are larger than that of

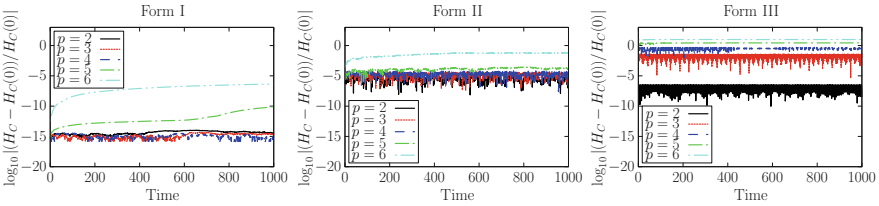


Fig. 1 Relative errors of the total Hamiltonian H_C against the initial value $H_C(0)$ for each p in the case of $H = 0$. The horizontal axis is time, and the vertical axis is $\log_{10}|(H_C - H_C(0))/H_C(0)|$. The left panel is drawn with Form I, the center panel with Form II, and the right panel with Form III

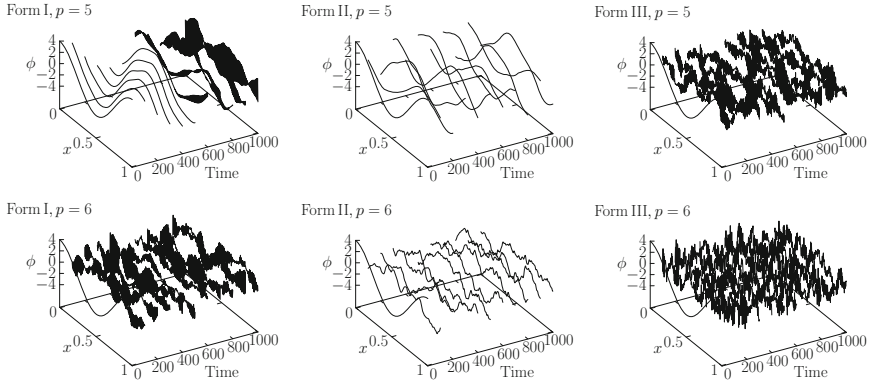


Fig. 2 ϕ with $p = 5$ and 6. The left panels are drawn with Form I, the center panels with Form II, and the right panels with Form III. The top panels are drawn for $p = 5$ and the bottom panels for $p = 6$. The vibrations occur at $t \geq 700$ in the top-left panel, $t \geq 100$ in the bottom-left panel, and $t \geq 100$ in the right panels

the left panel for each p. Thus, the simulations with Form I are more accurate than those with the other forms.

Then we show ϕ with $p = 5$ and 6 in Fig. 2 to investigate the stability of the simulations. The left panels are drawn with Form I, the center panels with Form II, and the right panels with Form III. The top panels are drawn with the exponent $p = 5$ and the bottom panels with $p = 6$. We see that the simulations of the top-left panel at $t \geq 700$, the bottom-left panel at $t \geq 100$, and the right panels at $t \geq 100$ are unstable because of the generated vibrations. On the other hand, the simulations shown in the center panels are stable until $t = 1000$. Thus, we determine that the simulations with Form II are more stable than those with the other formulations.

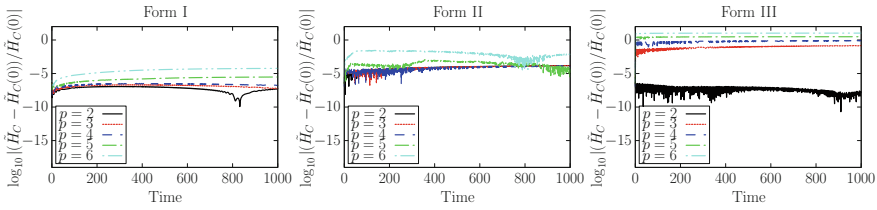


Fig. 3 Relative errors of the modified total Hamiltonian $\tilde{H}_C$ against the initial value $\tilde{H}_C(0)$ for each p in the case of $H = 10^{-3}$. The horizontal axis is time, and the vertical axis is $\log_{10}|(\tilde{H}_C - \tilde{H}_C(0))/\tilde{H}_C(0)|$. The left panel is drawn with Form I, the center panel with Form II, and the right panel with Form III

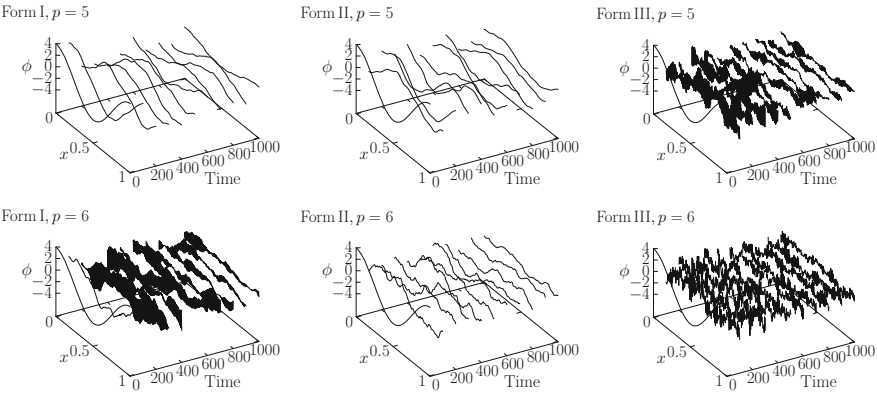


Fig. 4 The same as in Fig. 2 except for the value of the Hubble constant, which is 10^{-3}

4.2 Curved Spacetime

Here, we perform some simulations with the same settings as in Sec. 4.1 except for the Hubble constant. This time, we set the Hubble constant $H = 10^{-3}$.

We show the relative errors of the modified total Hamiltonian $\tilde{H}_C$ against the initial value $\tilde{H}_C(0)$ in Fig. 3. Note that $\tilde{H}_C$ is calculated approximately using Eq. (16) via the numerical solutions in time evolution. The left panel is drawn with Form I, the center panel with Form II, and the right panel with Form III. We see that the value of $p = 2$ with Form III is smaller than those in the other cases in the right panel. This tendency is consistent with the case of $H = 0$.

Figure 4 is the same as Fig. 2 except for the value of the Hubble constant. In the comparison between Figs. 2 and 4, no vibrations appear in the top-left panel in Fig. 4 and also in the bottom-left panel in Fig. 4 until $t = 100$. The other patterns of behavior are almost the same. These results indicate that the vibrations of the waveform of the solutions decrease in comparison with the case of $H = 0$. That is, the positive Hubble constant makes the simulation stable. This is also noted in [9].

5 Summary

We investigated the factors affecting the stability and accuracy of simulations of the semilinear Klein–Gordon equation in the de Sitter spacetime using SPS. We reviewed the canonical formulation of the equation and that of the discretized equation with SPS. To investigate the terms affecting the stability and accuracy in the discretized equations, we compared some simulations using three discretized formulations. The first formulation consists of the discretized equations with SPS, which is called Form I. This formulation was reported in [9]. The second formulation consists of the discretized equations with SPS, in which the second-order difference was replaced with a standard discretized second-order difference, which is called Form II. The third formulation consists of the discretized equations with SPS, in which the nonlinear term was replaced with a standard discretized term, which is called Form III. We monitored the total Hamiltonian or the modified one to see the accuracy of the simulations. As a result, we found that the stability and accuracy of the simulations using Form III are worse than those with Form I. This result indicates that the discretizations of the nonlinear term affect on the stability and accuracy of the simulations. In addition, the accuracy of the simulations with Form I is better than those with the other forms. On the other hand, the stability of the simulations with Form II is higher than those with the other forms. Moreover, we confirmed that the simulations with positive values of the Hubble constant are more stable than those in the flat spacetime.

The numerical stability of the simulations using Form I is lower than those using Form II. However, there are degrees of freedom in the selection of the discretized terms for Form I. Therefore, it seems that the formulation that enables stable and accurate numerical simulation can be constructed, which we will report in the near future.

Acknowledgments The authors thank the anonymous referees for their many helpful comments that improved the paper. T.T. and M.N. were partially supported by JSPS KAKENHI Grant Number 21K03354. T.T. was partially supported by JSPS KAKENHI Grant Number 20K03740 and Grant for Basic Science Research Projects from The Sumitomo Foundation. M.N. was partially supported by JSPS KAKENHI Grant Number 16H03940.

References

1. Furihata, D.: Finite difference schemes for $\frac{\partial u}{\partial t} = \left(\frac{\partial}{\partial x}\right)^{\alpha} \frac{\delta G}{\delta u}$ that inherit energy conservation or dissipation property. J. Comput. Phys. **156**, 181–205 (1999)
2. Furihata, D., Matsuo, T.: Discrete Variational Derivative Method. CRC Press/Taylor & Francis, London (2010)
3. Yagdjian, K., Galstian, A.: Fundamental solutions for the Klein–Gordon equation in de Sitter spacetime. Commun. Math. Phys. **285**(1), 293–344 (2009)
4. Yagdjian, K.: The semilinear Klein–Gordon equation in de Sitter spacetime. Discrete Contin. Dyn. Syst. Ser. S **2**(3), 679–696 (2009)

5. Yagdjian, K.: Global solutions of semilinear system of Klein–Gordon equations in de Sitter spacetime. In: Progress in Partial Differential Equations. Proceedings in Mathematics & Statistics, vol. 44, pp. 409–444. Springer, Berlin (2013)
6. Nakamura, M.: The Cauchy problem for semi-linear Klein–Gordon equations in de Sitter spacetime. J. Math. Anal. Appl. **410**(1), 445–454 (2014)
7. Nakamura, M.: The Cauchy problem for the Klein–Gordon equation under the quartic potential in the de Sitter spacetime. J. Math. Phys. **62**, 121509 (2021)
8. Yazici, M., Şengül, S.: Approximate solutions to the nonlinear Klein-Gordon equation in de Sitter spacetime. Open Phys. **14**(1), 314–320 (2016)
9. Tsuchiya, T., Nakamura, M.: On the numerical experiments of the Cauchy problem for semi-linear Klein-Gordon equations in the de Sitter spacetime. J. Comput. Appl. Math. **361**, 396–412 (2019)

Part X
Recent Progress in Evolution Equations

Global Small Data Solutions for an Evolution Equation with Structural Damping and Hartree-Type Nonlinearity

Marcello D'Abbicco

Abstract In this paper, we consider an evolution equation with structural damping and nonlocal nonlinearity of Hartree type, and we prove the existence of global small data solutions for supercritical powers.

1 Introduction

We consider a structurally damped evolution equation with a nonlocal nonlinearity

$$\begin{cases} u_{tt} + (-\Delta)^{\sigma} u + (-\Delta)^{\frac{\sigma}{2}} u_t = c\left(|x|^{-(n-\alpha)} * f(u)\right) g(u), \ t > 0, \ x \in \mathbb{R}^n, \\ u(0,x) = u_0(x), \\ u_t(0,x) = u_1(x), \end{cases} \tag{1}$$

where $c \neq 0$, $\alpha \in (0,n)$ and

$$|f(u) - f(v)| \leq C\,|u-v|\left(|u|^{p-1} + |v|^{p-1}\right), \quad \text{for some } p \geq 1, \tag{2}$$

$$|g(u) - g(v)| \leq C\,|u-v|\left(|u|^{q-1} + |v|^{q-1}\right), \quad \text{for some } q \geq 1. \tag{3}$$

A nonlinearity as in (1) generalizes Hartree-type nonlinearities [13–15] which appear in several physical models. We address the interested reader to [1, 3, 16] and the references therein.

We stress that, with no loss of generality, we may fix

$$c = c_{n,\alpha} = \pi^{\frac{n}{2}} 2^{\alpha} \frac{\Gamma(\alpha/2)}{\Gamma((n-\alpha)/2)},$$

M. D'Abbicco (✉)
University of Bari, Department of Mathematics, Bari, Italy
e-mail: marcello.dabbicco@uniba.it

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_43

in (1), so that

$$c_{n,\alpha}|x|^{-(n-\alpha)} * f = I_\alpha f$$

where I_α denotes the Riesz potential applied to f. The equation in (1) is then obtained by the following system:

$$\begin{cases} u_{tt} + (-\Delta)^\sigma u + (-\Delta)^{\frac{\sigma}{2}} u_t = U\, g(u), \\ (-\Delta)^{\frac{\alpha}{2}} U = f(u)\,. \end{cases} \tag{4}$$

By the Hardy-Littlewood-Sobolev theorem, for any $f \in L^{r^*}$, with $r^* \in (1, n/\alpha)$, it holds

$$I_\alpha f \in L^r, \quad \|I_\alpha f\|_{L^r} \leq C\, \|f\|_{L^{r^*}}, \quad \frac{1}{r^*} = \frac{1}{r} + \frac{\alpha}{n}. \tag{5}$$

In this paper, we prove that global-in-time solutions to (1) exist, for sufficiently small data in a suitable space, if

$$p + q > 1 + \frac{2\sigma + \alpha}{n - \sigma}. \tag{6}$$

It remains open to determine if the existence exponent $1+(2\sigma+\alpha)/(n-\sigma)$ in (6) is *critical,* that is, global-in-time solutions in general do not exist for subcritical powers. In [12], R. Filippucci and M. Gherghu studied the nonexistence exponent for quasilinear parabolic inequalities with a nonlinearity of type $(K * u^p)u^q$. In particular, for the parabolic m-Laplacian equation

$$u_t - \Delta_m u = (|x|^{-(n-\alpha)} * u^p)u^q,$$

where $\Delta_m u = \nabla \cdot (|\nabla u|^{m-2}\nabla u)$, with $0 < n - \alpha < m/2$, they proved the nonexistence of nonnegative nontrivial smooth solutions for

$$p + q < m - 1 + \frac{m + \alpha}{2n - \alpha}.$$

The critical exponent for structurally damped evolution equations with power nonlinearities

$$\begin{cases} u_{tt} + (-\Delta)^{\sigma} u + (-\Delta)^{\theta} u_t = g(u), \ t > 0, \ x \in \mathbb{R}^n, \\ u(0,x) = u_0(x), \\ u_t(0,x) = u_1(x), \end{cases} \tag{7}$$

has been well-investigated recently, and it has been highlighted how it depends on the strength of damping. If the damping is effective, i.e., $0 < 2\theta < \sigma$, the critical exponent is $1 + 2\sigma/(n - 2\theta)$, the same of the corresponding parabolic equation (see [6], see also [2] and the references therein), as it happens for the classical damped wave equation [18] (see also [11] for a deeper analysis of critical nonlinearities of type $g(u)$), i.e., $\theta = 0$ and $\sigma = 1$. In the noneffective case $\theta \in (\sigma/2, \sigma]$, oscillations come into play [9, 17] in the asymptotic profile of the solution, and it has been recently shown [7] that the critical exponent becomes $1+2\sigma/(n-\sigma)$, the same of the undamped evolution equation [10], that is, the same in (6) when $\alpha = 0$.

At the threshold of effectiveness, i.e., $\theta = \sigma/2$, the equation in (7) has the simplest structure, since oscillations are very slow, and the critical exponent is $1 + 2\sigma/(n - \sigma)$ as in the case of noneffective damping. If the power nonlinearity is replaced by a nonlocal-in-time power nonlinearity as

$$g(t,u) = \int_0^t (t-s)^{-(1-\alpha)} |u(s,x)|^q \, ds,$$

with $\alpha \in (0, 1)$, then the critical exponent becomes (see [4], see also [8])

$$\max\{\tilde{q}, (1-\alpha)^{-1}\}, \quad \tilde{q} = 1 + \frac{(2+\alpha)\sigma}{n - (1+\alpha)\sigma}.$$

In this paper, we show that the influence from a nonlocal-in-space nonlinearity is quite different.

Solutions in $C([0,\infty), L^{p_0})$
In our first result, for any given supercritical value $p + q$, we look for the less restrictive assumption on the regularity of initial data, such that we may construct a global-in-time solution to (1) in $C([0,\infty), L^{p_0})$, for some p_0.

Theorem 1 *Let $n \geq 1$ and $\sigma \in (0, n)$. Assume that p, q in (2) and (3) verify the following:*

$$\frac{\alpha}{n}\, p < q < \frac{n}{\alpha}\, p, \tag{8}$$

and that (6) holds. Moreover, if $n > 2\sigma$ we also assume that

$$p + q < 1 + \frac{2\sigma + \alpha}{n - 2\sigma}. \tag{9}$$

Fix

$$p_0 = \frac{n}{n+\alpha}\,(p+q), \quad \frac{1}{p_0^*} = \frac{1}{p_0} + \frac{\sigma}{n}.$$

Then there exists $\varepsilon > 0$ such that for any

$$(u_0, u_1) \in \mathcal{A} = \left(L^{m_0} \cap L^{p_0}\right) \times \left(L^{m_1} \cap L^{p_0^*}\right), \quad \|(u_0, u_1)\|_{\mathcal{A}} \leq \varepsilon, \tag{10}$$

where $1 \leq m_1 \leq m_0$ verify

$$p + q \geq 1 + \frac{2\sigma + \alpha}{(n/m_1) - \sigma}, \quad \frac{n}{\sigma}\left(\frac{1}{m_1} - \frac{1}{m_0}\right) \leq 1, \tag{11}$$

there is a unique global-in-time solution $u \in C([0, \infty), L^{m_0} \cap L^{p_0})$. Moreover, the following long-time decay estimate holds

$$\|u(t, \cdot)\|_{L^{p_0}} \leq C\,(1+t)^{1 - \frac{n}{\sigma}\left(\frac{1}{m_1} - \frac{1}{p_0}\right)}\, \|(u_0, u_1)\|_{\mathcal{A}}, \tag{12}$$

and the following estimate holds

$$\|u(t, \cdot)\|_{L^{m_0}} \leq C\,(1+t)^{1 - \frac{n}{\sigma}\left(\frac{1}{m_1} - \frac{1}{m_0}\right)}\, \|(u_0, u_1)\|_{\mathcal{A}}. \tag{13}$$

Remark 1 We stress that condition (11) follows as a consequence of (6) if $m_1 = 1$. In particular, in this case, (12) reads as

$$\|u(t, \cdot)\|_{L^{p_0}} \leq C\,(1+t)^{1 - \frac{n}{\sigma}\left(1 - \frac{1}{p_0}\right)}\, \|(u_0, u_1)\|_{\mathcal{A}}, \tag{14}$$

Moreover, if $m_0 = 1$, then (13) reads as:

$$\|u(t, \cdot)\|_{L^1} \leq C\,(1+t)\, \|(u_0, u_1)\|_{\mathcal{A}}. \tag{15}$$

Remark 2 We stress that $m_1 < p_0^*$ in (10), as a consequence of the fact that the exponent in the estimate in (12) is negative due to (11), that is,

$$\frac{1}{m_1} > \frac{1}{p_0} + \frac{\sigma}{n} = \frac{1}{p_0^*}.$$

An analogous reasoning shows that $m_0 < p_0$ in (10).

Solutions in $C([0,\infty), L^1 \cap L^\infty)$
We may remove assumptions (8) and (9) by strengthening the regularity assumption on the initial data. In particular, we may easily construct solutions in $L^1 \cap L^\infty$, proceeding as in [5].

Theorem 2 *Let $n \geq 1$ and $\sigma \in (0, n)$. Assume that p, q in (2) and (3) verifies (6). Then there exists $\varepsilon > 0$ such that for any*

$$(u_0, u_1) \in \mathcal{A} = \left(L^1 \cap L^\infty\right) \times \left(L^1 \cap L^{\frac{n}{\sigma}}\right), \quad \|(u_0, u_1)\|_{\mathcal{A}} \leq \varepsilon, \tag{16}$$

there is a unique global-in-time solution $u \in C([0,\infty), L^1 \cap L^\infty)$. Moreover, the following long-time decay estimate holds

$$\|u(t,\cdot)\|_{L^\infty} \leq C\,(1+t)^{1-\frac{n}{\sigma}}\,\|(u_0, u_1)\|_{\mathcal{A}}, \tag{17}$$

and estimate (15) holds.

Is the Exponent Critical?
Following the same approach in this paper, one may clearly study a large class of evolution equations with Hartree-type nonlinearity $c\left(|x|^{-(n-\alpha)} * f(u)\right)g(u)$. For instance, the existence exponent for the heat equation with that nonlinearity is obtained by the solution of the following equation:

$$1 = \frac{n}{2}\left(1 - \frac{1}{p_0}\right)(p+q) = \frac{n}{2}(p+q-1) - \frac{\alpha}{2}.$$

That is, the existence exponent is a shifted Fujita exponent corresponding to the equation

$$n(p+q-1) = 2 + \alpha.$$

(continued)

We stress that a gap remains open, comparing this result with the nonexistence exponent obtained in [12]. The question naturally arising is then: "Is this shifted Fujita exponent critical?"

The case $q = p - 1$
A model case is when $f(u) = |u|^p$ and $g(u) = |u|^{p-2}u$, that is, $q = p-1$ for some $p \geq 2$. In this case, condition (8) holds if, and only if, $p > n/(n-\alpha)$. Condition (6) reads as

$$p > 1 + \frac{\sigma + \alpha/2}{n - \sigma},$$

and $p_0 = (2p-1)n/(n+\alpha)$. In particular, $p = 2$ in the classical Hartree type inequality, so that our result applies for $\alpha < 2(n-2\sigma)$ if $n < 4\sigma$, and for any $\alpha \in (0, n)$ if $n \geq 4\sigma$.

2 Proof of Theorem 1

In order to prove our results, we consider the linear problem

$$\begin{cases} u_{tt} + (-\Delta)^{\sigma} u + (-\Delta)^{\frac{\sigma}{2}} u_t = 0, \ t > 0, \ x \in \mathbb{R}^n, \\ u(0,x) = u_0(x), \\ u_t(0,x) = u_1(x). \end{cases} \tag{18}$$

We address the reader to [6] and the references therein for the proof of the following estimates for the solution to (18):

$$\|u(t,\cdot)\|_{L^{q_2}} \leq C_0\, t^{-\frac{n}{\sigma}\left(\frac{1}{q_0}-\frac{1}{q_2}\right)} \|u_0\|_{L^{q_0}} + C_1\, t^{1-\frac{n}{\sigma}\left(\frac{1}{q_1}-\frac{1}{q_2}\right)} \|u_1\|_{L^{q_1}}, \tag{19}$$

where $q_2 \in [1,\infty]$ and $q_0, q_1 \in [1, q_2]$, with C_j independent of $t > 0$ and $\|u_j\|_{L^{q_j}}$, $j = 0, 1$.

Lemma 1 *Let p_0, m_0, m_1 as in the statement of Theorem 1. Then the solution to (18) verifies the long-time decay estimate (12) and the estimate (13).*

Proof For $t \leq 1$, we apply (19) with $q_2 = q_0 = p_0$ and $q_1 = p_0^*$, so that we obtain

$$\|u(t,\cdot)\|_{L^{p_0}} \leq C_0 \|u_0\|_{L^{p_0}} + C_1 \|u_1\|_{L^{p_0^*}},$$

whereas, for $t \geq 1$, we apply (19) with $q_2 = p_0$, $q_0 = m_0$ and $q_1 = m_1$, so that we obtain

$$\|u(t,\cdot)\|_{L^{p_0}} \leq t^{1-\frac{n}{\sigma}\left(\frac{1}{m_1}-\frac{1}{p_0}\right)} \left(C_0 \|u_0\|_{L^{m_0}} + C_1 \|u_1\|_{L^{m_1}}\right),$$

thanks to the second half of (11). Summing the previous two estimates, we derive (12). On the other hand, if we apply (19) with $q_2 = q_0 = m_0$ and $q_1 = m_1$, we obtain

$$\|u(t,\cdot)\|_{L^{m_0}} \leq C_0 \|u_0\|_{L^{m_0}} + C_1\, t^{1-\frac{n}{\sigma}\left(\frac{1}{m_1}-\frac{1}{m_0}\right)} \|u_1\|_{L^{m_1}},$$

from which we get (13), thanks to the second half of (11). □

The Interplay Between Integrability and Desired Decay
In Lemma 1, the assumption of $u_0 \in L^{p_0}$ and $u_1 \in L^{p_0^*}$ comes into play to guarantee that we may find a solution with $u(t,\cdot) \in L^{p_0}$ for any $t \geq 0$, whereas the assumption $u_0 \in L^{m_0}$ and $u_1 \in L^{m_1}$ provides the desired decay rate as $t \to \infty$ for the solution. The assumption that $u(t,\cdot) \in L^{p_0}$ and that $\|u(t,\cdot)\|_{L^{p_0}}$ has a certain decay rate, will be crucial to treat the nonlinear problem. On the other hand, as a bonus consequence, the previous assumption that $u_0 \in L^{m_0}$ and $u_1 \in L^{m_1}$ guarantees "for free" that also $u(t,\cdot) \in L^{m_0}$ for any $t \geq 0$.

We define the solution space

$$X = u \in C([0,\infty), L^{m_0} \cap L^{p_0}),$$

equipped with the norm

$$\|u\|_X = \sup_{t\in[0,\infty)} \Big((1+t)^{-1+\frac{n}{\sigma}\left(\frac{1}{m_1}-\frac{1}{p_0}\right)} \|u(t,\cdot)\|_{L^{p_0}} + (1+t)^{-1+\frac{n}{\sigma}\left(\frac{1}{m_1}-\frac{1}{m_0}\right)} \|u(t,\cdot)\|_{L^{m_0}}\Big).$$

In particular, for any $u \in X$, the following estimate holds

$$\|u(t,\cdot)\|_{L^{p_0}} \leq (1+t)^{1-\frac{n}{\sigma}\left(\frac{1}{m_1}-\frac{1}{p_0}\right)} \|u\|_X. \tag{20}$$

As a consequence of Lemma 1, the solution u^{lin} to (18) is in X and

$$\|u^{\mathrm{lin}}\|_X \leq C_1 \|(u_0,u_1)\|_{\mathcal{A}}. \tag{21}$$

We now consider the operator

$$N : X \to X, \qquad Nu = \int_0^t K(t-s,\cdot)(I_\alpha f(u(s,\cdot)))g(u(s,\cdot))\,ds,$$

where $K(t,\cdot)$ is the fundamental solution to (18), that is, the solution with $u_0 = 0$ and $u_1 = \delta$.

By Duhamel's principle, a function $u \in C([0,\infty), L^{m_0} \cap L^{p_0})$ is the solution to (1) if, and only if,

$$u(t,x) = u^{\mathrm{lin}}(t,x) + Nu(t,x), \quad \text{for any } t \geq 0 \text{ and for a.e. } x, \tag{22}$$

where u^{lin} is the solution to (18).

Lemma 2 *Let $u, v \in X$. Then*

$$\|Nu - Nv\|_X \leq C\,\|u-v\|_X\,(\|u\|_X^{p+q-1} + \|v\|_X^{p+q-1}). \tag{23}$$

Proof Let $r \in (n/(n-\alpha), \infty)$ and let $r' = r/(r-1)$, its Hölder conjugate. Also, $p' = p/(p-1)$ and $q' = q/(q-1)$. Then, using Hölder inequality, (2), (3) and (5), we may estimate the L^1 norm of

$$\begin{aligned} h(u,v) &= (I_\alpha f(u))\,g(u) - (I_\alpha f(v))\,g(v) \\ &= (I_\alpha(f(u)-f(v))\,g(u) + (I_\alpha f(v))\,(g(u)-g(v)) \end{aligned}$$

as follows:

$$\begin{aligned} \|h(u,v)\|_{L^1} &\leq \|I_\alpha(f(u)-f(v))\|_{L^r}\,\|g(u)\|_{L^{r'}} \\ &\quad + \|I_\alpha f(v)\|_{L^r}\,\|g(u)-g(v)\|_{L^{r'}} \\ &\leq C_1\,\|f(u)-f(v)\|_{L^{r^*}}\,\|g(u)\|_{L^{r'}} \\ &\quad + C_1\,\|f(v)\|_{L^{r^*}}\,\|g(u)-g(v)\|_{L^{r'}} \\ &\leq C_2\,\|u-v\|_{L^{r^*p}}\,\||u|^{p-1}+|v|^{p-1}\|_{L^{r^*p'}}\,\||u|^q\|_{L^{r'}} \\ &\quad + C_2\,\||v|^p\|_{L^{r^*}}\,\|u-v\|_{L^{r'q}}\,\|(|u|^{q-1}+|v|^{q-1})\|_{L^{r'q'}} \\ &\leq C_3\,\|u-v\|_{L^{r^*p}}\left(\|u\|_{L^{r^*p}}^{p-1} + \|v\|_{L^{r^*p}}^{p-1}\right)\|u\|_{L^{r'q}}^q \\ &\quad + C_3\,\|v\|_{L^{r^*p}}^p\,\|u-v\|_{L^{r'q}}\left(\|u\|_{L^{r'q}}^{q-1} + \|v\|_{L^{r'q}}^{q-1}\right). \end{aligned}$$

We fix

$$r = \frac{n}{np-\alpha q}\,(p+q),$$

so that

$$r^* p = r' q = p_0.$$

We notice that $r \in (n/(n-\alpha), \infty)$, thanks to (8).

Now, using that $u, v \in X$, by (20), replacing $p_0 = n(p+q)/(n+\alpha)$, we obtain the estimate

$$\begin{aligned}
&\|h(u,v)(t,\cdot)\|_{L^1} \\
&\quad \le C\,\|u-v\|_X \left(\|u\|_X^{p+q-1} + \|v\|_X^{p+q-1}\right)(1+t)^{p+q-\frac{n(p+q)}{\sigma}\left(\frac{1}{m_1}-\frac{1}{p_0}\right)} \\
&\quad = C\,\|u-v\|_X \left(\|u\|_X^{p+q-1} + \|v\|_X^{p+q-1}\right)(1+t)^{(p+q)\left(1-\frac{n}{\sigma m_1}\right)+\frac{n+\alpha}{\sigma}},
\end{aligned}$$

for any $t \ge 0$. Using (19) with $u_0 = 0$, $q_2 = p_0$ and $q_1 = 1$, we get

$$\begin{aligned}
\|(Nu - Nv)(t,\cdot)\|_{L^{p_0}} &\le C \int_0^t (t-s)^{1-\frac{n}{\sigma}\left(1-\frac{1}{p_0}\right)} \|h(u,v)(s,\cdot)\|_{L^1}\, ds \\
&\le C\,\|u-v\|_X \left(\|u\|_X^{p+q-1} + \|v\|_X^{p+q-1}\right) I(t),
\end{aligned}$$

where

$$I(t) = \int_0^t (t-s)^{1-\frac{n}{\sigma}\left(1-\frac{1}{p_0}\right)} (1+s)^{(p+q)\left(1-\frac{n}{\sigma m_1}\right)+\frac{n+\alpha}{\sigma}}\, ds\,.$$

We notice that, thanks to (9),

$$1 - \frac{n}{\sigma}\left(1 - \frac{1}{p_0}\right) > -1.$$

We now distinguish two cases. If $m_1 = 1$, then, using (6) we may estimate

$$(p+q)\left(1 - \frac{n}{\sigma}\right) + \frac{n+\alpha}{\sigma} < -1.$$

Therefore (see, for instance, [7, Lemma 3.1]), we get

$$I(t) \approx (1+t)^{1-\frac{n}{\sigma}\left(1-\frac{1}{p_0}\right)}.$$

If $m_1 > 1$, thanks to the first half of (11), we may estimate

$$(p+q)\left(1-\frac{n}{\sigma m_1}\right)+\frac{n+\alpha}{\sigma} \leq -\frac{n+\sigma m_1+\alpha m_1}{\sigma m_1}+\frac{n+\alpha}{\sigma}$$
$$=\frac{n}{\sigma}\left(1-\frac{1}{m_1}\right)-1.$$

Therefore (see, for instance, [7, Lemma 3.1]), we get

$$I(t) \leq C(1+t)^{1-\frac{n}{\sigma}\left(\frac{1}{m_1}-\frac{1}{p_0}\right)}.$$

We proceed similarly to estimate $\|(Nu-Nv)(t,\cdot)\|_{L^{m_0}}$, and we conclude the proof. □

We may now conclude the proof of Theorem 1.

Proof ***(Theorem 1)*** We now define

$$R = 2C_1 \|(u_0,u_1)\|_{\mathcal{A}},$$

where C_1 is as in (21). For sufficiently small data, $2CR^{p+q-1} \leq 1/2$, where C is as in (23). Then, by (21) it follows that the operator $u^{\mathrm{lin}}(t,x)+N$ maps the ball $B_R = \{u : \|u\|_X \leq R\}$ in itself. Due to (23), it is a contraction. Therefore, there is a unique fixed point for $u^{\mathrm{lin}}(t,x)+F$ in B_R, that is, a unique solution to (22). Moreover, $\|u\|_X \leq 2C_1 \|(u_0,u_1)\|_{\mathcal{A}}$, that is, we get (12) and (13). This concludes the proof. □

3 Proof of Theorem 2

To prove Theorem 2, we follow the proof of Theorem 1, with some modifications.

Lemma 3 *Let m_0, m_1 as in the statement of Theorem 1. Then the solution to (18) verifies the long-time decay estimate (17) and the estimate (15).*

Proof The proof is as the proof of Lemma 1, formally setting $p_0 = \infty$, and with $m_0 = m_1 = 1$. □

We define the solution space

$$X = u \in C([0,\infty), L^1 \cap L^\infty),$$

equipped with the norm

$$\|u\|_X = \sup_{t\in[0,\infty)} \left((1+t)^{-1+\frac{n}{\sigma}} \|u(t,\cdot)\|_{L^\infty} + (1+t)^{-1} \|u(t,\cdot)\|_{L^1}\right).$$

In particular, by interpolation, for any $u \in X$, the following estimate holds

$$\|u(t,\cdot)\|_{L^m} \le (1+t)^{1-\frac{n}{\sigma}\left(1-\frac{1}{m}\right)} \|u\|_X, \quad \forall m \in [1,\infty]. \tag{24}$$

As a consequence of Lemma 3, the solution u^{lin} to (18) is in X and (21) holds. We now look for $u \in C([0,\infty), L^1 \cap L^\infty)$ verifying (22).

The proof of Lemma 2 is now modified.

Proof ***(Lemma 2)*** Letting $r \in (n/(n-\alpha), \infty)$, $r' = r/(r-1)$, $p' = p/(p-1)$ and $q' = q/(q-1)$, and proceeding as in the proof in Sect. 2, we get

$$\begin{aligned}\|h(u,v)\|_{L^1} &\le C_3 \|u-v\|_{L^{r*p}} \left(\|u\|_{L^{r*p}}^{p-1} + \|v\|_{L^{r*p}}^{p-1}\right) \|u\|_{L^{r'q}}^{q} \\ &\quad + C_3 \|v\|_{L^{r*p}}^{p} \|u-v\|_{L^{r'q}} \left(\|u\|_{L^{r'q}}^{q-1} + \|v\|_{L^{r'q}}^{q-1}\right).\end{aligned}$$

Since we did not assume (8), we cannot fix

$$r = \frac{n}{np-\alpha q}(p+q),$$

in general, as we did in the proof in Sect. 2. However, thanks to (24) it is now not important to fix a specific value for r, and we may choose any $r \in (n/(n-\alpha), \infty)$. Indeed, independently on the choice of r, we obtain the estimate

$$\begin{aligned}&\|h(u,v)(t,\cdot)\|_{L^1} \\ &\quad \le C \|u-v\|_X \left(\|u\|_X^{p+q-1} + \|v\|_X^{p+q-1}\right) (1+t)^{(p+q)(1-\frac{n}{\sigma})+\frac{n+\alpha}{\sigma}},\end{aligned}$$

for any $t \ge 0$. Similarly, we may derive

$$\begin{aligned}&\|h(u,v)(t,\cdot)\|_{L^{\frac{n}{\sigma}}} \\ &\quad \le C \|u-v\|_X \left(\|u\|_X^{p+q-1} + \|v\|_X^{p+q-1}\right) (1+t)^{(p+q)(1-\frac{n}{\sigma})+\frac{\alpha}{\sigma}}.\end{aligned}$$

Using (19) with $u_0 = 0$, $q_2 = \infty$ and $q_1 = 1$ if $s \in [0, t/2]$ and $q_1 = n/\sigma$ if $s \in [t/2, t]$, we get

$$\begin{aligned}\|(Nu-Nv)(t,\cdot)\|_{L^\infty} &\le C \int_0^{t/2} (t-s)^{1-\frac{n}{\sigma}} \|h(u,v)(s,\cdot)\|_{L^1}\, ds \\ &\quad + C \int_{t/2}^{t} \|h(u,v)(s,\cdot)\|_{L^{\frac{n}{\sigma}}}\, ds \\ &\le C \|u-v\|_X \left(\|u\|_X^{p+q-1} + \|v\|_X^{p+q-1}\right) \big(I_1(t) + I_2(t)\big),\end{aligned}$$

where

$$I_1(t) = \int_0^{t/2} (t-s)^{1-\frac{n}{\sigma}} (1+s)^{(p+q)(1-\frac{n}{\sigma})+\frac{n+\alpha}{\sigma}} \, ds,$$
$$I_2(t) = \int_{t/2}^{t} (1+s)^{(p+q)(1-\frac{n}{\sigma})+\frac{\alpha}{\sigma}} \, ds \, .$$

Thanks to (6), that is,

$$(p+q)\left(1-\frac{n}{\sigma}\right)+\frac{n+\alpha}{\sigma} < -1,$$

we may estimate:

$$\begin{aligned} I_1(t) &\leq C\,(1+t)^{1-\frac{n}{\sigma}} \int_0^{t/2} (1+s)^{(p+q)(1-\frac{n}{\sigma})+\frac{n+\alpha}{\sigma}} \, ds \\ &\leq C'\,(1+t)^{1-\frac{n}{\sigma}}, \\ I_2(t) &\leq C\,(1+t)^{1+(p+q)(1-\frac{n}{\sigma})+\frac{\alpha}{\sigma}} \leq C\,(1+t)^{1-\frac{n}{\sigma}} \, . \end{aligned}$$

We proceed similarly to estimate $\|(Nu - Nv)(t,\cdot)\|_{L^1}$, and we conclude the proof. □

The conclusion of the proof of Theorem 2 is as the conclusion of the proof of Theorem 1.

References

1. Chadam, J.M., Glassey, R.T.: Global existence of solutions to the Cauchy problem for time-dependent Hartree equations. J. Math. Phys. **16**, 1122–1130 (1975)
2. Chen, W., D'Abbicco, M., Girardi, G.: Global small data solutions for semilinear waves with two dissipative terms. Ann. Mat. (2021). https://doi.org/10.1007/s10231-021-01128-z
3. Cingolani, S., Secchi, S., Squassina, M.: Semiclassical limit for Schrödinger equations with magnetic field and Hartree-type nonlinearities. Proc. R. Soc. Edinburgh A **140**, 973–1009 (2010)
4. D'Abbicco, M.: A wave equation with structural damping and nonlinear memory. Nonlinear Differential Equations Appl. **21**(5), 751–773 (2014)
5. D'Abbicco, M.: A benefit from the L^1 smallness of initial data for the semilinear wave equation with structural damping. In: Mityushev, V., Ruzhansky, M. (eds.) Current Trends in Analysis and its Applications. Proceedings of the 9th ISAAC Congress, pp. 209–216. Krakow (2015). http://www.springer.com/br/book/9783319125763
6. D'Abbicco, M., Ebert, M.R.: A new phenomenon in the critical exponent for structurally damped semi-linear evolution equations. Nonlinear Anal. Theory Methods Appl. **149**, 1–40 (2017)

7. D'Abbicco, M., Ebert, M.R.: The critical exponent for semilinear sigma-evolution equations with a strong non-effective damping. Nonlinear Anal. **215**, 112637 (2022). https://doi.org/10.1016/j.na.2021.112637
8. D'Abbicco, M., Girardi, G.: A structurally damped σ- evolution equation with nonlinear memory. Math Methods. Appl. Sci. (2020). https://doi.org/10.1002/mma.6633
9. D'Abbicco, M., Girardi, G., Liang, J.: $L^1 - L^1$ estimates for the strongly damped plate equation. J. Math. Anal. Appl. **478**, 476–498 (2019)
10. Ebert, M.R., Lourenço, L.M.: The critical exponent for evolution models with power nonlinearity. In: Trends in Mathematics, New Tools for Nonlinear PDEs and Applications, pp. 153–177. Birkhäuser, Basel (2019)
11. Ebert, M.R., Girardi, G., Reissig, M.: Critical regularity of nonlinearities in semilinear classical damped wave equations. Math. Ann. **378**(3–4), 1311–1326 (2020)
12. Filippucci, R., Ghergu, M.: Fujita type results for quasilinear parabolic inequalities with nonlocal terms. Discrete Contin. Dyn. Syst. **42**(4), 1817–1833 (2022). https://doi.org/10.3934/dcds.2021173
13. Hartree, D.R.: The wave mechanics of an atom with a non-Coulomb central field, Part I. Theory and methods. Math. Proc. Camb. Philos. Soc. **24**, 89–110 (1928)
14. Hartree, D.R.: The wave mechanics of an atom with a non-Coulomb central field, Part II. Some results and discussion. Math. Proc. Camb. Philos. Soc. **24**, 111–132 (1928)
15. Hartree, D.R.: The wave mechanics of an atom with a non-Coulomb central field, Part III. Term values and intensities in series in optical spectra. Math. Proc. Camb. Philos. Soc. **24**, 426–437 (1928)
16. Hayashi, N., Ozawa, T.: Time decay of solutions to the cauchy problem for time-dependent Schrödinger-Hartree equations. Commun. Math. Phys. **110**(3), 467–478 (1987)
17. Shibata, Y.: On the rate of decay of solutions to linear viscoelastic equation. Math. Methods Appl. Sci. **23**, 203–226 (2000)
18. Todorova, G., Yordanov, B.: Critical exponent for a nonlinear wave equation with damping. J. Differential Equations **174**, 464–489 (2001)

A Note on Continuity of Strongly Singular Calderón-Zygmund Operators in Hardy-Morrey Spaces

Marcelo de Almeida, Tiago Picon, and Claudio Vasconcelos

Abstract In this note we address the continuity of strongly singular Calderón-Zygmund operators on Hardy-Morrey spaces $\mathcal{HM}^{\lambda}_{q}(\mathbb{R}^n)$, assuming weaker integral conditions on the associated kernel. Important examples that falls into this scope are pseudodifferential operators on the Hörmander classes $OpS^{m}_{\sigma,\mu}(\mathbb{R}^n)$ with $0 < \sigma \leq 1$, $0 \leq \mu < 1$, $\mu \leq \sigma$ and $m \leq -n(1-\sigma)/2$.

1 Introduction

J. Álvarez and M. Milman [1] introduced a new class of Calderón-Zygmund operators, called *strongly singular Calderón-Zygmund operator* and established the continuity of these operators in real Hardy spaces $H^q(\mathbb{R}^n)$. More precisely, a continuous function $K \in C(\mathbb{R}^{2n}\backslash\Delta)$, where $\Delta = \{(x,x) : x \in \mathbb{R}^n\}$ is a δ-kernel of type σ, if there exists some $0 < \delta \leq 1$ and $0 < \sigma \leq 1$ such that

$$|K(x,y) - K(x,z)| + |K(y,x) - K(z,x)| \leq C\frac{|y-z|^{\delta}}{|x-z|^{n+\frac{\delta}{\sigma}}}, \tag{1}$$

for all $|x-z| \geq 2|y-z|^{\sigma}$. A bounded and linear operator $T : \mathcal{S}(\mathbb{R}^n) \to \mathcal{S}'(\mathbb{R}^n)$ is called a *strongly singular Calderón-Zygmund operator*, if it is associated to a

M. de Almeida
Departamento de Matemática, Universidade Federal de Sergipe, Aracajú, SE, Brasil
e-mail: marcelo@mat.ufs.br

T. Picon (✉)
Departamento de Computação e Matemática, Universidade de São Paulo, Ribeirão Preto, SP, Brasil
e-mail: picon@ffclrp.usp.br

C. Vasconcelos
Departamento de Matemática, Universidade Federal de São Carlos, São Carlos, SP, Brasil
e-mail: claudio.vasconcelos@estudante.ufscar.br

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_44

δ-kernel of type σ in the sense $\langle Tf, g\rangle = \int\int K(x, y) f(y)g(x)dydx$, for all $f, g \in \mathcal{S}(\mathbb{R}^n)$ with disjoint supports; it has bounded extension from $L^2(\mathbb{R}^n)$ to itself and in addition T and T^* extend to a continuous operator from $L^p(\mathbb{R}^n)$ to $L^2(\mathbb{R}^n)$, where $\frac{1}{p} = \frac{1}{2} + \frac{\beta}{n}$ for some $(1-\sigma)\frac{n}{2} \le \beta < \frac{n}{2}$. When $\sigma = 1$ and $\beta = 0$ we recover the standard non-convolution Calderón-Zygmund operators (see [3]).

The authors in [1, Theorem 2.2] established the continuity of those classes of operators in real Hardy spaces $H^q(\mathbb{R}^n)$ as follows: under the condition $T^*(1) = 0$, strongly singular Calderón-Zygmund operators associated to a kernel satisfying (1) are bounded from $H^q(\mathbb{R}^n)$ to itself for every $q_0 < q \le 1$ where

$$\frac{1}{q_0} := \frac{1}{2} + \frac{\beta\left(\frac{\delta}{\sigma} + \frac{n}{2}\right)}{n\left(\frac{\delta}{\sigma} - \delta + \beta\right)}. \tag{2}$$

The case $q = q_0$ is still open, however the conclusion continues to hold replacing the target space by $L^{q_0}(\mathbb{R}^n)$ (see [2, Theorem 3.9]).

In this note, we establish results on continuity of strongly singular Calderón-Zygmund operators on Hardy-Morrey spaces $\mathcal{HM}^{\lambda}_q(\mathbb{R}^n)$ assuming weaker integral conditions on the kernel, introduced by the second and third authors in [12]. Let $0 < \sigma \le 1$, $r \ge 1$ and $\delta > 0$. We say that $K(x, y)$, associated to T, is a $D_{\delta,r}$ kernel of type σ if

$$\left(\int_{C_j(z,\ell)} |K(x, y) - K(x, z)|^r + |K(y, x) - K(z, x)|^r dx\right)^{\frac{1}{r}} \lesssim |C_j(z, \ell)|^{\frac{1}{r}-1} 2^{-j\delta} \tag{3}$$

for $\ell \ge 1$ and

$$\left(\int_{C_j(z,\ell^\rho)} |K(x, y) - K(x, z)|^r + |K(y, x) - K(z, x)|^r dx\right)^{\frac{1}{r}} \lesssim |C_j(z, \ell^\rho)|^{\frac{1}{r}-1+\frac{\delta}{n}\left(\frac{1}{\rho}-\frac{1}{\sigma}\right)} 2^{-\frac{j\delta}{\rho}} \tag{4}$$

for $\ell < 1$, where $z \in \mathbb{R}^n$, $|y - z| < \ell$, $0 < \rho \le \sigma$ and $C_j(z, \eta) := \{x \in \mathbb{R}^n : 2^j\eta < |x - z| \le 2^{j+1}\eta\}$. These conditions also covers the standard case $\sigma = 1$, by choosing $\rho = \sigma$ in (4), and in that case both conditions are the same. It is easy to check that D_{δ,r_1} condition is stronger than D_{δ,r_2} for $r_1 > r_2$ and any $\delta > 0$ and $0 < \sigma \le 1$. Moreover, $\delta-$kernels of type σ satisfying (1) also satisfies $D_{\delta,r}$ condition for all $r \ge 1$. It has also been shown in [12, Proposition 5.3] that pseudodifferential operators associated to symbols in the Hörmander classes $S^m_{\sigma,\mu}(\mathbb{R}^n)$ with $0 < \sigma \le 1$, $0 \le \mu < 1$, $\mu \le \sigma$ and $m \le -n(1-\sigma)/2$, satisfy the $D_{1,r}$ condition for $1 \le r \le 2$. We refer to [12] for more details. In particular, the continuity of operators associated

to symbols given by $e^{i|\xi|^\sigma}|\xi|^{-m}$ away from the origin are also examples of this type of operators and have been extensively studied, for instance in [4, 6, 8, 13].

Our main result is the following:

Theorem 1 *Let T to be a strongly singular Calderón-Zygmund operator associated to a $D_{\delta,r}$ kernel of type σ for some $1 \leq r \leq 2$. Under the assumption that $T^*(x^\alpha) = 0$ for every $|\alpha| \leq \lfloor \delta \rfloor$, T can be extended to a bounded operator from $\mathcal{HM}^\lambda_q(\mathbb{R}^n)$ to itself for any $0 < q \leq \lambda < r$ and $q_0 < q \leq 1$, where q_0 is given by (2).*

The proof relies on showing that T maps atoms into molecules and a molecular decomposition in $\mathcal{HM}^\lambda_q(\mathbb{R}^n)$ for $0 < q \leq 1$ and $q \leq \lambda < \infty$ under restriction $\lambda < r$ (see Theorem 3 and the Remark 2). As an immediate consequence of previous theorem, we also obtain the continuity of standard non-convolution Calderón-Zygmund operators ($\sigma = 1$) associated to kernels satisfying integral conditions. The corresponding result in the convolution setting for kernels satisfying derivative conditions can be found in [10, Section 2.2].

Corollary 1 *Under the same hypothesis of the previous theorem, if T is a standard Calderón-Zygmund operator, then it is bounded from $\mathcal{HM}^\lambda_q(\mathbb{R}^n)$ to itself provided that $n/(n+\delta) < q \leq 1$.*

The organization of the paper is as follows. In Sect. 2 we recall some basic definitions and a general atomic and molecular decomposition of Hardy-Morrey spaces. In particular, in Sect. 2.1 we present an atomic decomposition in terms of L^r-atoms by showing the equivalence with classical L^∞ atomic space and in Sect. 2.2 we show an appropriate molecular decomposition of Hardy-Morrey spaces. Finally, in Sect. 3 we present the proof of Theorem 1 showing that T maps atoms into molecules.

Notation throughout this work, the symbol $f \lesssim g$ means that there exist a constant $C > 0$, not depending on f nor g, such that $f \leq C\, g$. By a dyadic cube we mean cubes on $\mathbb{R}^n$, open on the right whose vertices are adjacent points of the lattice $(2^{-k}\mathbb{Z})^n$ for some $k \in \mathbb{Z}$. Given a set $A \subset \mathbb{R}^n$ we denote by $|A|$ its Lebesgue measure. Given a cube Q (dyadic or not), we will always denote its center and side-length by x_Q and ℓ_Q respectively. By Q^* we mean the cube with same center as Q and side-length $2\ell_Q$. We also denote by $\fint_Q f(x)dx := \frac{1}{|Q|}\int_Q f(x)dx$.

2 Hardy-Morrey Spaces $\mathcal{HM}^\lambda_q(\mathbb{R}^n)$

In this section, we recall and present some properties of Hardy-Morrey spaces. For $0 < q \leq \lambda < \infty$, the Morrey spaces, denoted by $\mathcal{M}^\lambda_q(\mathbb{R}^n)$, are defined to be the set

of measurable functions $f \in L^q_{loc}(\mathbb{R}^n)$ such that

$$\|f\|_{\mathcal{M}^\lambda_q} := \sup_J |J|^{\frac{1}{\lambda}-\frac{1}{q}} \left(\int_J |f(y)|^q dy \right)^{\frac{1}{q}} < \infty,$$

where the supremum is taken over all cubes $J \subset \mathbb{R}^n$.

For any tempered distribution $f \in \mathcal{S}'(\mathbb{R}^n)$ and any fixed $\varphi \in \mathcal{S}(\mathbb{R}^n)$ with $\int \varphi \neq 0$, consider the smooth maximal function $M_\varphi f(x) = \sup_{t>0} |(\varphi_t * f)(x)|$, where $\varphi_t(x) = t^{-n}\varphi(x/t)$. For any $0 < q \leq \lambda < \infty$, we say that $f \in \mathcal{S}'(\mathbb{R}^n)$ belongs to Hardy-Morrey space $\mathcal{HM}^\lambda_q(\mathbb{R}^n)$ if the smooth maximal function $M_\varphi f \in \mathcal{M}^\lambda_q(\mathbb{R}^n)$. The functional $\|f\|_{\mathcal{HM}^\lambda_q} := \|M_\varphi f\|_{\mathcal{M}^\lambda_q}$ defines a quasi-norm as $0 < q < 1$ and a norm if $q \geq 1$.

In the same way as Hardy spaces, the Hardy-Morrey spaces have also equivalent maximal characterizations (see [9, Section 2]). Clearly, Hardy-Morrey spaces cover the classical Hardy spaces $H^p(\mathbb{R}^n)$ when $\lambda = q$ and Morrey spaces $\mathcal{M}^\lambda_q(\mathbb{R}^n)$ if $1 < q \leq \lambda < \infty$.

2.1 Atomic Decomposition in Hardy-Morrey Spaces

Definition 1 [10, Definiton 2.2]. Let $0 < q \leq 1 \leq r \leq \infty$ with $q < r$ and $q \leq \lambda < \infty$. A measurable function a_Q is called a $(q, \lambda, r)-$atom if it is supported on a cube $Q \subset \mathbb{R}^n$ and satisfies: (1) $\|a_Q\|_{L^r} \leq |Q|^{\frac{1}{r}-\frac{1}{\lambda}}$ and (2) $\int_{\mathbb{R}^n} x^\alpha a_Q(x)dx = 0$ for all $\alpha \in \mathbb{N}^n_0$ such that $|\alpha| \leq N_q := \lfloor n\,(1/q - 1) \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function.

The following lemma is an extension of [5, Proposition 2.5] and the proof will be presented for completeness.

Proposition 1 *Let $0 < q \leq 1 \leq r \leq \infty$ with $q < r$ and $q \leq \lambda < \infty$ with $\lambda \leq r$. If f is a compactly supported function in $L^r(\mathbb{R}^n)$ satisfying the moment condition*

$$\int_{\mathbb{R}^n} x^\alpha f(x)dx = 0 \text{ for all } |\alpha| \leq N_q, \tag{5}$$

then it belongs to $\mathcal{HM}^\lambda_q(\mathbb{R}^n)$ and moreover $\|f\|_{\mathcal{HM}^\lambda_q} \lesssim \|f\|_{L^r}|Q|^{1/\lambda-1/r}$ for all cube $Q \supseteq supp\,(f)$. In particular, if $f = a_Q$, then $\|a_Q\|_{\mathcal{HM}^\lambda_q} \lesssim 1$ uniformly.

Proof Let $J \subset \mathbb{R}^n$ be an arbitrary cube and Q a cube such that $\text{supp}\,(f) \subseteq Q$. Split the integral over J into $J \cap Q^*$ and $J \setminus Q^*$. Since the maximal function M_φ is bounded from $L^r(\mathbb{R}^n)$ to itself for every $1 < r \leq \infty$, it follows that

$$\int_{J\cap Q^*} |M_\varphi f(x)|^q dx \leq \|M_\varphi f\|^q_{L^r} |J \cap Q^*|^{1-\frac{q}{r}} \lesssim \|f\|^q_{L^r} |J \cap Q^*|^{1-\frac{q}{r}}.$$

For $r = 1$ and $0 < q < 1$, setting $R = \|f\|_{L^1}|J \cap Q^*|^{-1}$ and using that M_φ satisfies weak (1, 1) inequality we get the analogous inequality:

$$\int_{J\cap Q^*} |M_\varphi f(x)|^q dx \simeq \int_0^\infty \omega^{q-1} \left|\{x \in J \cap Q^* : |M_\varphi f(x)| > \omega\}\right| d\omega$$
$$\lesssim |J \cap Q^*| \int_0^R \omega^{q-1} d\omega + \|f\|_{L^1} \int_R^\infty \omega^{q-2} d\omega \lesssim \|f\|_{L^1}^q |J \cap Q^*|^{1-q}. \tag{6}$$

If $|Q| < |J|$, since $q/\lambda - 1 \leq 0$ and $1 - q/r > 0$ for all $1 \leq r < \infty$, one has $|J|^{q/\lambda-1}|J \cap Q^*|^{1-q/r} \leq |Q|^{q/\lambda-q/r}$. On the other hand, if $|J| < |Q|$, using that $\lambda \leq r$ it follows $|J|^{\frac{q}{\lambda}-1}|J \cap Q^*|^{1-\frac{q}{r}} = |J|^{\frac{q}{\lambda}-\frac{q}{r}} \left(\frac{|J\cap Q^*|}{|J|}\right)^{1-\frac{q}{r}} \leq |Q|^{1-\frac{q}{r}}$. Hence $|J|^{\frac{q}{\lambda}-1} \int\!\!\int_{J\cap Q^*} |M_\varphi f(x)|^q dx \lesssim \|f\|_{L^r}^q |Q|^{q/\lambda-q/r}$.

To estimate the integral on $J \setminus Q^*$, using the moment condition (5) we write $\varphi_t * f(x) = \int\!\!\int f(y) \left(\varphi_t(x-y) - P_{\varphi_t}(y)\right) dy$, where $P_{\varphi_t}(y) = \sum_{|\alpha|\leq N_q} C_\alpha \, \partial^\alpha \varphi_t(x) \, (-y)^\alpha$ denotes the Taylor polynomial of degree N_q of the function $y \mapsto \varphi_t(x - y)$. The standard estimate of the remainder term (see [11, p. 106]) yields $\left|\varphi_t(x-y) - P_{\varphi_t}(y)\right| \lesssim |y - x_Q|^{N_q+1}|x - x_Q|^{-(n+N_q+1)}$ and since $\operatorname{supp}(f) \subseteq Q$, we have the pointwise control

$$\left|M_\varphi f(x)\right| \lesssim \frac{\ell_Q^{N_q+1}}{|x-x_Q|^{n+N_q+1}} \int_Q |f(y)| dy \lesssim \frac{\ell_Q^{N_q+1}}{|x-x_Q|^{n+N_q+1}} \|f\|_{L^r} |Q|^{1-\frac{1}{r}}.$$

If $|Q| < |J|$, since $N_q + 1 > n\,(1/q - 1)$, we estimate $|J|^{\frac{q}{\lambda}-1} \int\!\!\int_{J\cap Q^*} |M_\varphi f(x)|^q dx$ by

$$\|f\|_{L^r}^q |Q|^{q\left(\frac{1}{\lambda}-\frac{1}{r}+\frac{N_q}{n}+\frac{1}{n}+1\right)-1} \int_{(Q^*)^c} |x - x_Q|^{-q(n+N_q+1)} dx \lesssim \|f\|_{L^r}^q |Q|^{\frac{q}{\lambda}-\frac{q}{r}}.$$

Finally, if $|J| < |Q|$

$$|J|^{\frac{q}{\lambda}-1} \int_{J\setminus Q^*} |M_\varphi f(x)|^q dx \lesssim \|f\|_{L^r}^q |J|^{\frac{q}{\lambda}-1} |Q|^{q-\frac{q}{r}} \ell_Q^{-nq} \, |J \setminus Q^*| \lesssim \|f\|_{L^r}^q |Q|^{\frac{q}{\lambda}-\frac{q}{r}},$$

which concludes the proof. □

Given $1 \leq r \leq \infty$, we denote the atomic space $\mathbf{at}\mathcal{HM}_q^{\lambda,r}(\mathbb{R}^n)$ by the collection of $f \in \mathcal{S}'(\mathbb{R}^n)$ such that $f = \sum_Q : \text{dyadic } s_Q a_Q$ in $\mathcal{S}'(\mathbb{R}^n)$, where $\{a_Q\}_Q$ are

(q,λ,r)-atoms and $\{s_Q\}_Q$ is a sequence of complex scalars satisfying

$$\|\{s_Q\}_Q\|_{\lambda,q} := \sup_J \left\{ \left(|J|^{\frac{q}{\lambda}-1} \sum_{Q\subseteq J} \left(|Q|^{\frac{1}{q}-\frac{1}{\lambda}} |s_Q| \right)^q \right)^{\frac{1}{q}} \right\} < \infty.$$

The functional $\|f\|_{\mathbf{at}\mathcal{HM}_q^{\lambda,r}} := \inf\left\{ \|\{s_Q\}_Q\|_{\lambda,q} : f = \sum_Q s_Q a_Q \right\}$, where the infimum is taken over all such atomic representations, defines a quasi-norm in $\mathbf{at}\mathcal{HM}_q^{\lambda,r}(\mathbb{R}^n)$. Clearly, if $1 \le r_1 < r_2 \le \infty$ then $\mathbf{at}\mathcal{HM}_q^{\lambda,r_2}(\mathbb{R}^n)$ is continuously embedded in $\mathbf{at}\mathcal{HM}_q^{\lambda,r_1}(\mathbb{R}^n)$. The converse of this simple embedding is the content of the next result.

Lemma 1 *Let $0 < q \le 1 \le r$ with $q < r$ and $q \le \lambda < \infty$. Then $\mathbf{at}\mathcal{HM}_q^{\lambda,r}(\mathbb{R}^n) = \mathbf{at}\mathcal{HM}_q^{\lambda,\infty}(\mathbb{R}^n)$ with comparable quasi-norms.*

Proof The proof is based on the corresponding theorem for Hardy spaces (see [7, Theorem 4.10]). Let a_Q to be a $(q,\lambda,r)-$atom and we show that $a_Q = \sum_j s_{Q_j} a_{Q_j}$, where $\{a_{Q_j}\}_j$ are $(q,\lambda,\infty)-$atoms and $\|\{s_{Q_j}\}_j\|_{q,\lambda} \le C$ independently. Consider $b_Q = |Q|^{1/\lambda} a_Q$ and since $\fint_Q |b_Q(x)|^r dx \le |Q|$, from Calderón-Zygmund decomposition applied for $|b_Q|^r \in L^1(Q)$ at level $\alpha^r > 0$, there exists a sequence $\{Q_j\}_j$ of disjoint dyadic cubes (subcubes of Q) such that $|b_Q(x)| \le \alpha$, $\forall\, x \notin \bigcup_j Q_j$, $\alpha^r \le \fint_{Q_j} |b_Q(x)|^r dx \le 2^n \alpha^r$ and $\left|\bigcup_j Q_j\right| \le \alpha^{-r} \fint_Q |b_Q(x)|^r dx \le |Q|\,\alpha^{-r}$. Let $\mathcal{P}_{N_q}$ to be the space of polynomials in $\mathbb{R}^n$ with degree at most N_q and $\mathcal{P}_{N_q,j}$ its restriction to Q_j. Since $\mathcal{P}_{N_q,j}$ is a subspace of the Hilbert space $L^2(Q_j)$, let $P_{Q_j} b \in \mathcal{P}_{N_q,j}$ to be the unique polynomial such that $\fint_{Q_j} [b_Q(x) - P_{Q_j}(b)(x)] x^\beta dx = 0$ for all $|\beta| \le N_q$.

Now we write $b_Q = g_0 + \sum_j h_j$, where $h_j(x) = [b_Q(x) - P_{Q_j}(b)(x)]\mathbb{1}_{Q_j}(x)$ and $g_0(x) = b_Q(x)$ if $x \notin \bigcup_j Q_j$ and $g_0(x) = P_{Q_j}(b)(x)$ if $x \in Q_j$. Clearly $\int h_j(x) x^\beta dx = 0$ and since $|g_0(x)| \le c\alpha$ almost everywhere (see [11, Remark 2.1.4 p. 104]), this implies

$$\left(\fint_{Q_j} |h_j(x)|^r dx \right)^{1/r} \le \left(\fint_{Q_j} |b_Q(x)|^r dx \right)^{1/r} + \left(\fint_{Q_j} |g_0(x)|^r dx \right)^{1/r} \le c\alpha.$$

For each $j_0 \in \mathbb{N}$, let $b_{j_0}(x) := (c\alpha)^{-1} h_{j_0}(x)$ and write $b_Q(x) = g_0(x) + (c\alpha)\sum_{j_0} b_{j_0}(x)$, where $\fint_{Q_{j_0}} |b_{j_0}(x)|^r dx \le |Q_{j_0}|$. Applying the previous argument for each b_{j_0} we obtain the identity

$$b_Q = g_0 + (c\alpha)\sum_{j_0} b_{j_0} = g_0 + c\alpha \sum_{j_0} g_{j_0} + (c\alpha)^2 \sum_{j_0,j_1} b_{j_0,j_1},$$

where $\int\!\!\!\!\!\!\diagup_{Q_{j_0,j_1}} |b_{j_0,j_1}(x)|^r dx \leq |Q_{j_0,j_1}|$ and $\{Q_{j_0,j_1}\}_{j_1}$ is a sequence of disjoint dyadic cubes (subcubes of Q_{j_0}) such that $|g_{j_0}(x)| \leq c\alpha$ a.e., $\alpha^r \leq \int\!\!\!\!\!\!\diagup_{Q_{j_0,j_1}} |b_{j_0}(x)|^r dx \leq 2^n \alpha^r$ and $\big|\bigcup_{j_1} Q_{j_0,j_1}\big| \leq c\,\alpha^{-r} \int\!\!\!\!\!\!\diagup_{Q_{j_0}} |b_{j_0}(x)|^r dx \leq c|Q_{j_0}|\,\alpha^{-r}$. Employing an induction argument, we can find a family $\{Q_{i_{k-1},j}\}_j := \{Q_{j_0,\cdots,j_{k-1},j}\}_j$ of disjoint dyadic subcubes of $Q_{i_{k-1}} := Q_{j_0,\cdots,j_{k-1}}$ for $k = 1, 2, \cdots$ with $i_{k-1} = \{j_0, j_1, \cdots, j_{k-1}\}$ such that

$$b_Q = g_{i_0} + c\alpha \sum_{i_1} g_{i_1} + (c\alpha)^2 \sum_{i_2} g_{i_2} + \cdots + (c\alpha)^{k-1} \sum_{i_{k-1}} g_{i_{k-1}} + (c\alpha)^k \sum_{i_k} h_{i_k}, \tag{7}$$

in which $g_{i_{k-1}}$ and h_{i_k}, for every $i_k = (j_0, j_1, \cdots, j_{k-1}, j)$, satisfy $|g_{i_{k-1}}(x)| \leq c\alpha$ a.e. $x \in \mathbb{R}^n$, $\alpha^r \leq \int\!\!\!\!\!\!\diagup_{Q_{i_{k-1},j}} |h_{i_k}(x)|^r dx \leq 2^n \alpha^r$ and $\big|\bigcup_j Q_{i_{k-1},j}\big| \leq c|Q_{i_{k-1}}|\,\alpha^{-r}$. The sum at (7) is interpreted as $\sum_{i_{k-1}} g_{i_{k-1}} := \sum_{j_0\in\mathbb{N}} \cdots \sum_{j_{k-1}\in\mathbb{N}} g_{j_0,\cdots,j_{k-1}}$ (analogously to $\sum_{i_k} h_{i_k}$). We claim that the reminder term $(c\alpha)^k \sum_{i_k} h_{i_k}$ in (7) goes to zero in $L^1(\mathbb{R}^n)$ as $k \to \infty$. Indeed, writing $Q_{i_k} := Q_{i_{k-1},j}$ for some fixed j we have $\int_{\mathbb{R}^n} |h_{i_k}(x)|dx = \int\!\!\!\!\!\!\diagup_{Q_{i_k}} |h_{i_k}(x)|dx \leq \Big(\int\!\!\!\!\!\!\diagup_{Q_{i_k}} |h_{i_k}(x)|^r dx\Big)^{\frac{1}{r}} |Q_{i_k}|^{1-\frac{1}{r}} \leq c\alpha|Q_{i_k}|$ and iterating $(k+1)$-times the previous argument one has

$$\sum_{i_k} |Q_{i_k}| \leq \Big(\frac{c}{\alpha^r}\Big)^{k+1} |Q|. \tag{8}$$

Thus, $\int \Big|(c\alpha)^k \sum_{i_k} h_{i_k}(x)\Big| dx \leq (c\alpha)^{k+1} \sum_{i_k} |Q_{i_k}| \leq (c^2\alpha^{1-r})^{(k+1)} |Q|$. That means, $(c\alpha)^k \sum_{i_k} h_{i_k}(x)$ goes to 0 in $L^1(\mathbb{R}^n)$ as $k \to \infty$, provided that $c^2\alpha^{1-r} < 1$. Therefore,

$$b_Q = g_{i_0} + c\alpha \sum_{i_1} g_{i_1} + (c\alpha)^2 \sum_{i_2} g_{i_2} + \cdots + (c\alpha)^{k-1} \sum_{i_{k-1}} g_{i_{k-1}} + (c\alpha)^k \sum_{i_k} g_{i_k} + \cdots$$

in $L^1(\mathbb{R}^n)$, where $|g_{i_k}(x)| \leq c\alpha$ a.e. and for all $|\beta| \leq N_q$ we have $\int\!\!\!\!\!\!\diagup x^\beta g_{i_k}(x)dx = \int\!\!\!\!\!\!\diagup x^\beta b_{i_k}(x)dx + \sum_j \int\!\!\!\!\!\!\diagup_{Q_{i_{k-1},j}} x^\beta P_{Q_{i_k,j}} b(x)dx = \int\!\!\!\!\!\!\diagup x^\beta b_{i_k}(x)dx = 0$. From the above considerations it is clear that $a_{i_0} := (c\alpha)^{-1} |Q|^{-1/\lambda} g_{i_0}$ and $a_{i_k} := (c\alpha)^{-1} |Q_{i_k}|^{-1/\lambda} g_{i_k}$ are $(q, \lambda, \infty)-$atoms for all $k = 1, 2, \cdots$. Moreover, we can write

$$a_Q = s_{i_0} a_{i_0} + \sum_{i_1} s_{i_1} a_{i_1} + \sum_{i_2} s_{i_2} a_{i_2} + \cdots + \sum_{i_k} s_{i_k} a_{i_k} + \cdots \tag{9}$$

where each coefficient $\{s_{i_k}\}$ is defined by $s_{i_k} = (c\alpha)^{k+1}|Q|^{-1/\lambda}|Q_{i_k}|^{1/\lambda}$. It remains to show that $\|\{s_{i_k}\}_k\|_{\lambda,q} \leq C$, uniformly. Fixed $J \subset \mathbb{R}^n$ a dyadic cube, we may estimate

$$
\begin{aligned}
|J|^{\frac{q}{\lambda}-1} \sum_{k=0}^{\infty} \sum_{Q_{i_k} \subseteq J} |s_{i_k}|^q |Q_{i_k}|^{1-\frac{q}{\lambda}} &= |J|^{\frac{q}{\lambda}-1} |Q|^{-\frac{q}{\lambda}} \sum_{k=0}^{\infty} (c\alpha)^{q(k+1)} \Big(\sum_{Q_{i_k} \subseteq J} |Q_{i_k}| \Big) \\
&\lesssim |J|^{\frac{q}{\lambda}-1} |Q|^{-\frac{q}{\lambda}} |J \cap Q| \sum_{k=0}^{\infty} (c\alpha)^{q(k+1)} \left(\tfrac{c}{\alpha^r}\right)^{k+1} \\
&\leq C
\end{aligned}
$$

provided $c^{q+1}\alpha^{q-r} < 1$ (weaker than the previous one) and $q \leq \lambda$. Note that here we have used a refinement of (8) given by $\sum_{i_k \,:\, Q_{i_k} \subseteq J} |Q_{i_k}| \lesssim \left(\frac{c}{\alpha^r}\right)^{k+1} |J \cap Q|$ and the uniform control $|J|^{q/\lambda-1}|Q|^{-q/\lambda}|J \cap Q| \lesssim 1$. □

The previous lemma allows us to study Hardy-Morrey spaces $\mathcal{HM}^{\lambda}_{q}(\mathbb{R}^n)$ with any of the atomic spaces $\mathbf{at}\mathcal{HM}^{\lambda,r}_{q}(\mathbb{R}^n)$ for $1 \leq r \leq \infty$ provided that $q < r$. In addition, we announce an atomic decomposition in terms of $(q, \lambda, r)-$atoms, which is a direct consequence of the one proved in [9, p. 100] for $(q, \lambda, \infty)-$atoms and Lemma 1, since they are in particular $(q, \lambda, r)-$atoms.

Theorem 2 *Let $0 < q \leq 1 \leq r \leq \infty$ with $q < r$ and $q \leq \lambda < \infty$. Then, $f \in \mathcal{HM}^{\lambda}_{q}(\mathbb{R}^n)$ if and only if there exist a collection of $(q, \lambda, r)-$atoms $\{a_Q\}_Q$ and a sequence of complex numbers $\{s_Q\}_Q$ such that $f = \sum_Q s_Q a_Q$ in $\mathcal{S}'(\mathbb{R}^n)$ and $\|f\|_{\mathbf{at}\mathcal{HM}^{\lambda}_{q}} \approx \|f\|_{\mathcal{HM}^{\lambda}_{q}}$.*

2.2 Molecular Decomposition in Hardy-Morrey Spaces

Definition 2 Let $0 < q \leq 1 \leq r < \infty$ with $q < r$, $q \leq \lambda < \infty$, and $s > n\,(r/q - 1)$. A function $m(x)$ is called a $(q, \lambda, s, r)-$molecule in $\mathcal{HM}^{\lambda}_{q}(\mathbb{R}^n)$, or simply an L^r-molecule, if there exist a cube Q such that

$$
(M_1) \int_{\mathbb{R}^n} |m(x)|^r dx \lesssim \ell_Q^{\,n(1-\frac{r}{\lambda})} \quad (M_2) \int_{\mathbb{R}^n} |m(x)|^r |x - x_Q|^s dx \lesssim \ell_Q^{\,s+n(1-\frac{r}{\lambda})}
$$

and also satisfies the cancellation condition (M_3) $\iint\!\!\!\!\!\!\diagup_{\mathbb{R}^n} m(x)x^{\alpha}dx = 0$ for all $|\alpha| \leq N_q$.

Remark 1 Equivalently, we can replace the previous global estimates by (M_1) on $2Q$ and (M_2) on $(2B)^c$.

Lemma 2 *Let $m(x)$ to be an L^r-molecule. Then $m = \sum_Q d_Q\, a_Q + \sum_Q t_Q\, b_Q$ in $L^r(\mathbb{R}^n)$, where each $\{a_Q\}_Q$ are $(q, \lambda, r)-$atoms and $\{b_Q\}_Q$ are $(q, \lambda, \infty)-$atoms, for a suitable sequence of scalars $\{d_Q\}_Q$ and $\{t_Q\}_Q$.*

Proof The proof follows from the corresponding result for Hardy spaces [7, Theorem 7.16]. Let m to be a $(q,\lambda,s,r)-$molecule centered in the cube Q. For each $j \in \mathbb{N}$, let $Q_j := Q(x_Q, \ell_j)$ in which $\ell_j = 2^j \ell_Q$. Consider the collection of annulus $\{E_j\}_{j\in\mathbb{N}_0}$ given by $E_0 = Q$ and $E_j = Q_j \backslash Q_{j-1}$ for $j \geq 1$, and let $m_j(x) := m(x)\,\mathbb{1}_{E_j}(x)$. By the same arguments presented in the proof of Lemma 1, there exist polynomials $\{\phi_\gamma^j(x)\}_{|\gamma|\leq N_q}$ uniquely determined in E_j such that

$$(2^j \ell_Q)^{|\gamma|} |\phi_\gamma^j(x)| \lesssim 1 \quad \text{and} \quad \frac{1}{|E_j|}\int_{E_j} \phi_\gamma^j(x) x^\beta dx = \begin{cases} 1, & \gamma = \beta \\ 0, & \gamma \neq \beta \end{cases} \tag{10}$$

where the implicit constant is uniformly on E_j. Let $m_\gamma^j = \fint_{E_j} m_j(x) x^\gamma dx$ and consider $P_j(x) = \sum_{|\gamma|\leq N_q} m_\gamma^j \phi_\gamma^j(x)$. Splitting $m = \sum_{j=0}^\infty \left(m_j - P_j\right) + \sum_{j=0}^\infty P_j$, with convergence in $L^r(\mathbb{R}^n)$, we claim that for each j, $m_j - P_j$ is multiple of a (q,λ,r)-atom and P_j is a finite linear combination of (q,λ,∞)-atoms.

For the first sum, since m_j and P_j are supported on E_j, so is $m_j - P_j$ and by definition one has the desired vanish moments up to the order N_q. It remains to show that $m_j - P_j$ satisfies the size estimate. Indeed, from conditions (M_1) and (M_2) it follows that for every $j \in \mathbb{N}_0$

$$\|m_j\|_{L^r} \lesssim |E_j|^{\frac{1}{r}-\frac{1}{\lambda}} (2^j)^{-\frac{s}{r}+n\left(\frac{1}{\lambda}-\frac{1}{r}\right)}. \tag{11}$$

Also, from (10) it follows $|P_j(x)| \leq \left(\sum_{|\beta|\leq N_q} |\phi_\beta^j(x)| 2^{j|\beta|}\right) \fint_{E_j} |m_j(x)|dx \lesssim |E_j|^{-\frac{1}{r}} \|m_j\|_{L^r}$, where the implicit constants are independent of j. Hence, if we write $(m_j - P_j)(x) = d_j\, a_{Q_j}(x)$ for $d_j = \|m_j - P_j\|_{L^r} |Q_j|^{\frac{1}{\lambda}-\frac{1}{r}}$ and $a_{Q_j} = \frac{m_j - P_j}{\|m_j - P_j\|_{L^r}} |Q_j|^{\frac{1}{r}-\frac{1}{\lambda}}$, for each $j \in \mathbb{N}_0$, it is clear that $\{a_{Q_j}\}_j$ is a sequence of (q,λ,r)-atoms supported on Q_j. Moreover, from (11) we have $\|m_j - P_j\|_{L^r} \lesssim \|m_j\|_{L^r} \lesssim |Q_j|^{\frac{1}{r}-\frac{1}{\lambda}} (2^j)^{-\frac{s}{r}+n\left(\frac{1}{\lambda}-\frac{1}{r}\right)}$. Hence, since $s > n\,(r/q - 1)$

$$\sum_{j=0}^\infty |d_j|^q |Q_j|^{1-\frac{q}{\lambda}} \lesssim |Q|^{1-\frac{q}{\lambda}} \sum_{j=0}^\infty (2^j)^{q\left[-\frac{s}{r}+n\left(\frac{1}{q}-\frac{1}{r}\right)\right]} \lesssim |Q|^{1-\frac{q}{\lambda}}.$$

For the second sum, let $\psi_\gamma^j(x) := N_\gamma^{j+1}\left[|E_{j+1}|^{-1}\phi_\gamma^{j+1}(x) - |E_j|^{-1}\phi_\gamma^j(x)\right]$, where $N_\gamma^j = \sum_{k=j}^\infty m_\gamma^k |E_k| = \sum_{k=j}^\infty \fint_{E_k} m_Q(x) x^\gamma dx$. Then, we can represent P_j (using the vanish moments (M_3)) as $\sum_{j=0}^\infty P_j(x) = \sum_{j=0}^\infty \sum_{|\gamma|\leq N_q} \psi_\gamma^j(x)$. The function ψ_α^j is supported on E_{j+1} and by construction also satisfies vanishing moments conditions up to the order N_q. It remain to check the size condition.

Since $|\gamma| \leq n(1/\lambda - 1)$ and $s > n(r/q - 1)$, $|N_\gamma^{j+1}| \leq |Q_j|^{1-1/\lambda} (2^j \ell_Q)^{|\gamma|} (2^j)^{-s/r+n(1/\lambda-1/r)}$. The previous estimate and $(2^j \ell_Q)^{|\gamma|} |\phi_\gamma^j(x)| \leq C$ yields for all $x \in E_j$

$$\left| N_\gamma^{j+1} |E_j|^{-1} \phi_\gamma^j(x) \right| \leq C |Q_j|^{-\frac{1}{\lambda}} (2^j)^{-\frac{s}{r}+n\left(\frac{1}{\lambda}-\frac{1}{r}\right)}.$$

Let $\psi_\gamma^j = t_j\, b_\gamma^j$, where $t_j = (2^j)^{-s/r+n(1/\lambda-1/r)}$ and $b_\gamma^j(x) = (2^j)^{s/r-n(1/\lambda-1/r)} \psi_\gamma^j(x)$. Hence, we can write $\sum_{j=0}^\infty P_j(x) = \sum_{j=0}^\infty \sum_{|\gamma|\leq N_q} t_j\, b_\gamma^j(x)$, and for each $j \in \mathbb{N}_0$ the function $b_\gamma^j(x)$ is a $(q, \lambda, \infty)-$atom, since is supported on E_{j+1} and satisfies $|b_\gamma^j(x)| \lesssim |Q_j|^{-\frac{1}{\lambda}}$, as desired. Moreover from $s > n\,(r/q - 1)$ one has

$$\sum_{j=0}^\infty |t_j|^q |Q_j|^{1-\frac{q}{\lambda}} = |Q|^{1-\frac{q}{\lambda}} \sum_{j=0}^\infty (2^j)^{q\left(-\frac{s}{r}+n\left(\frac{1}{q}-\frac{1}{r}\right)\right)} \lesssim |Q|^{1-\frac{q}{\lambda}}.$$

□

Now we ready to announce a molecular decomposition in Hardy-Morrey spaces.

Theorem 3 *Let $\{m_Q\}_Q$ be a collection of L^r-molecules and $\{s_Q\}_Q$ be a sequence of complex numbers such that $\|\{s_Q\}_Q\|_{\lambda,q} < \infty$. If the series $f = \sum_Q s_Q m_Q$ converges in $\mathcal{S}'(\mathbb{R}^n)$ and $\lambda < r$, then $f \in \mathcal{HM}_q^\lambda(\mathbb{R}^n)$ and moreover, $\|f\|_{\mathcal{HM}_q^\lambda} \lesssim \|\{s_Q\}_Q\|_{\lambda,q}$ with implicit constant independent of f.*

Proof Suppose $f = \sum_Q s_Q m_Q$ in $\mathcal{S}'(\mathbb{R}^n)$ and $\|\{s_Q\}_Q\|_{\lambda,q} < \infty$. Since $0 < q \leq 1$, for a fixed dyadic cube $J \subset \mathbb{R}^n$ we may estimate $\int\!\!\!\!\!\!\int_J M_\varphi f(x)|^q dx$ by

$$\sum_{Q\subseteq J} |s_Q|^q \int_J |M_\varphi m_Q(x)|^q dx + \sum_{J\subset Q} |s_Q|^q \int_J |M_\varphi m_Q(x)|^q dx := I_1 + I_2.$$

Estimate of I_1 From Lemma 2, write $m_Q = \sum_{j=0}^\infty d_j a_{Q_j}$ (convergence in L^r) where $\{a_{Q_j}\}_j$ are $(q, \lambda, r)-$atoms and moreover $\sum_{j=0}^\infty |d_j|^q |Q_j|^{1-\frac{q}{\lambda}} \lesssim |Q|^{1-\frac{q}{\lambda}}$. It follows from analogous estimates of Proposition 1 that

$$\begin{aligned} I_1 &\lesssim \sum_{Q\subseteq J} |s_Q|^q \sum_{j=0}^\infty |d_{Q_j}|^q \int_J |M_\varphi a_{Q_j}(x)|^q dx \lesssim \sum_{Q\subseteq J} |s_Q|^q \sum_{j=0}^\infty |d_{Q_j}|^q |Q_j|^{1-\frac{q}{\lambda}} \\ &\lesssim \sum_{Q\subseteq J} |s_Q|^q |Q|^{1-\frac{q}{\lambda}} \lesssim |J|^{1-\frac{q}{\lambda}} \|\{s_Q\}_Q\|_{\lambda,q}^q. \end{aligned}$$

Estimate of I_2 Since $1 < r < \infty$ and M_φ is bounded on $L^r(\mathbb{R}^n)$, it follows

$$
\begin{aligned}
&|J|^{\frac{q}{\lambda}-1} \int_J |M_\varphi m_Q(x)|^q dx \\
&\quad \le |J|^{q\left(\frac{1}{\lambda}-\frac{1}{r}\right)} \left(\sum_{j=0}^{\infty} (2^j \ell_Q)^{-s} \int_{E_j} |m_Q(x)|^r |x - x_Q|^s dx \right)^{\frac{q}{r}} \\
&\quad \lesssim |J|^{q\left(\frac{1}{\lambda}-\frac{1}{r}\right)} |Q|^{q\left(\frac{1}{r}-\frac{1}{\lambda}\right)} \left(\sum_{j=0}^{\infty} 2^{-js} \right)^{\frac{q}{r}} \simeq \left(\frac{|J|}{|Q|} \right)^{q\left(\frac{1}{\lambda}-\frac{1}{r}\right)}.
\end{aligned}
$$

If $r = 1$ and $0<q<1$, we proceed like in (6) and then $|J|^{\frac{q}{\lambda}-1} \int\!\!\int_J |M_\varphi m_Q(x)|^q dx \lesssim$

$$
\begin{aligned}
&|J|^{\frac{q}{\lambda}-1} \left[|J| \int_0^{|Q|^{1-\frac{1}{\lambda}}|J|^{-1}} \omega^{q-1} d\omega + |Q|^{-1+\frac{1}{\lambda}} \int_{|Q|^{1-\frac{1}{\lambda}}|J|^{-1}}^{\infty} \omega^{q-2} d\omega \right] \\
&\quad \lesssim \left(\frac{|J|}{|Q|} \right)^{q\left(\frac{1}{\lambda}-1\right)}.
\end{aligned}
$$

Fixed a dyadic cube J, we point out that there exists a subset $N \subseteq \mathbb{N}$ such that each cube $J \subset Q$ is uniquely determined by a dyadic cube $Q_{k,J} \in \{Q \text{ dyadic} : J \subset Q \text{ and } \ell_Q = 2^k \ell_J\}$. Hence, we can write $\sum_{J \subset Q} |s_Q|^q \left(\frac{|J|}{|Q|} \right)^\gamma = \sum_{k \in N} |s_{Q_{k,J}}|^q \, 2^{-kn\gamma}$ with $\gamma := 1/\lambda - 1/r > 0$. Then,

$$
\begin{aligned}
|J|^{\frac{q}{\lambda}-1} \sum_{J \subset Q} |s_Q|^q \|M_\varphi(a_Q)\|_{L^q(J)}^q &\lesssim \sum_{k \in N} \left(|s_{Q_{k,J}}|^q |Q_{k,J}|^{1-\frac{q}{\lambda}} \right) |Q_{k,J}|^{\frac{q}{\lambda}-1} 2^{-kn\gamma q} \\
&\le \sum_{k \in N} \left(\sum_{Q \subseteq Q_{k,J}} |s_Q|^q |Q|^{1-\frac{q}{\lambda}} \right) |Q_{k,J}|^{\frac{q}{\lambda}-1} 2^{-kn\gamma q} \\
&\lesssim \| \{s_Q\}_Q \|_{\lambda,q}^q \sum_{k \in N} 2^{-kn\gamma q} \lesssim \| \{s_Q\}_Q \|_{\lambda,q}^q.
\end{aligned}
$$

□

Remark 2 The Theorem 3 covers [10, Theorem 2.6] when $r = 2$ where the natural restriction $\lambda < 2$ was omitted (see also Proposition 1).

3 Proof of Theorem 1

Proof Let a be a (q,λ,r)-atom supported in the cube Q. From Theorem 3, it suffices to show that Ta is a $(q,\lambda,s,r)-$ molecule associated to Q. Suppose first that $\ell_Q \geq 1$. Since T is bounded in $L^2(\mathbb{R}^n)$ to itself and $1 \leq r \leq 2$, condition (M_1) follows by

$$\int_{2Q} |Ta(x)|^r dx \leq |2Q|^{1-\frac{r}{2}} \|Ta\|_{L^2}^r \lesssim |Q|^{1-\frac{r}{2}} \|a\|_{L^2}^r \lesssim |Q|^{1-\frac{r}{\lambda}} \simeq \ell_Q^{n\left(1-\frac{r}{\lambda}\right)}. \tag{12}$$

For (M_2) using the moment condition of the atom a, Minkowski inequality and (3), we estimate $\int_{2Q^c} |Ta(x)|^r |x-x_Q|^s dx$ by

$$\begin{aligned}
&\sum_{j=1}^{\infty} \int_{C_j(x_Q,\ell_Q)} \left| \int_Q [K(x,y)-K(x,x_Q)]a(y)dy \right|^r |x-x_Q|^s dx \\
&\leq \sum_{j=1}^{\infty} (2^j \ell_Q)^s \left\{ \int_Q |a(y)| \left[\int_{C_j(x_Q,\ell_Q)} |K(x,y)-K(x,x_Q)|^r dx \right]^{\frac{1}{r}} dy \right\}^r \\
&\lesssim \sum_{j=1}^{\infty} (2^j \ell_Q)^{s-n(r-1)}\ 2^{-jr\delta}\ \ell_Q^{rn\left(1-\frac{1}{\lambda}\right)} \\
&\simeq \ell_Q^{r+n\left(1-\frac{r}{\lambda}\right)} \sum_{j=1}^{\infty} 2^{j[s-n(r-1)-r\delta]} \simeq \ell_Q^{s+n\left(1-\frac{r}{\lambda}\right)},
\end{aligned}$$

assuming $s < n(r-1)+r\delta$. We remark that for the case $r=1$, one needs to consider $(q,\lambda,s,1)-$molecules and hence $0 < q \leq \lambda < 1$. Suppose now that $\ell_Q < 1$. Since T is a bounded operator from $L^p(\mathbb{R}^n)$ to $L^2(\mathbb{R}^n)$ and $1 < r \leq 2$, condition (M_1) follows by

$$\begin{aligned}
\int_{2Q} |Ta(x)|^r dx \leq |2Q|^{1-\frac{r}{2}} \|Ta\|_{L^2}^r \lesssim |Q|^{1-\frac{r}{2}} \|a\|_{L^p}^r \lesssim |Q|^{1-\frac{r}{\lambda}+r\left(\frac{1}{p}-\frac{1}{2}\right)} \\
\lesssim |Q|^{1-\frac{r}{\lambda}}.
\end{aligned}$$

To estimate the global (M_2) condition, we consider $0 < \rho \leq \sigma \leq 1$ a parameter that will be chosen conveniently later, denote by $2Q^\rho := Q(x_Q, 2\ell_Q^\rho)$ and split the integral over $\mathbb{R}^n$ into $2Q^\rho$ and $(2Q^\rho)^c$. For $2Q^\rho$ we use the boundedness from

$L^p(\mathbb{R}^n)$ to $L^2(\mathbb{R}^n)$ again and obtain

$$\int_{2Q^\rho} |Ta(x)|^r \, |x - x_Q|^s dx \lesssim \ell_Q^{s\rho} |4Q^\rho|^{1-\frac{r}{2}} \|Ta\|_{L^2}^r \lesssim \ell_Q^{\rho s + n\rho\left(1-\frac{r}{2}\right)} \|a\|_{L^p}^r$$
$$\lesssim \ell_Q^{\rho s + n\left[\rho - \frac{r\rho}{2} + r\left(\frac{1}{p} - \frac{1}{\lambda}\right)\right]} \lesssim \ell_Q^{s + n\left(1-\frac{r}{\lambda}\right)},$$

assuming $s \leq -n\left(1 - \frac{r}{2}\right) + \frac{nr}{1-\rho}\left(\frac{1}{p} - \frac{1}{2}\right)$. For $(2Q^\rho)^c$, we use (4) to estimate $\int_{(2Q^\rho)^c} |Ta(x)|^r |x - x_Q|^s dx$ by

$$\sum_{j=1}^{\infty} (2^j \ell_Q^\rho)^s \left\{ \int_Q |a(y)| \left[\int_{C_j(x_Q, \ell_Q^\rho)} |K(x,y) - K(x,x_Q)|^r dx \right]^{\frac{1}{r}} dy \right\}^r$$
$$\lesssim \sum_{j=1}^{\infty} (2^j \ell_Q^\rho)^s \left(|C_j(x_Q, \ell_Q^\rho)|^{\frac{1}{r} - 1 + \frac{\delta}{n}\left(\frac{1}{\rho} - \frac{1}{\sigma}\right)} 2^{-\frac{j\delta}{\rho}} \right)^r \ell_Q^{rn\left(1-\frac{1}{\lambda}\right)}$$
$$\simeq \ell_Q^{\rho s + n\left[r + \frac{r\delta}{n} - r\rho\left(1 - \frac{1}{r} + \frac{\delta}{n\sigma}\right) - \frac{r}{\lambda}\right]} \sum_{j=1}^{\infty} 2^{j\left[s - n(r-1) - \frac{r\delta}{\sigma}\right]}$$
$$\lesssim \ell_Q^{\rho s + n\left[\rho\left(1-\frac{r}{2}\right) + r\left(\frac{1}{p} - \frac{1}{\lambda}\right)\right]} \leq \ell_Q^{s + n\left(1-\frac{r}{\lambda}\right)},$$

where the convergence follows assuming $s < n(r-1) + \frac{r\delta}{\sigma}$ and we choose ρ to be such that $r + \frac{r\delta}{n} - \rho\left(r - 1 + \frac{r\delta}{n\sigma}\right) = \rho\left(1 - \frac{r}{2}\right) + \frac{r}{p} \Leftrightarrow \rho := \frac{n\left(1 - \frac{1}{p}\right) + \delta}{\frac{n}{2} + \frac{\delta}{\sigma}}$. By the choice of ρ we have

$$-n\left(1 - \frac{r}{2}\right) + \frac{nr}{1-\rho}\left(\frac{1}{p} - \frac{1}{2}\right) < n(r-1) + r\delta < n(r-1) + \frac{r\delta}{\sigma}.$$

In particular, collecting the restrictions on s we get

$$n\left(\frac{r}{q} - 1\right) < s \leq -n\left(1 - \frac{r}{2}\right) + \frac{nr}{1-\rho}\left(\frac{1}{p} - \frac{1}{2}\right)$$
$$\Rightarrow \quad \frac{1}{q} < \frac{1}{2} + \frac{\beta\left(\frac{\delta}{\sigma} + \frac{n}{2}\right)}{n\left(\frac{\delta}{\sigma} - \delta + \beta\right)} = \frac{1}{q_0}.$$

We point out that when $\sigma = 1$, only condition $s < n(r-1) + r\delta$ is imposed to verify (M_1) and (M_2). Condition (M_3), given formally by $T^*(x^\alpha) = 0$ for all $|\alpha| \leq N_q$, is trivially valid, since $n/(n+\delta) < q_0 < q \leq 1$ implies $N_q \leq \lfloor \delta \rfloor$. □

Remark 3 The previous proof remains the same if one consider integral conditions incorporating derivatives of the kernel. For a complete discussion and the precise definition of such conditions we refer [12, Section 4.2].

References

1. Álvarez, J., Milman, M.: H^p Continuity properties of Calderón-Zygmund-type operators. J. Math. Anal. Appl. **118**, 63–79 (1986)
2. Álvarez, J., Milman, M.: Vector valued inequalities for strongly singular Calderón-Zygmund operators. Rev. Mat. Iber. **2**, 405–426 (1986)
3. Coifman, R., Meyer, Y.: Au-delà des opérateurs pseudo-différentiels. Astérisque, no. 57, 210 p. (1978)
4. D'Abbicco, M., Ebert, M., Picon, T.: The critical exponent(s) for the semilinear fractional diffusive equation. J. Fourier Anal. Appl. **25**, 696–731 (2019)
5. de Almeida, M., Picon, T.: Fourier transform decay of distributions in Hardy-Morrey spaces. https://arxiv.org/abs/2011.10176 (2021)
6. Fefferman, C.: Inequalities for strongly singular convolution operators. Acta Math. **124**(1), 9–36 (1970)
7. Garcia-Cuerva, J., Rubio de Francia, J.L.: Weighted Norm Inequalities and Related Topics. North-Holland Math. Stud. 116. North-Holland, Amsterdam (1985)
8. Hirschman, I.: On multiplier transformations. Duke Math. J. **26**(2), 221–242 (1959)
9. Jia, H., Wang, H.: Decomposition of Hardy-Morrey spaces. J. Math. Anal. Appl. **354**(1), 99–110 (2009)
10. Jia, H., Wang, H.: Singular integral operator, Hardy-Morrey space estimates for multilinear operators and Navier-Stokes equations. Math. Methods Appl. Sci. **33**(14), 1661–1684 (2010)
11. Stein, E.: Harmonic Analysis: Real-variable Methods, Orthogonality, and Oscillatory Integrals. Princeton Mathematical Series, 43. Princeton University Press, Princeton, NJ (1993)
12. Picon, T., Vasconcelos, C.: On the continuity of strongly singular Calderón-Zygmund-type operators on Hardy spaces. Integral Equ. Oper. Theory **95**(9), (2023)
13. Wainger, S.: Special Trigonometric Series in k Dimensions. Memoirs of the American Mathematical Society, no. 59. American Mathematical Society (1965)

The Asymptotic Estimates of the Solutions to the Linear Damping Models with Spatial Dependent Coefficients

Pham Trieu Duong

Abstract We study the Cauchy problem

$$u_{tt} + a_1(x)(-\Delta)^{\sigma} u + a_2(x)u_t = 0,\ t > 0,\ x \in \mathbb{R}^n,$$
$$u(0, x) = u_0(x), u_t(0, x) = u_1(x), \quad x \in \mathbb{R}^n,$$

with $\sigma \in (0, 1)$, where the coefficients $a_1(x), a_2(x)$ are continuous functions of the spatial variable x. We will derive the decay estimates for the solutions for this linear problem, as well as for the solution of the corresponding Cauchy - Dirichlet problem in the exterior domain $\Omega \subset \mathbb{R}^n$.

1 Introduction

The damping models of the form

$$u_{tt} + (-\Delta)^{\sigma} u + \mu(-\Delta)^{\delta} u_t = F(u, |D|^{\alpha} u, u_t), \quad u(0, x) = u_0(x),$$
$$u_t(0, x) = u_1(x),$$

have been studied by many authors (see [6, 11]). The diffusion phenomenon for damped wave equations with time dependent coefficient $a(t)$ was obtained by Wirth in [14, 15]. In [5] D'Abbicco and Ebert study the asymptotic profiles of the solution to the Cauchy problem for the linear plate equation with a decreasing coefficient $\lambda(t)$. Recently, in [10], the authors considered the following Cauchy-Dirichlet problem

$$\begin{aligned} &u_{tt} + a_1(x)(-\Delta)^{\sigma} u + a u_t = 0, \quad t > 0, \quad x \in \Omega, \\ &u(t, x) = 0, \quad t > 0, \quad x \in \partial\Omega, \\ &u(0, x) = u_0(x), \quad u_t(0, x) = u_1(x), \quad x \in \Omega, \end{aligned} \tag{1}$$

P. T. Duong (✉)
Department of Mathematics, Hanoi National University of Education, Hanoi, Vietnam
e-mail: duongptmath@hnue.edu.vn

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_45

where $a = \text{const} > 0$, Ω is the exterior domain in $\mathbb{R}^n$: $\Omega \equiv \mathbb{R}^n \setminus K$, and obtained the linear estimates for both cases, when $K = \emptyset$ and $K \neq \emptyset$. In the above models, the fractional Laplacian $(-\Delta)^\sigma$ for $\sigma \in (0, 1)$ can be defined as

$$(-\Delta)^\sigma u(x) = C \int_{\mathbb{R}^n} \frac{u(x) - u(y)}{|x - y|^{n+2\sigma}} dy$$

for sufficiently smooth u with a positive normalization constant

$$C := C_{n,\sigma} = \frac{2^{2\sigma}\sigma\Gamma(\frac{n}{2}+\sigma)}{\pi^{\frac{n}{2}}\Gamma(1-\sigma)}.$$

The main tools for deriving decay estimates of the model (1) is construction of the diffusion phenomenon and transferring the decay rate of the solution of the evolution equation

$$\begin{cases} \rho(x)v_t + (-\Delta)^\sigma v = 0, \quad t > 0, \ x \in \Omega \\ v(t, x) = 0, \quad t > 0, \ x \in \partial\Omega \\ v_t(0, x) = v_0(x), \ x \in \Omega \end{cases} \tag{2}$$

with suitable $\rho(x)$ and $v_0(x)$, to the corresponding rate of the solution of problem (1). It should be noted that the condition $a = \text{const}$ is quite essential in [10] in order to get the desired diffusion phenomenon.

This article is devoted to the study of the models

$$\begin{cases} u_{tt} + a_1(x)(-\Delta)^\sigma u + a_2(x)u_t = 0, \quad (t, x) \in [0, \infty) \times \mathbb{R}^n, \\ u(0, x) = u_0(x), \ u_t(0, x) = u_1(x), \quad x \in \mathbb{R}^n, \end{cases} \tag{3}$$

and

$$\begin{cases} u_{tt} + a_1(x)(-\Delta)^\sigma u + a_2(x)u_t = 0, \quad (t, x) \in [0, \infty) \times \Omega \\ u(t, x) = 0, \quad t > 0, \quad x \in \partial\Omega, \\ u(0, x) = u_0(x), \quad u_t(0, x) = u_1(x), \quad x \in \Omega, \end{cases} \tag{4}$$

where both $a_1(x)$ and $a_2(x)$ are depending on x variables. We will show that some reasonable decay rates of solutions for problems (3) and (4) are still available and they can be obtained by the notion of the generalized diffusion phenomenon that has been introduced by Radu et al. in [13].

2 Main Results

We study the linear Cauchy problem (3) in $[0,\infty)\times\mathbb{R}^n$, where the coefficient $a_1 = a_1(x), a_2 = a_2(x)$ are continuous functions on $\mathbb{R}^n$. Moreover, we assume that there exist positive constants $c_i, i = 1,2,3,4$, such that

$$c_1 \le a_1(x) \le c_2, \quad c_3 \le a_2(x) \le c_4, \quad \forall x \in \mathbb{R}^n. \tag{5}$$

The linear estimates of the solution for problem (3) are stated in the following.

Theorem 1 *Consider the linear problem (3) in $(0,\infty)\times\mathbb{R}^n$ with $\sigma \in (0,1)$, $(u_0,u_1) \in \big(H^\sigma(\mathbb{R}^n)\cap L^1(\mathbb{R}^n)\big)\times\big(L^2(\mathbb{R}^n)\cap L^1(\mathbb{R}^n)\big)$. Assume that $n > 2\sigma$, $a_i = a_i(x), i = 1,2$ satisfy (5). Let u be the unique weak solution to (3). Then the following estimates*

$$\begin{aligned}\|u(t)\|_2 \lesssim{}& \big(\|u_0\|_2+\|u_0\|_{\dot H^\sigma}+\|u_1\|_2+\|u_0\|_q+\|u_1\|_q\big)(t+1)^{-\frac{n}{2\sigma}(1/q-1/2)-1}+\\ &+\big(\|u_0\|_2+\|u_1\|_2+\|u_0\|_1+\|u_1\|_1\big)(t+1)^{-\frac{n}{4\sigma}},\end{aligned} \tag{6}$$

$$\begin{aligned}\|\partial_t u(t)\| \lesssim{}& (\|u_0\|_2+\|u_0\|_{\dot H^\sigma}+\|u_1\|_2+\|u_0\|_1+\|u_1\|_1)(t+1)^{-\frac{n}{4\sigma}-1}\\ &+(\|u_0\|_2+\|u_0\|_{\dot H^\sigma}+\|u_1\|_2+\|u_0\|_q+\|u_1\|_q)(t+1)^{-\frac{n}{2\sigma}(1/q-1/2)-3/2},\end{aligned} \tag{7}$$

$$\begin{aligned}\|u(t)\|_{\dot H^\sigma} \lesssim{}& (\|u_0\|_2+\|u_0\|_{\dot H^\sigma}+\|u_1\|_2+\|u_0\|_1+\|u_1\|_1)(t+1)^{-\frac{n}{4\sigma}-1}\\ &+(\|u_0\|_2+\|u_0\|_{\dot H^\sigma}+\|u_1\|_2+\|u_0\|_q+\|u_1\|_q)(t+1)^{-\frac{n}{2\sigma}(1/q-1/2)-3/2}.\end{aligned} \tag{8}$$

hold for $q \in (1,2]$ and for all $t > 0$.

We also are interested in the decay estimates of solutions for the Cauchy-Dirichlet problem (4) in $[0,\infty)\times\Omega$. In this problem the exterior domain Ω is defined as $\Omega = \mathbb{R}^n \setminus K$, where K is a compact with sufficiently smooth ∂K. The linear estimates of the solution for problem (4) are contained in the following theorem.

Theorem 2 *Consider the linear Cauchy-Dirichlet problem (4) with $\sigma \in (0,1)$ and $n > 2\sigma$ with the compactly supported data $((u_0,u_1) \in \big(\dot H^\sigma(\overline{\Omega})\cap L^1(\Omega)\big)\times\big(L^2(\Omega)\cap L^1(\Omega)\big)$. Assume that the coefficients $a_i = a_i(x), i = 1,2$, satisfy (5). Let $(u,\partial_t u)\in C\big((0,\infty),\dot H^\sigma(\overline{\Omega})\times L^2(\Omega)\big)$ be the unique weak solution to (4) and B be the self-adjoint operator defined through (19)–(23) and the Dirichlet condition. Then the following estimates*

$$\|u(t)\|_2 \lesssim \big(\|u_0\|_2+\|u_0\|_{\dot H^\sigma}+\|u_1\|_2+\|u_0\|_q+\|u_1\|_q\big)(t+1)^{-\frac{n}{2\sigma}(1/q-1/2)-1}$$

$$+\big(\|u_0\|_2+\|u_1\|_2+\|u_0\|_1+\|u_1\|_1\big)(t+1)^{-\frac{n}{4\sigma}}, \tag{9}$$

$$\begin{aligned}\|\partial_t u(t)\|_2 \lesssim &(\|u_0\|_2+\|u_0\|_{\dot H^\sigma}+\|u_1\|_2+\|u_0\|_1+\|u_1\|_1)(t+1)^{-\frac{n}{4\sigma}-1}\\ &+(\|u_0\|_2+\|u_0\|_{\dot H^\sigma}+\|u_1\|_2+\|u_0\|_q+\|u_1\|_q)(t+1)^{-\frac{n}{2\sigma}(1/q-1/2)-3/2},\end{aligned} \tag{10}$$

$$\begin{aligned}\|\sqrt{B}u(t)\|_2 \lesssim &(\|u_0\|_2+\|u_0\|_{\dot H^\sigma}+\|u_1\|_2+\|u_0\|_1+\|u_1\|_1)(t+1)^{-\frac{n}{4\sigma}-1}\\ &+(\|u_0\|_2+\|u_0\|_{\dot H^\sigma}+\|u_1\|_2+\|u_0\|_q+\|u_1\|_q)(t+1)^{-\frac{n}{2\sigma}(1/q-1/2)-3/2},\end{aligned} \tag{11}$$

hold for $q \in (1, 2]$ *and for all* $t > 0$.

3 The Generalized Diffusion Phenomenon

In this section we recall the notion of the generalized diffusion phenomenon that has been well studied by Radu et al. in [13] in the abstract setting with two non-commuting self-adjoin operators.

Let $B : \mathcal{D}(B) \to \mathcal{H}$ and $C : \mathcal{H} \to \mathcal{H}$ be two self-adjoint nonnegative definite operators on the real Hilbert space $(\mathcal{H}, \|\cdot\|)$. We consider the Cauchy problem in $(0, \infty)$

$$C\partial_t^2 u + \partial_t u + Bu = 0, \quad u(0) = u_0, \quad \partial_t u(0) = u_1, \tag{12}$$

where $(u_0, u_1) \in \mathcal{D}(\sqrt{B}) \times \mathcal{H}$. The generalized diffusion phenomenon is described by the approximation $u(t) \approx e^{-tB}(u_0 + Cu_1), \quad t \to \infty$.

In more details, let us introduce the following conditions on operators B and C.

$$\begin{aligned}&\text{(H1)} \quad \mathcal{D}(B) \text{ is dense in } \mathcal{H} \text{ and } C \text{ is a bounded operator on } \mathcal{H};\\ &\text{(H2)} \quad \langle Bu, u\rangle > 0 \text{ for } u \in \mathcal{D}(B) \text{ and } u \neq 0;\\ &\text{(H3)} \quad C_1\|u\|^2 \geq \langle Cu, u\rangle \geq C_0\|u\|^2 \text{ for } u \in \mathcal{H}, \text{ where } C_1 \geq C_0 > 0.\end{aligned} \tag{13}$$

In [13] it was proved that conditions $(H1)$–$(H3)$ are sufficient to ensure the existence and uniqueness of mild solutions $(u, \partial_t u) \in C(\mathbb{R}_+, \mathcal{D}(B) \times \mathcal{H})$ for problem (12) (see Appendix A in [13]).

Let $\mathcal{H} = L^2(\Omega, \mu)$, where (Ω, μ) is a $\sigma-$ finite measure space. To obtain the generalized diffusion phenomenon for problem (12), we assume further that the next

conditions on B and C are satisfied.

(H4) $-B$ generates a Markov semigroup $\{e^{-tB}\}_{t\geq 0}$ on $L^q(\Omega,\mu)$, $q\in[1,2]$.

(H5) $\exists m>0$ such that $\|e^{-tB}g\|_2 \leq c_q t^{-m/2(1/q-1/2)}(\|g\|_q+\|g\|_2)$, for $g\in L^q(\Omega,\mu)\cap L^2(\Omega,\mu)$, $t>0$, $q\in[1,2]$. (14)

(H6) C is a bounded operator $L^q(\Omega,\mu)\to L^q(\Omega,\mu)$ for all $q\in[1,2]$.

Recall that the semigroup $\{e^{-tB}\}_{t>0}$ on $L^1(\mathcal{X},\mu)$, where $(\mathcal{X},\mathcal{A},\mu)$ is a σ-finite measure space, is said to be Markov if

$$f\in L^1(\mathcal{X},\mu),\ f\geq 0 \ \text{ implies } \ e^{-tB}f\geq 0 \ \text{ and } \ \|e^{-tB}f\|_{L^1}\leq \|f\|_{L^1} \tag{15}$$

for all $t>0$. The property (15) will be called also *the Markov property* of the semigroup. The norm $\|\cdot\|_1$ is the norm in $L^1(\mathcal{X},\mu)$, meanwhile $f\geq 0$ holds almost everywhere with respect to the measure μ.

We introduce the energy $E_v(s)$ associated with $(v,\partial_s v)\in C(\mathbb{R}_+,\mathcal{D}(B)\times\mathcal{H})$:

$$E_v(s)=\frac{1}{2}\big(\|\sqrt{C}\partial_s v(s)\|^2+\|\sqrt{B}v(s)\|^2\big),\quad s\geq 0.$$

The decay rates of the solution for problem (12) obtained by the generalized diffusion phenomenon are described as follows.

Proposition 1 (Corollary 1.5 in [13]) *Assume that (H1)–(H6) are satisfied and let* $(u,\partial_t u)\in C(\mathbb{R}_+,\mathcal{D}(\sqrt{B})\times\mathcal{H})$ *be the unique mild solution of* (12). *If* $q\in(1,2]$, *then*

$$\begin{aligned}\|u(t)-e^{-tB}(u_0+Cu_1)\|_2 &\\ \lesssim (\|u_0\|_2+\|\sqrt{B}u_0\|_2+\|u_1\|_2+\|u_0\|_q+\|u_1\|_q)&(t+1)^{-m/2(1/q-1/2)-1},\end{aligned} \tag{16}$$

$$\begin{aligned}\|u(t)\|_2 \lesssim (\|u_0\|_2+\|u_1\|_2+\|u_0\|_1+\|u_1\|_1)&(t+1)^{-m/4}\\ +(\|u_0\|_2+\|\sqrt{B}u_0\|_2+\|u_1\|_2+\|u_0\|_q+\|u_1\|_q)&(t+1)^{-m/2(1/q-1/2)-1},\end{aligned} \tag{17}$$

$$\begin{aligned}E_u^{1/2}(t) \lesssim (\|u_0\|_2+\|\sqrt{B}u_0\|_2+\|u_1\|_2+\|u_0\|_1+\|u_1\|_1)&(t+1)^{-m/4-1}\\ +(\|u_0\|_2+\|\sqrt{B}u_0\|_2+\|u_1\|_2+\|u_0\|_q+\|u_1\|_q)&(t+1)^{-m/2(1/q-1/2)-3/2}.\end{aligned} \tag{18}$$

4 Proofs of Theorems 1 and 2

Since the proofs Theorems 1 and 2 will follow almost exactly same ways, if we consider the case $\Omega = \mathbb{R}^n$ as a special case of the exterior domain $\Omega = \mathbb{R}^n \setminus K$, with $K = \emptyset$, we will present below the proof of Theorem 2. In order to apply the generalized diffusion phenomenon in the abstract setting, we will construct the Hilbert space $\mathcal{H}$ and the corresponding self-adjoint operators B and C as follows.

We denote $\rho(x) := \dfrac{a_2(x)}{a_1(x)}$. Let $\mathcal{H}$ be the space $L^2(\Omega, \rho(x)dx)$. Thanks to the condition (5) on $a_i(x), i = 1, 2$, the norm $\|\cdot\|_q$ in $L^q_\rho := L^q(\Omega, \rho(x)dx)$ is equivalent to the usual L^q- norm in $L^q(\Omega)$, for all $q \in [1, \infty]$. Thus for simplicity, we will omit ρ in the notation of L^q_ρ, whenever the norms $\|\cdot\|_q$ are involved.

The self-adjoint operator C is chosen as $C := \dfrac{1}{a_2(x)}$, that means $Cu = \dfrac{1}{a_2(x)}u$. As for the operator B, we will chose it to be

$$B = \frac{a_1(x)}{a_2(x)}(-\Delta)^\sigma. \tag{19}$$

More precisely, consider $B_0 u := \dfrac{a_1(x)}{a_2(x)}(-\Delta)^\sigma u$ for $u \in C_0^\infty(\Omega)$. We will show that B_0 is a symmetric nonnegative operator in $\mathcal{H}$. Indeed, for $u, v \in C_0^\infty(\Omega)$ we have

$$\begin{aligned}
(B_0u, v)_{\mathcal{H}} &= ((-\Delta)^\sigma u, v)_{L^2(\Omega)} = C\int_\Omega\int_{\mathbb{R}^n} \frac{u(x)-u(y)}{|x-y|^{n+2\sigma}}v(x)dydx \\
&= C\int_\Omega\int_{\mathbb{R}^n} \frac{(u(x)-u(y))(v(x)-v(y))}{|x-y|^{n+2\sigma}}dydx + C\int_\Omega\int_{\mathbb{R}^n} \frac{u(x)-u(y)}{|x-y|^{n+2\sigma}}v(y)dydx \\
&= C\int_\Omega\int_{\mathbb{R}^n} \frac{(u(x)-u(y))(v(x)-v(y))}{|x-y|^{n+2\sigma}}dydx + C\int_\Omega\int_{\Omega} \frac{u(x)-u(y)}{|x-y|^{n+2\sigma}}v(y)dydx,
\end{aligned} \tag{20}$$

due to $v \in C_0^\infty(\Omega)$.

Renaming the variables x, y by y, x in the last double integral we obtain

$$\begin{aligned}
(B_0u, v)_{\mathcal{H}} = \; & C\int_\Omega\int_{\mathbb{R}^n} \frac{(u(x)-u(y))(v(x)-v(y))}{|x-y|^{n+2\sigma}}dydx \\
& - C\int_\Omega\int_{\Omega} \frac{u(x)-u(y)}{|x-y|^{n+2\sigma}}v(x)dydx
\end{aligned}$$

$$= C \int_{\Omega} \int_{\mathbb{R}^n} \frac{(u(x)-u(y))(v(x)-v(y))}{|x-y|^{n+2\sigma}} dydx$$
$$- C \int_{\Omega} \int_{\mathbb{R}^n} \frac{u(x)-u(y)}{|x-y|^{n+2\sigma}} v(x) dydx$$
$$+ C \int_{\Omega} \int_{\mathbb{R}^n \setminus \Omega} \frac{u(x)-u(y)}{|x-y|^{n+2\sigma}} v(x) dydx. \tag{21}$$

Again, thanks to $u \in C_0^\infty(\Omega)$, from (21) it follows that

$$(B_0 u, v)_{\mathcal{H}} = \frac{C}{2} \Big[\int_{\Omega} \int_{\mathbb{R}^n} \frac{(u(x)-u(y))(v(x)-v(y))}{|x-y|^{n+2\sigma}} dydx$$
$$+ \int_{\Omega} dx \int_{\mathbb{R}^n \setminus \Omega} \frac{u(x)v(x)}{|x-y|^{n+2\sigma}} dy \Big], \tag{22}$$

which implies $(B_0 u, v)_{\mathcal{H}} = (u, B_0 v)_{\mathcal{H}}$ and $(B_0 u, u)_{\mathcal{H}} \geq 0$ for all $u, v \in C_0^\infty(\Omega)$.

Now we define B as the Friedrichs extension for the densely defined symmetric nonnegative operator B_0 in $\mathcal{H}$. Then B itself is a nonnegative self-adjoint operator, with

$$\mathcal{D}(B) = V \cap \{u \in \mathcal{H} : B_0 u \in \mathcal{H}\}, \tag{23}$$

where V is the completion of $C_0^\infty(\Omega)$ with respect to the norm

$$u \mapsto \Big((B_0 u, u)_{\mathcal{H}} + \|u\|_{\mathcal{H}}^2 \Big)^{\frac{1}{2}}.$$

The square root $\sqrt{B}$ of B and other powers B^k are defined by operator calculus.

Now, let us verify the validation of conditions $(H1)$–$(H6)$.

Conditions $(H1)$–$(H3)$ and $(H6)$ are obviously satisfied, thanks to the boundedness assumptions on coefficients $a_i(x)$, $i = 1, 2$.

Condition $(H4)$ can be verified by the well-known approach proposed in [2, 7, 12] to apply the realization of the fractional Laplacian $(-\Delta)^\sigma$ through the $2\sigma-$harmonic extension to the upper half-space. Indeed, the Markov property of the semigroup $\{e^{-tB}\}_{t\geq 0}$, that is condition $(H4)$, follows from Propositions 5.12 and 5.18 in [10] that summarize the most important properties of solutions, including

positivity and L^1_ρ-contraction, to the evolution model

$$\begin{cases} \rho(x)v_t + (-\Delta)^\sigma v = 0, & (t,x) \in (0,\infty)\times\Omega, \\ v(0,x) = v_0(x), & x \in \Omega, \\ v(t,x) = 0, & (t,x) \in (0,\infty)\times\partial\Omega, \end{cases} \tag{24}$$

under the following assumptions on the density $\rho = \rho(x)$ and the initial condition $v_0 = v_0(x)$

$$\begin{cases} \rho \in C(\mathbb{R}^n),\ \ \rho > 0 \text{ in } \ \mathbb{R}^n, \\ v_0 \in L^\infty(\mathbb{R}^n) \cap L^+_\rho(\mathbb{R}^n), \\ 0 < \sigma < 1, \end{cases} \tag{$\mathbf{A}_0$}$$

where $L^+_\rho(\mathbb{R}^n) := \Big\{ f \in L^1_\rho(\mathbb{R}^n) : f \geq 0 \Big\}$.

Since $\{e^{-tB}\}_{t\geq 0}$ is also a symmetric semigroup of contraction in $L^2_\rho(\Omega)$, by interpolation between

$$\|e^{-tB} f\|_{L^2_\rho} \leq \|f\|_{L^2_\rho}, \quad \text{and} \quad \|e^{-tB} f\|_{L^1_\rho} \leq \|f\|_{L^1_\rho},$$

a standard duality argument allows the possibility to extend the semigroup $\{e^{-tB}\}_{t\geq 0}$ to all $L^q_\rho(\Omega)$ with $q \geq 1$, such that

$$\|e^{-tB} f\|_{L^q_\rho} \leq \|f\|_{L^q_\rho}.$$

Hence we obtain a symmetric Markov semigroup $\{e^{-tB}\}_{t\geq 0}$ on $L^q_\rho(\Omega)$, for all $q \in [1,\infty)$. Thus condition $(H4)$ is verified.

The validity of condition $(H5)$ is a consequence of the following lemmas.

Lemma 1 *For all functions $f \in L^1(\Omega) \cap H^\sigma_0(\Omega)$ the following estimate holds:*

$$\|f\|_{L^2}^{2+\frac{4\sigma}{n}} \lesssim (Bf, f)_{L^2_\rho} \|f\|_{L^1}^{\frac{4\sigma}{n}}. \tag{25}$$

Proof In the case $\Omega \equiv \mathbb{R}^n$ the statement of this result follows from Hölder's and the fractional Sobolev inequalities. Indeed, by Hölder's inequality

$$\|f\|_{L^2}^2 \leq \|f\|_{L^1}^{\frac{1}{p}} \|f\|_{L^{\frac{2p-1}{p-1}}}^{\frac{2p-1}{p}},$$

choosing the parameter $p = \frac{n+2\sigma}{4\sigma}$ and then raising the last inequality to $1 + \frac{2\sigma}{n}$, we derive

$$\|f\|_{L^2}^{2+\frac{4\sigma}{n}} \leq \|f\|_{L^1}^{\frac{4\sigma}{n}} \|f\|^2_{L^{\frac{2n}{n-2\sigma}}}.$$

In order to estimate the norm $\|f\|_{L^{\frac{2n}{n-2\sigma}}}$ by the scalar product $(Bf, f)^{\frac{1}{2}}_{L^2_\rho}$, we apply the following well-known Sobolev inequality (see [8])

$$\|f\|^p_{L^{p^\star}(\mathbb{R}^n)} \leq C_1 \int\limits_{\mathbb{R}^n} \int\limits_{\mathbb{R}^n} \frac{|f(x) - f(y)|^p}{|x-y|^{n+sp}} dxdy,$$

that is valid for $s \in (0,1)$, $p \geq 1, sp < n$ and $p^\star = \frac{np}{n-sp}$, with a constant $C_1 = C_1(n, s, p)$. If we chose $s = \sigma$, $p = 2$, then the above fractional Sobolev inequality with $p^\star = \frac{2n}{n-2\sigma}$ implies $\|f\|_{L^{\frac{2n}{n-2\sigma}}} \lesssim (Bf, f)^{\frac{1}{2}}_{L^2_\rho}$ from the definition of B.

The case when $K \neq \emptyset$ requires an attention, since we are working with the non-local fractional Laplacian $(-\Delta)^\sigma$, that means we need to estimate the following scalar product in $L^2_\rho(\Omega)$

$$(Bf, f)_{\mathcal{H}} = C \int\limits_{\Omega} \int\limits_{\mathbb{R}^n} \frac{(f(x) - f(y))f(x)}{|x-y|^{n+2\sigma}} dydx, \tag{26}$$

where $C = \text{const}$, for $f \in D(B)$.

Let us denote $I = \int\limits_{\Omega} \int\limits_{\mathbb{R}^n} \frac{(f(x) - f(y))f(x)}{|x-y|^{n+2\sigma}} dydx$. By (22) we obtain

$$\begin{aligned} 2I &= \int\limits_{\Omega} \int\limits_{\mathbb{R}^n \setminus \Omega} \frac{f^2(x)}{|x-y|^{n+2\sigma}} dydx + \int\limits_{\Omega} \int\limits_{\mathbb{R}^n} \frac{(f(x) - f(y))(f(x) - f(y))}{|x-y|^{n+2\sigma}} dydx \\ &\geq \int\limits_{\Omega} \int\limits_{\mathbb{R}^n} \frac{(f(x) - f(y))(f(x) - f(y))}{|x-y|^{n+2\sigma}} dydx \\ &\geq \int\limits_{\Omega} \int\limits_{\Omega} \frac{(f(x) - f(y))(f(x) - f(y))}{|x-y|^{n+2\sigma}} dydx, \end{aligned} \tag{27}$$

since all integrands are non-negative.

Now we apply the following version of fractional Sobolev inequalities for domains (see Thm. 1.1 in [9] and Thm. 6.7 in [8] for reference)

$$\iint_{\Omega \times \Omega} \frac{|f(x) - f(y)|^p}{|x-y|^{n+2\sigma}} dydx \geq C_{n,p,\sigma} \Big(\int_\Omega |f(x)|^{p^\star} dx \Big)^{p/p^\star},$$

that is valid for all open $\Omega \subset \mathbb{R}^n$, $p \geq 2, n \geq 2, 0 < \sigma < 1, n > p\sigma$ and $f \in \overset{\circ}{W}{}^{\sigma}_{p}(\Omega)$. With the choice $p = 2$, the above inequality combined with (27) implies the estimate for (Bf, f), and thus, the statement of Lemma 1 in the case $\Omega \neq \mathbb{R}^n$ as well, after applying Hölder's inequality as in the case $\Omega \equiv \mathbb{R}^n$.

The following ultracontractivity result is classical and can be found in [1, 3, 4]

Lemma 2 *Assume that $\{e^{-tB}\}_{t\geq 0}$ is a symmetric Markov semigroup on $L^q(\Omega, \mu), q \in [1, \infty]$, where (Ω, μ) is a σ- finite measure space. If the following condition*

$$\|f\|_{L^2}^{2+\frac{4}{m}} \lesssim (Bf, f)_{L^2} \|f\|_{L^1}^{\frac{4}{m}}, \quad \forall f \in D(B) \cap L^1(\Omega, \mu), \tag{28}$$

holds with $m > 0$, then

i) $\|e^{-tB} f\|_p \leq C_{p.q} t^{-m/2(1/q-1/p)} \|f\|_q, \quad t > 0,$

ii) $\|f\|_p \leq C_p \|f\|_2^{1-m/k(1/2-1/p)} \|B^{k/2} f\|_2^{m/k(1/2-1/p)},$

for all $p \in [2, \infty], q \in [1, p]$ and $k > m(1/2 - 1/p)$.

From Lemma 1 it follows that condition (28) is satisfied for $m = \frac{n}{\sigma}$. The estimate *i)* with $p = 2$ in Lemma 2 implies the validity of condition $(H5)$ for the operator B.

Thus we have checked conditions $(H1)$–$(H6)$ for operators B and C. By Proposition 1 we obtain the diffusion phenomenon

$$\begin{aligned} &\|u(t) - e^{-tB}(u_0 + Cu_1)\|_2 \\ &\lesssim (\|u_0\|_2 + \|\sqrt{B}u_0\|_2 + \|u_1\|_2 + \|u_0\|_q + \|u_1\|_q)(t+1)^{-m/2(1/q-1/2)-1}, \end{aligned} \tag{29}$$

for $q \in (1, 2]$. Now we will transfer the decay rate from $e^{-tB}(u_0+Cu_1)$ on the decay of $u(t)$. Recalling condition $(H5)$ for $q = 1$ and the assumptions on $a_i(x), i = 1, 2$, we can estimate

$$\|e^{-tB}(u_0 + Cu_1)\|_2 \lesssim t^{-\frac{n}{4\sigma}} \big(\|u_0\|_2 + \|u_1\|_2 + \|u_0\|_1 + \|u_1\|_1\big). \tag{30}$$

Combining estimates (29) and (30) we obtain

$$\begin{aligned} \|u(t)\|_2 \lesssim &\big(\|u_0\|_2 + \|u_0\|_{\dot{H}^\sigma} + \|u_1\|_2 + \|u_0\|_q + \|u_1\|_q\big)(t+1)^{-\frac{n}{2\sigma}(1/q-1/2)-1} + \\ &+ \big(\|u_0\|_2 + \|u_1\|_2 + \|u_0\|_1 + \|u_1\|_1\big)(t+1)^{-\frac{n}{4\sigma}}, \end{aligned} \tag{31}$$

for $q \in (1, 2]$.

Thanks to the assumptions on $a_i(x), i = 1, 2$ and the construction of B and C, we have an obvious approximation

$$E_v(s) = \frac{1}{2}\big(\|\sqrt{C}\partial_s v(s)\|^2 + \|\sqrt{B}v(s)\|^2\big) \approx \Big(\|\partial_s v(s)\|_2 + \|\sqrt{B}v\|_2\Big)^2.$$

Hence, by Proposition 1, we get

$$\|\partial_t u(t)\|_2 + \|\sqrt{B}u(t)\|_2 \lesssim (\|u_0\|_2 + \|u_0\|_{\dot{H}^\sigma} + \|u_1\|_2 + \|u_0\|_1 + \|u_1\|_1)(t+1)^{-\frac{n}{4\sigma}-1}$$
$$+ (\|u_0\|_2 + \|u_0\|_{\dot{H}^\sigma} + \|u_1\|_2 + \|u_0\|_q + \|u_1\|_q)(t+1)^{-\frac{n}{2\sigma}(1/q-1/2)-3/2}, \tag{32}$$

for $q \in (1, 2]$. The proof of Theorem 2 thus is complete.

Acknowledgments The author thanks the Organizers of the session Recent Progress in Evolution Equations, for the opportunity to participate in 13th ISAAC Congress, Ghent 2021. The research from this article is supported by the Ministry of Education and Training of Vietnam (Project no. B2022 - SPH - 01).

References

1. Bauelos, R., Davies, B.: Heat kernel, eigenfunctions, and conditioned Brownian motion in planar domains. J. Funct. Anal. **84**(1), 188–200 (1989)
2. Caffarelli, L., Silvestre, L.: An extension problem related to the fractional Laplacian. Commun. Partial Differential Equations **32**, 1245–1260 (2007)
3. Coulhon, T.: Inegalites de Gagliardo-Nirenberg pour les semi-groupes d'operateurs et applications. Potential Anal. **1**, 343–353 (1992)
4. Coulhon, T.: Ultracontractivity and Nash type inequalities. J. Funct. Anal. **141**, 510–539 (1996)
5. D'Abbicco, M., Ebert, M.R.: Asymptotic profiles and critical exponents for a semilinear damped plate equation with time-dependent coefficients. Asympt. Anal. **123**(1–2), 1–40 (2021)
6. D'Abbicco, M., Reissig, M.: Semilinear structural damped waves. Math. Methods Appl. Sci. **37**, 1570–1592 (2014)
7. De Pablo, A., Quiros, F., Rodrigues, A., Vazquez, J.L.: A general fractional porous medium equation. Commun. Pure Appl. Math. **65**, 1242–1284 (2012)
8. Di Nezza, E., Palatucci, G., Valdinoci, E.: Hitchhiker's guide to the fractional Sobolev spaces. Bull. Sci. Math. **136**, 521–573 (2012)
9. Dyda, B., Frank, R.L.: Fractional Hardy–Sobolev–Maz'ya inequality for domains. Stud. Math. **208**, 151–166 (2012)
10. Pham, T.D., Reissig, M.: Semilinear mixed problems in exterior domains for σ-evolution equations with friction and coefficients depending on spatial variables. J. Math. Anal. Appl. **494**(1) (2021). https://doi.org/10.1016/j.jmaa.2020.124587
11. Pham, T.D., Mozadek, M.K., Reissig, M.: Global existence for semi-linear structurally damped σ-evolution models. J. Math. Anal. Appl. **431**, 569–596 (2015)
12. Punzo, F., Terrone, G.: On the Cauchy problem for a general fractional porous medium equation with variable density. Nonlinear Anal. Theory Methods Appl. **98**, 27–47 (2014)
13. Radu, P., Todorova, G., Yordanov, B.: The generalized diffusion phenomenon and applications. SIAM J. Math. Anal. **48**(1), 174–203 (2016)

14. Wirth, J.: Wave equations with time-dependent dissipation. I: Non-effective dissipation. J. Differential Equations **222**(2), 487–514 (2006)
15. Wirth, J.: Wave equations with time-dependent dissipation. II: Effective dissipation. J. Differential Equations **223**(3), 74–103 (2007)

A Klein-Gordon Model with Time-Dependent Coefficients and a Memory-Type Nonlinearity

Giovanni Girardi

Abstract In this paper we consider a nonlinear Klein-Gordon model with a constant dissipation, a time-dependent positive mass term and a memory type nonlinearity, assuming the initial data to be small in the energy space and in L^m, for some $m \in (1, 2]$. We investigate how the presence of the mass term influences the critical exponent compared with the one found in the purely dissipative case, in low space dimension $n = 1, 2$. Such critical exponent arises from the interplay between the additional decay rate produced by the presence of the mass term and the loss of decay rate due to the presence of the nonlinear memory and to the assumption of initial data in L^m instead of L^1.

1 Introduction

In this paper, we look for global (in time) small data energy solutions to the Cauchy problem

$$\begin{cases} u_{tt} - \Delta u + u_t + \frac{\delta^2}{1+t} u = F(t,u), & t \geq 0,\ x \in \mathbb{R}^n, \\ u(0,x) = u_0(x), \quad u_t(0,x) = u_1(x), \end{cases} \tag{1}$$

where $\delta \geq 0$ and the right-hand side is defined by

$$F(t,u) = \int_0^t (t-s)^{-\gamma} |u(s,x)|^p \, ds, \tag{2}$$

G. Girardi (✉)
Università degli Studi di Bari, Bari, Italy
e-mail: giovanni.girardi@uniba.it

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_46

for some $\gamma \in (0, 1)$ and $p > 1$. In order to do that, we first collect suitable decay estimates for solutions to the corresponding linear Cauchy problem

$$\begin{cases} u_{tt} - \Delta u + u_t + \frac{\delta^2}{1+t} u = 0, \\ u(0, x) = u_0(x), \quad u_t(0, x) = u_1(x), \end{cases} \tag{3}$$

and then, we apply a contraction argument to construct the solution to (1). Problem (3) is a special case of a more general Klein-Gordon model with time-dependent coefficients

$$\begin{cases} v_{tt} - \Delta v + b(t) v_t + m^2(t) v = 0, \\ v(0, x) = v_0(x), \quad v_t(0, x) = v_1(x). \end{cases} \tag{4}$$

Decay estimates for the solution to wave models of the form (4) have been investigated by many authors, under different assumptions on the coefficients $b(t)$ and $m(t)$. One could refer to the survey articles [22] and [28] for an overview of results; the case of zero mass, $m(t) \equiv 0$, is deeply studied in [24] and [25]; here, the author has introduced a classification of the dissipation term $b(t)u_t$ as *non-effective* or *effective*, which distinguishes the dissipation terms according to their strength and influence on the large-time behaviour of solutions. The constant dissipation term u_t is an instance of *effective* damping; in this case the solution $v = v(t, x)$ to the classical damped wave model behaves asymptotically like the solution $w = w(t, x)$ to the heat equation $b(t)w_t - \Delta w = 0$ with suitable initial condition $w(0, \cdot)$ depending on v_0, v_1 and $b(0)$, i.e. $v(t, x) \sim w(t, x)$ in an appropriate L^p-sense. This can be made precise in the form of the so-called *diffusion phenomenon* for damped waves (see [20, 21, 26]).

In [10] the authors studied the Cauchy problem (4) assuming that the damping term is effective and dominates the mass term $m^2(t)u$, i.e. $m(t) = o(b(t))$ as $t \to \infty$, under control assumptions on the oscillations of the coefficients. In that paper it has been proved that under simple conditions on the interaction between $b(t)$ and $m(t)$, the solutions to (4) satisfies the estimate

$$\|v(t, \cdot)\|_{L^2} \leq C\, \omega(t)\, \|(v_0, v_1)\|_{H^1 \times L^2}, \tag{5}$$

where

$$\omega(t) = \exp\left(- \int_0^t \frac{m^2(\tau)}{b(\tau)}\, d\tau \right). \tag{6}$$

The decreasing function $\omega = \omega(t)$ in (6) describes how the interplay between the damping term and the mass term influences the energy decay estimates. In particular,

if we introduce the parameter

$$\beta = \liminf_{t\to\infty} \left(\int_0^t \frac{1}{b(\tau)}\, d\tau \right) m(t)^2, \tag{7}$$

as a consequence of (5), we get

$$\|v(t,\cdot)\|_{L^2} \le C \left(1 + \int_0^t \frac{1}{b(\tau)}\, d\tau\right)^{-\alpha} \|(v_0, v_1)\|_{H^1\times L^2}, \tag{8}$$

for any $\alpha \in [0, \beta)$. Estimate (8) shows that the presence of the mass term produces an additional polynomial decay which becomes faster as the mass term becomes more influent; in particular, if $\beta = \infty$ in (7), one can obtain a polynomial decay as fast as you want. Moreover, if the mass term is assumed to be dominant with respect to the damping term, the solution decays exponentially, that is

$$\|u(t,\cdot)\|_{L^2} \le C\, \exp\left(-\delta \int_0^t b(\tau)\, d\tau\right) \|(u_0, u_1)\|_{H^1\times L^2}, \tag{9}$$

provided that $\liminf_{t\to\infty} m(t)/b(t) > 1/4$ (see [13]). In all the cited papers some conditions on derivatives of the coefficients are assumed to avoid a bad influence of oscillations. However, some long time decay estimates for wave models of the form (4) can be still derived if one considers time periodic coefficients, without further assumptions on derivatives (see [15, 27]).

The decay estimates obtained for the solution to the linear model (4) may be applied to investigate global (in time) existence results for the associated nonlinear problem

$$\begin{cases} v_{tt} - \Delta v + b(t)v_t + m^2(t)v = |v|^p; \\ v(0,x) = v_0(x), \quad v_t(0,x) = v_1(x). \end{cases} \tag{10}$$

In [11], the model without mass has been considered,

$$\begin{cases} v_{tt} - \Delta v + b(t)v_t = |v|^p \\ v(0,x) = v_0(x), \quad v_t(0,x) = v_1(x), \end{cases} \tag{11}$$

and it has been proved that if $b(t)$ is *effective* the critical exponent for global (in time) small data energy solutions to (11) remains the same as for the Cauchy problem with $b = 1$ (see [16–19, 23, 29]). In particular, global existence holds for $p > 1 + 2/n$ if initial data are assumed to be small in exponentially weighted energy spaces. In the subcritical and critical range, $1 < p \le 1 + 2/n$, no global in time small data Sobolev solutions exist, under a suitable sign assumption for the data (for example, see [9]). If smallness of the data is assumed only in the standard energy space $H^1 \times L^2$ and in L^1, then the same result holds in space dimension $n = 1, 2$.

If the additional L^1 smallness is replaced by an additional L^m regularity, then the critical exponent becomes $1 + 2m/n$.

The presence of the mass term in problem (10) can influence the critical exponent with respect to the purely dissipative case (11); indeed, if the effective damping term in (10) dominates the mass term $m^2(t)u$, i.e. $m(t) = o(b(t))$, then estimate (8) holds for any $\alpha \in [0, \beta)$ with β defined as in (7); assuming small initial data in the energy space $H^1 \times L^2$ and in L^m for some $m \in [1, 2]$, the additional decay factor $(1+\int_0^t 1/b(\tau)\,d\tau)^{-\alpha}$ allows to find a scale of critical exponents, which continuously move from $1 + 2m/n$ to 1, as the mass becomes more influent, with respect to the damping term. In particular, the global (in time) existence of small data solutions can be proved for any $p > p_{\beta,m}(n)$ where

$$p_{\beta,m}(n) = \begin{cases} 1 + \frac{2m}{n+2m\beta} & \text{if } \beta \in [0, \infty), \\ 1 & \text{if } \beta = \infty \end{cases}$$

assuming that $\beta > -1 + n/4$ if $n \geq 4$ and $p < n/(n-2)$ if $n \geq 3$ (see Theorem 3 in [10]).

Such result applies to the case in which b and m are defined as in our problem (1): for any $\delta \geq 0$ the Cauchy problem

$$\begin{cases} v_{tt} - \Delta v + v_t + \frac{\delta^2}{1+t} v = |v|^p, \\ v(0, x) = v_0(x), \quad v_t(0, x) = v_1(x), \end{cases} \tag{12}$$

admits a global (in time) solution for any $p > p_\delta(n)$ where

$$p_{\delta,m}(n) := 1 + \frac{2m}{n + 2m\delta^2}, \tag{13}$$

assuming small initial data (v_0, v_1) in $(H^1 \cap L^m) \times (L^2 \cap L^m)$; in particular, in this special case the function ω introduced in (6) corresponds to $(1 + t)^{-\delta^2}$ and the parameter in (7) take value $\beta = \delta^2$ (see Theorem 3). We remark that, the nonlinear term in (10) may be replaced by a more general power nonlinearity of the form $h(t, u) = (1 + \int_0^t 1/b(\tau)\,d\tau)^\omega |u(t, \cdot)|^p$ with $\omega \in [-1, \infty)$ (see [4] and [14]). Finally, we mention that problem (10) has been recently deeply investigated in the scale invariant case $b(t) = \mu_1(1 + t)^{-1}$ and $m(t) = \mu_2(1 + t)^{-1}$ (see, for instance, [2] and [12]).

In this work we consider the memory type nonlinearity defined in (2). Recently, many authors investigated fractional PDEs from different points of view, since they are particularly interesting for the real world applications and they are useful to describe memory and hereditary processes. In particular, it is of interest to understand how to treat nonlinear evolution problems in which the nonlinearity is represented by some memory term like $F(t, u)$ defined in (2).

In [1] the authors have considered the Cauchy problem for the heat equation

$$\begin{cases} u_t - \Delta u = F(t,u), \quad x \in \mathbb{R}^n,\ t > 0, \\ u(0,x) = u_0(x), \end{cases} \tag{14}$$

and they have proved that the critical exponent for (14) is given by

$$\bar{p}_1 := \max\left\{\gamma^{-1}, p_\gamma(n)\right\}, \quad p_\gamma(n) := 1 + \frac{2(2-\gamma)}{n - 2(1-\gamma)}. \tag{15}$$

In [6] the author studied the nonlinear Cauchy problem

$$\begin{cases} u_{tt} - \Delta u + \mu u_t = F(t,u), & x \in \mathbb{R}^n,\ t > 0, \\ u(0,x) = 0,\ \ u_t(0,x) = u_1(x); \end{cases} \tag{16}$$

he has proved the existence of global (in time) solutions again for $p > \bar{p}_1$ with p_1 as in (15), as for the Cauchy problem (14), for $n \le 5$; this is reasonable due to the *diffusion phenomena* discussed above. In particular, as $\gamma \to 1$ the critical exponent $\bar{p}_1$ for problems (14) and (16) tends to $1 + 2/n$ which is the critical exponent for corresponding problem with power nonlinearity $|u|^p$.

On the other hand, if one assumes the initial data to be in L^m instead of L^1, a new critical exponent appears for problems (14) and (16), whose shape is quite different from the one of the critical exponent for L^m theory for the corresponding problem with power nonlinearity $|u|^p$ (see [7]); in particular, in space dimension $n = 1, 2$ the critical exponent becomes

$$\bar{p}_0(n) := \begin{cases} p_\gamma(n), & \text{if } 0 < \gamma < 1 - \frac{n}{2}\left(1 - \frac{1}{m}\right), \\ p_{0,m}(\gamma, n), & \text{if } 1 - \frac{n}{2}\left(1 - \frac{1}{m}\right) < \gamma < 1, \end{cases} \tag{17}$$

where $p_\gamma(n)$ is defined in (15) and

$$p_{0,m}(\gamma, n) := 1 + \frac{2m(2-\gamma)}{n}; \tag{18}$$

this means that, if γ is sufficiently small, the loss of decay due to the assumption of L^m smallness of the initial data becomes irrelevant with respect to the loss of decay rate related to the presence of the nonlinear memory term. In this case, one may easily prove the global existence of solutions for $p > \bar{p}_1$ even replacing the L^1 assumption of the data by the L^m assumption.

In this paper we want to study the critical exponent for the Cauchy problem (1). In particular, we will investigate how the mass term in (1) influences the critical exponent compared with the one found in the purely dissipative case (16), that is $\bar{p}_0(n)$ defined by (17), in low space dimension $n = 1, 2$.

2 Main Results

For $n = 1, 2$ and $\gamma \in (0, 1)$, assuming the initial data in the energy space $H^1 \times L^2$ with additional regularity L^m for some $m \in (1, 2]$, we prove the global existence of small data solutions to problem (1) for any $p > \bar{p}$,

$$\bar{p} := \max\{p_\gamma(n),\, p_{\delta,m}(\gamma, n)\},$$

where $p_\gamma(n)$ is defined by (15) and

$$p_{\delta,m}(\gamma, n) := 1 + \frac{2m(2-\gamma)}{n + 2m\delta^2}.$$

We note that $\bar{p} = \bar{p}_1 = p_\gamma(n)$ if, and only if, γ is sufficiently small, namely,

$$0 < \gamma < 1 - \frac{n}{2}\left(1 - \frac{1}{m}\right) + \delta^2; \tag{19}$$

indeed in this case, on the one hand the presence of the nonlinear memory destroys the benefits which derive by the additional decay factor $(1+t)^{-\delta^2}$ related to the mass term (see Theorem 3; a technical motivation is given in Remark 3); on the other hand, the loss of decay resulting from the assumption of L^m regularity of the initial data becomes negligible with respect to the loss of decay rate related to the presence of the nonlinear memory term: explicitly, it holds $(1+t)^{-\delta^2 + \frac{n}{2}\left(1-\frac{1}{m}\right)} \le (1+t)^{1-\gamma}$. In particular, we note that if $2\delta^2 > n(1 - 1/m)$, then condition (19) is always satisfied, and then the critical exponent $\bar{p} = \bar{p}_1$ is independent of both the mass term coefficient $\delta^2/(1+t)$ and the L^m regularity.

In the following we give our main results; here and hereafter, we denote by $\mathcal{A}$ the space of initial data, i.e.

$$\mathcal{A} := (H^1 \cap L^m) \times (L^2 \cap L^m).$$

Theorem 1 *Let $n = 1, 2$ and $m \in (1, 2]$. Assume that*

$$1 - \frac{n}{2}\left(1 - \frac{1}{m}\right) + \delta^2 < \gamma < 1, \quad \delta^2 < \frac{n}{2}\left(1 - \frac{1}{m}\right), \tag{20}$$

and that $p \ge p_{\delta,m}(\gamma, n)$. Then, there exists $\varepsilon > 0$ such that for any initial data

$$(u_0, u_1) \in \mathcal{A}, \quad \|(u_0, u_1)\| \le \varepsilon,$$

there exists a unique global (in time) solution to (1)

$$u \in C([0, \infty), H^1) \cap C^1([0, \infty), L^2).$$

Moreover, the solution satisfies the following decay estimates

$$\|u(t,\cdot)\|_{L^2} \le C(1+t)^{-\delta^2-\frac{n}{2}\left(\frac{1}{m}-\frac{1}{2}\right)} \|(u_0,u_1)\|_{\mathcal{A}};$$

additionally, its derivatives satisfies

$$\begin{aligned}
\|u_x(t,\cdot)\|_{L^2} &\le C(1+t)^{-\delta^2-\frac{1}{2m}-\frac{1}{4}} \|(u_0,u_1)\|_{\mathcal{A}}, \quad \text{if } n=1,\\
\|\nabla u(t,\cdot)\|_{L^2} &\le C(1+t)^{-\delta^2-\frac{1}{m}} \ln(e+t)\, \|(u_0,u_1)\|_{\mathcal{A}}, \quad \text{if } n=2,\\
\|u_t(t,\cdot)\|_{L^2} &\le C(1+t)^{-\delta^2+\frac{n}{2}\left(1-\frac{1}{m}\right)-1} \|(u_0,u_1)\|_{\mathcal{A}}, \quad \text{if } n=1,2.
\end{aligned}$$

Remark 1 We notice that $p_{\delta,m}(\gamma,n) > 2$ for $n = 1, 2$, if one considers δ^2 as in (20). This allows to prove the desired results working only with energy solutions, without the need to use L^1-L^p estimates with $p < 2$; one may do this to investigate the same problem in higher dimension $n = 3, 4$.

Theorem 2 *Let $n = 1, 2$, $m \in [1, 2]$ and $\delta^2 > 0$. Assume that*

$$1-\frac{n}{2} < \gamma < 1-\frac{n}{2}\left(1-\frac{1}{m}\right)+\delta^2 \tag{21}$$

and that $p > p_\gamma(n)$. Then, there exists $\varepsilon > 0$ such that for any initial data

$$(u_0,u_1) \in \mathcal{A}, \quad \|(u_0,u_1)\| \le \varepsilon,$$

there exists a unique global solution to (1)

$$u \in C([0,\infty), H^1) \cap C^1([0,\infty), L^2).$$

Moreover, the solution satisfies the following decay estimates

$$\|u(t,\cdot)\|_{L^2} \le C(1+t)^{-\frac{n}{4}+1-\gamma} \|(u_0,u_1)\|_{\mathcal{A}};$$

additionally, its derivatives satisfies

$$\begin{aligned}
\|u_x(t,\cdot)\|_{L^2} &\le C(1+t)^{-\gamma+\frac{1}{4}} \|(u_0,u_1)\|_{\mathcal{A}}, \quad \text{if } n=1,\\
\|\nabla u(t,\cdot)\|_{L^2} &\le C(1+t)^{-\gamma} \ln(e+t)\, \|(u_0,u_1)\|_{\mathcal{A}}, \quad \text{if } n=2,\\
\|u_t(t,\cdot)\|_{L^2} &\le C(1+t)^{-\gamma} \|(u_0,u_1)\|_{\mathcal{A}} \quad \text{if } n=1,2.
\end{aligned}$$

Remark 2 The assumption $\gamma > 1 - n/2$ in (21) guarantees that $p_\gamma(n) < \infty$.

3 Open Problems

It would be interesting to prove the sharpness of the global existence results: the presence of a time-dependent mass term $m^2(t)u$ in Cauchy problem (1) makes very difficult the application of a test function method to investigate some non existence results for $p < \bar{p}$.

Also, the study of analogous results for different Cauchy problems with nonlinear memory with not integrable data would be of interest. We recall, for instance, that in [5] the author considers the Cauchy problem

$$\begin{cases} u_{tt} - \Delta u + \mu(-\Delta)^{\frac{1}{2}} u_t = F(t,u), \quad x \in \mathbb{R}^n,\ t > 0, \\ u(0,x) = 0,\ \ u_t(0,x) = u_1(x), \end{cases} \tag{22}$$

and he proves that global (in time) small data energy solutions exist for

$$p > \max\left\{\gamma^{-1}, 1 + \frac{3-\gamma}{n+\gamma-2}\right\}, \tag{23}$$

for any $n \geq 2$, under L^1 smallness assumption for the initial data; the same problem (22), with $|u_t|^p$ in place of $|u|^p$ in the nonlinear memory term, was studied in [8]. One may study how the obtained critical exponents change if the L^1 regularity of the initial data is replaced by L^m regularity.

4 Proof of the Main Results

In order to prove our main results, it is useful to recall the following decay estimates for the solution to the associated linear problem (see [10]):

Theorem 3 *Let $(u_0, u_1) \in \mathcal{A}$. Then, the solution $u(t,\cdot)$ to the linear Cauchy problem (3) satisfies the following decay estimates:*

$$\|u(t,\cdot)\|_{L^2} \leq C(1+t)^{-\delta^2 - \frac{n}{2}\left(\frac{1}{m}-\frac{1}{2}\right)} \|(u_0,u_1)\|_{\mathcal{A}},$$

$$\|\nabla u(t,\cdot)\|_{L^2} \leq C(1+t)^{-\delta^2 - \frac{n}{2}\left(\frac{1}{m}-\frac{1}{2}\right)-\frac{1}{2}} \|(u_0,u_1)\|_{\mathcal{A}},$$

$$\|u_t(t,\cdot)\|_{L^2} \leq C(1+t)^{-\delta^2 - \frac{n}{2}\left(\frac{1}{m}-\frac{1}{2}\right)-1} \|(u_0,u_1)\|_{\mathcal{A}};$$

here, the constant C is independent of t.

Due to the presence of time-dependent coefficients, the equation in (3) is not invariant by time translations. Having in mind to apply the Duhamel's principle, we need the decay estimates for the solution to a family of parameter-dependent Cauchy

problems

$$\begin{cases} u_{tt} - \Delta u + u_t + \frac{\delta^2}{1+t} u = 0, & t \geq s, \\ u(s,x) = 0, \quad u_t(s,x) = g(s,x), \end{cases} \tag{24}$$

where $s \geq 0$, obtaining decay rates which depend on both t and s (see Lemma 3.1 in [10]):

Lemma 1 *Let $g(s,\cdot) \in L^1 \cap L^2$. Then, the solution to Cauchy problem (24) satisfies the following estimates:*

$$\|u(t,\cdot)\|_{L^2} \leq C(1+t-s)^{-\frac{n}{4}} \left(\frac{1+s}{1+t}\right)^{\delta^2} \|g(s,\cdot)\|_{L^1 \cap L^2},$$

$$\|\nabla u(t,\cdot)\|_{L^2} \leq C(1+t-s)^{-\frac{n}{4}-\frac{1}{2}} \left(\frac{1+s}{1+t}\right)^{\delta^2} \|g(s,\cdot)\|_{L^1 \cap L^2},$$

$$\|u_t(t,\cdot)\|_{L^2} \leq C(1+t-s)^{-\frac{n}{4}-1} \left(\frac{1+s}{1+t}\right)^{\delta^2} \|g(s,\cdot)\|_{L^1 \cap L^2},$$

where the constant C is independent of s.

Remark 3 We notice that in the decay estimates for the solution to the parameter dependent Cauchy problem (24), the influence of the mass term $\delta^2(1+t)^{-1}u$ is described by the additional factor $(1+s)^{\delta^2}/(1+t)^{\delta^2}$; when δ^2 becomes sufficiently large, namely condition (19) holds, such influence becomes irrelevant with respect to the influence of the nonlinear memory term.

In the proof of each theorem we will introduce the solution space $X(T) := C([0,T], H^1) \cap C([0,T], L^2)$, equipped with an appropriate norm. Then, we may introduce the operator

$$N : \ u \in X(T) \to u^{\text{lin}} + Gu, \qquad Gu(t,x) := \int_0^t \Phi(t,s,\cdot) *_{(x)} F(s,u(s,\cdot))(x) ds,$$

where u^{lin} is the solution to the linear Cauchy problem (3), and by $\Phi(t,s,\cdot) *_{(x)} F(s,u(s,\cdot))(x)$ we are denoting the solution to problem (24) with $g(s,\cdot) = F(s,u(s,\cdot))$. According to the Duhamel's principle, we will prove the existence of a unique global (in time) solution to (1) as the fixed point of the operator N. Hence, in order to get the global (in time) existence and uniqueness of the solution in $X(T)$, we need to prove the following two crucial estimates:

$$\|Nu\|_{X(T)} \leq C \|(u_0,u_1)\|_{\mathcal{A}} + \|u\|^p_{X(T)}, \tag{25}$$

$$\|Nu - Nv\|_{X(T)} \leq C \|u-v\|_{X(T)} \left(\|u\|^{p-1}_{X(T)} + \|v\|^{p-1}_{X(T)}\right), \tag{26}$$

with $C > 0$, independent of T, where $\mathcal{A}$ denotes the space of the data. As a consequence of Banach's fixed point theorem, the conditions (25) and (26) guarantee the existence of a uniquely determined solution $u \in X(T)$ to the integral equation $u = u^{\text{lin}} + Gu$, provided that $\|(u_0, u_1)\|_{\mathcal{A}}$ is sufficiently small., i.e. there exists a uniquely determined solution to Cauchy problem (1), in $X(T)$, for small initial data. Since the constants in (25) and (26) do not depend on T, the solution is globally defined (in time).

In the proof of our result, it will be useful to apply the following straightforward estimates (see, for instance, [3]).

Lemma 2 *For any $\gamma \in (0, 1)$, $\delta > 1$ and $\omega \in \mathbb{R}$ it holds*

$$\int_0^t (1+t-s)^{-\omega} \int_0^s (s-\tau)^{-\gamma}(1+\tau)^{-\delta}\, d\tau\, ds \lesssim \begin{cases} (1+t)^{-\gamma+1-\omega} & \text{if } \omega < 1, \\ (1+t)^{-\gamma} & \text{if } \omega > 1, \\ (1+t)^{-\gamma}\ln(e+t) & \text{if } \omega = 1. \end{cases}$$

Lemma 3 *For any $\gamma, \delta \in (0, 1)$ and $\omega \in \mathbb{R}$ it holds*

$$\int_0^t (1+t-s)^{-\omega} \int_0^s (s-\tau)^{-\gamma}(1+\tau)^{-\delta}\, d\tau\, ds \lesssim \begin{cases} (1+t)^{-\gamma+2-\delta-\omega} & \text{if } \omega < 1, \\ (1+t)^{-\gamma+1-\delta} & \text{if } \omega > 1, \\ (1+t)^{-\gamma+1-\delta}\ln(e+t) & \text{if } \omega = 1; \end{cases}$$

Proof of Theorem 1 For any $T > 0$, we define the Banach spaces

$$X(T) = C([0, T], H^1) \cap C^1([0, T], L^2)$$

equipped with the norm

$$\|u\|_{X(T)} := \sup_{0\le t\le T} (1+t)^{\delta^2+\frac{1}{2m}-\frac{1}{4}} \Big\{ \|u(t,\cdot)\|_{L^2} + (1+t)^{\frac{1}{2}}\|u_x(t,\cdot)\|_{L^2} + (1+t)^{\frac{3}{4}}\|u_t(t,\cdot)\|_{L^2} \Big\},$$

if $n = 1$ and

$$\|u\|_{X(T)} := \sup_{0\le t\le T} \Big\{ \sup_{q\in[2,\infty)} (1+t)^{\delta^2+\frac{1}{m}-\frac{1}{q}} \|u(t,\cdot)\|_{L^q} + (1+t)^{\delta^2+\frac{1}{m}} \Big(\ln(e+t)^{-1}\|\nabla u(t,\cdot)\|_{L^2} + \|u_t(t,\cdot)\|_{L^2} \Big) \Big\},$$

if $n = 2$. Since $\gamma > n/(2m) - n/2 + 1 + \delta^2$, as an immediate consequence of Theorem 3, we obtain

$$\|u^{\text{lin}}\|_{X(T)} \le C\|(u_0, u_1)\|_{\mathcal{A}}, \tag{27}$$

where C is independent of T. By the definition of $\|\cdot\|_{X(T)}$, using the Gagliardo-Nirenberg inequality we get

$$\|u(t,\cdot)\|_{L^q} \lesssim (1+t)^{-\frac{n}{2}\left(\frac{1}{m}-\frac{1}{q}\right)-\delta^2}\|u\|_{X(T)}, \tag{28}$$

for any $q \in [2,\infty]$ if $n = 1$ and $q \in [2,\infty)$ if $n = 2$. For $j + k = 0, 1$ we have

$$\|\partial_t^k \nabla^j Gu(t,\cdot)\|_{L^2} \le \int_0^t \left\|\partial_t^k \nabla^j \left(\Phi(t,s,\cdot) *_{(x)} F(s,u(s,x))\right)\right\|_{L^2} ds,$$

where $F(s,u)$ is defined by (2):

$$F(s,u(s,x)) = \int_0^s (s-\tau)^{-\gamma}|u(\tau,x)|^p \, d\tau.$$

In order to prove estimate (25) we apply Lemma 1 to get:

$$\begin{aligned}&\|\partial_t^k \nabla^j Gu(t,\cdot)\|_{L^2}\\ &\quad\lesssim (1+t)^{-\delta^2}\int_0^t (1+t-s)^{-\frac{n}{4}-\frac{j}{2}-k}(1+s)^{\delta^2}\int_0^s (s-\tau)^{-\gamma}\|u(\tau,\cdot)\|_{L^p\cap L^{2p}}^p \, d\tau \, ds;\end{aligned} \tag{29}$$

By estimate (28), for any $p \ge p_{\delta,m}(\gamma,n)$ it holds

$$\|u(\tau,\cdot)\|_{L^p}^p \lesssim (1+\tau)^{-\frac{n}{2}\left(\frac{p}{m}-1\right)-\delta^2 p}\|u\|_{X(T)}^p \lesssim (1+\tau)^{-\frac{n}{2}\left(\frac{1}{m}-1\right)-2+\gamma-\delta^2}\|u\|_{X(T)}^p; \tag{30}$$

and

$$\|u(\tau,\cdot)\|_{L^{2p}}^p \lesssim (1+\tau)^{-\frac{n}{2}\left(\frac{p}{m}-\frac{1}{2}\right)-\delta^2 p}\|u\|_{X(T)}^p \lesssim (1+\tau)^{-\frac{n}{2}\left(\frac{1}{m}-1\right)-2+\gamma-\delta^2}\|u\|_{X(T)}^p; \tag{31}$$

indeed, $p \ge p_{\delta,m}(\gamma,n)$ and $p_{\delta,m}(\gamma,n) \ge 2$ for $n = 1, 2$ as a consequence of the assumption $\delta^2 \le n/2(1-1/m)$.

Then, being $1+s \le 1+t$ for any $s \in [0,t]$ and applying Lemma 3, we get:

$$\begin{aligned}
&\|Gu(t,\cdot)\|_{L^q} \\
&\quad \lesssim \int_0^t (1+t-s)^{-\frac{n}{2}\left(1-\frac{1}{q}\right)} \int_0^s (s-\tau)^{-\gamma}(1+\tau)^{-\frac{n}{2}\left(\frac{1}{m}-1\right)-2+\gamma-\delta^2}\|u\|_{X(T)}^p \, d\tau \, ds, \\
&\quad \lesssim (1+t)^{-\frac{n}{2}\left(\frac{1}{m}-\frac{1}{q}\right)-\delta^2}\|u\|_{X(T)}^p
\end{aligned}$$

for any $q \in [2,\infty)$ if $n=1$, and $q \in [2,\infty)$ if $n=2$; moreover, we have

$$\begin{aligned}
\|\partial_x Gu(t,\cdot)\|_{L^2} &\lesssim \int_0^t (1+t-s)^{-\frac{3}{4}} \int_0^s (s-\tau)^{-\gamma}(1+\tau)^{-\frac{1}{2m}-\frac{3}{2}+\gamma-\delta^2}\|u\|_{X(T)}^p \, d\tau \, ds \\
&\lesssim (1+t)^{-\delta^2-\frac{1}{2m}-\frac{1}{4}}\|u\|_{X(T)}^p,
\end{aligned}$$

if $n=1$, and similarly

$$\begin{aligned}
&\|\nabla Gu(t,\cdot)\|_{L^2} \\
&\quad \lesssim \int_0^t (1+t-s)^{-1} \int_0^s (s-\tau)^{-\gamma}(1+\tau)^{-\frac{1}{m}-1+\gamma-\delta^2}\|u\|_{X(T)}^p \, d\tau \, ds \\
&\quad \lesssim (1+t)^{-\delta^2-\frac{1}{m}} \ln(e+t)\|u\|_{X(T)}^p,
\end{aligned}$$

if $n=2$. Finally, we find

$$\begin{aligned}
&\|\partial_t Gu(t,\cdot)\|_{L^2} \\
&\quad \lesssim \int_0^t (1+t-s)^{-\frac{n}{4}-1} \int_0^s (s-\tau)^{-\gamma}(1+\tau)^{-\frac{n}{2}\left(\frac{1}{m}-1\right)-2+\gamma-\delta^2}\|u\|_{X(T)}^p \, d\tau \, ds \\
&\quad \lesssim (1+t)^{-\delta^2+\frac{n}{2}\left(1-\frac{1}{m}\right)-1}\|u\|_{X(T)}^p.
\end{aligned}$$

Summarizing estimate (25) follows for any $p \ge p_{\delta,m}(\gamma,n)$.

We proceed similarly to prove estimate (26). In particular, we replace (30) by

$$\begin{aligned}
&\||u(\tau,\cdot)|^p - |v(\tau,\cdot)|^p\|_{L^1} \\
&\qquad \lesssim \||u(\tau,\cdot)-v(\tau,\cdot)|\,(|u(\tau,\cdot)|^{p-1}+|v(\tau,\cdot)|^{p-1})\|_{L^1} \\
&\qquad \lesssim \|u(\tau,\cdot)-v(\tau,\cdot)\|_{L^p}\left(\|u(\tau,\cdot)\|_{L^p}^{p-1}+\|v(\tau,\cdot)\|_p^{p-1}\right) \\
&\qquad \lesssim (1+\tau)^{-\frac{n}{2}(p-1)+p-\gamma p}\|u-v\|_{X(T)}\left(\|u\|_{X(T)}^{p-1}+\|v\|_{X(T)}^{p-1}\right),
\end{aligned}$$

and, likewise, we replace (31) by

$$\||u(\tau,\cdot)|^p - |v(\tau,\cdot)|^p\|_{L^2} \lesssim (1+\tau)^{-\frac{n}{2}(p-\frac{1}{2})+p-\gamma p}\|u-v\|_{X(T)} \left(\|u\|_{X(T)}^{p-1} + \|v\|_{X(T)}^{p-1}\right).$$

This concludes the proof.

Proof of Theorem 2 For any $T > 0$, we equip the Banach spaces

$$X(T) = C([0,T], H^1) \cap C^1([0,T], L^2)$$

equipped with the norm

$$\|u\|_{X(T)} := \sup_{0\le t\le T} (1+t)^{\gamma}\left((1+t)^{-\frac{3}{4}}\|u(t,\cdot)\|_{L^2} + (1+t)^{-\frac{1}{4}}\|u_x(t,\cdot)\|_{L^2} + \|u_t(t,\cdot)\|_{L^2}\right),$$

if $n = 1$ and

$$\|u\|_{X(T)} := \sup_{0\le t\le T}\left\{(1+t)^{\gamma}\left(\sup_{q\in[2,\infty)}(1+t)^{-\frac{1}{q}}\|u(t,\cdot)\|_{L^q} + \ln(e+t)^{-1}\|\nabla u(t,\cdot)\|_{L^2} + \|u_t(t,\cdot)\|_{L^2}\right)\right\},$$

if $n = 2$. As a consequence of Theorem 3, being $\gamma < n/(2m) - n/2 + 1 + \delta^2$ we immediately get

$$\|u^{\,\mathrm{lin}}\|_{X(T)} \le C\|(u_0,u_1)\|_{\mathcal{A}} \tag{32}$$

where $C > 0$ does not depend on T. By applying the Gagliardo-Nirenberg inequality we get

$$\|u(t,\cdot)\|_{L^q} \lesssim (1+t)^{-\frac{n}{2}\left(1-\frac{1}{q}\right)+1-\gamma}\|u\|_{X(T)}, \tag{33}$$

for any $q \in [2,\infty]$ if $n = 1$, and $q \in [2,\infty)$ if $n = 2$. In order to prove estimate (25) we apply Lemma 1 to get:

$$\|Gu(t,\cdot)\|_{L^q} \lesssim (1+t)^{-\delta^2}\int_0^t (1+t-s)^{-\frac{n}{2}\left(1-\frac{1}{q}\right)}(1+s)^{\delta^2}\int_0^s (s-\tau)^{-\gamma}\|u(\tau,\cdot)\|_{L^p\cap L^{2p}}^p\,d\tau\,ds;$$

for any $q \in [2,\infty]$ if $n=1$, and $q \in [2,\infty)$ if $n=2$; moreover, for $j+k=1$ we have

$$\|\partial_t^k \nabla^j Gu(t,\cdot)\|_{L^2}$$
$$\lesssim (1+t)^{-\delta^2} \int_0^t (1+t-s)^{-\frac{n}{4}-\frac{j}{2}-k} (1+s)^{\delta^2} \int_0^s (s-\tau)^{-\gamma} \|u(\tau,\cdot)\|_{L^p\cap L^{2p}}^p \, d\tau \, ds.$$

By estimate (33), for any $p > p_\gamma(n)$ we have

$$\|u(\tau,\cdot)\|_{L^p\cap L^{2p}}^p \lesssim (1+\tau)^{-\frac{n}{2}(p-1)+p-\gamma p} \|u\|_{X(T)}^p,$$

indeed, it holds $p_\gamma(n) \geq 2$ for $n=1,2$. As a consequence, since $1+s \leq 1+t$ for any $s \in [0,t]$, we get

$$\|Gu(t,\cdot)\|_{L^q}$$
$$\lesssim \int_0^t (1+t-s)^{-\frac{n}{2}\left(1-\frac{1}{q}\right)} \int_0^s (s-\tau)^{-\gamma} (1+\tau)^{-\frac{n}{2}(p-1)+p-\gamma p} \|u\|_{X(T)}^p \, d\tau \, ds$$

for any $q \in [2,\infty]$ if $n=1$, and $q \in [2,\infty)$ if $n=2$; furthermore, we find

$$\|\partial_t^k \nabla^j Gu(t,\cdot)\|_{L^2}$$
$$\lesssim \int_0^t (1+t-s)^{-\frac{n}{4}-\frac{j}{2}-k} \int_0^s (s-\tau)^{-\gamma} (1+\tau)^{-\frac{n}{2}(p-1)+p-\gamma p} \|u\|_{X(T)}^p \, d\tau \, ds,$$

for $j+k=1$. For any $p > p_\gamma(n)$ it holds

$$\frac{n}{2}(p-1) - p + \gamma p > 1;$$

thus, we can apply Lemma 2 to conclude

$$\|Gu(t,\cdot)\|_{L^q} \leq (1+t)^{-\frac{n}{2}\left(1-\frac{1}{q}\right)+1-\gamma} \|u\|_{X(T)}^p,$$

for any $q \in [2,\infty]$ if $n=1$, and $q \in [2,\infty)$ if $n=2$; similarly, we get

$$\|\partial_x Gu(t,\cdot)\|_{L^2} \leq (1+t)^{\frac{1}{4}-\gamma} \|u\|_{X(T)}^p,$$

if $n=1$, and

$$\|\nabla Gu(t,\cdot)\|_{L^2} \leq (1+t)^{-\gamma} \ln(e+t) \|u\|_{X(T)}^p,$$

if $n = 2$. Finally, we obtain

$$\|\partial_t Gu(t,\cdot)\|_{L^2} \leq (1+t)^{-\gamma}\|u\|^p_{X(T)}.$$

Thus, estimate (25) follows for any $p > p_\gamma(n)$. As in the proof of Theorem 1, one can proceed similarly to prove that (26) holds.

Remark 4 We notice that in Theorems 1 and 2 the estimates obtained for $\|\partial_t^k \nabla^j u(t,\cdot)\|_{L^2}$ have in some cases a loss of decay with respect to the corresponding linear estimates (see Theorem 3). Indeed, in the proof of the estimates for the solution to the nonlinear problem (1), there is a competition between the benefit which derive by the application of the $L^1 \cap L^2 - L^2$ estimate (see, for instance, (29)) and the drawbacks due to the presence of the singularity $(s-\tau)^{-\gamma}$. More precisely, in the proof of each theorem, for any $u \in X(T)$ we estimate

$$\|u(\tau,\cdot)\|_{L^p} \lesssim (1+\tau)^{-\beta_p}\|u\|_{X(T)},$$

where β_p is given by (28) in Theorem 1 and (33) in Theorem 2; then, applying Lemmas 2 and 3 we obtain

$$\|\partial_t^k \nabla^j Gu(t,\cdot)\|_{L^2} \lesssim (1+t)^{\left(1-\frac{n}{4}-\frac{j}{2}-k\right)_+ - \gamma + (1-p\beta_p)_+}\ell(t)\|u\|^p_{X(T)},$$

where

$$\ell(t) = \begin{cases} \ln(e+t) & \text{if } \frac{n}{4}+\frac{j}{2}+k = 1, \\ 1 & \text{otherwise;} \end{cases}$$

thus, if γ and δ satisfy (20) we are able to prove our global existence result for any $p \geq p_{\delta,m}(\gamma,n)$; moreover, if $n/4 + j/2 + k < 1$ we obtain for $\|\partial_t^k \nabla^j u(t,\cdot)\|_{L^2}$ the same estimate as for the solution to the linear problem; instead, if $n/4 + j/2 + k \geq 1$ the loss $(1+t)^{\frac{n}{4}+\frac{j}{2}+k-1}\ell(t)$ appears.

On the other hand, if γ is small, namely (21) holds, we can prove the existence of global (in time) small data solutions only for $p > p_\gamma(n)$ and the estimate of $\|\partial_t^k \nabla^j u(t,\cdot)\|_{L^2}$ always contains a loss of decay with respect to the estimate for the solution to the linear problem; such loss of decay is given by $(1+t)^{\delta^2+\frac{n}{2}\left(\frac{1}{m}-1\right)+1-\gamma}$ if $n/4 + j/2 + k < 1$, and $(1+t)^{\delta^2+\frac{n}{2}\left(\frac{1}{m}-\frac{1}{2}\right)+\frac{j}{2}+k-\gamma}\ell(t)$ if $n/4 + j/2 + k \geq 1$.

Acknowledgments The author is "Titolare di un Assegno di Ricerca dell'Istituto Nazionale di Alta Matematica (INdAM)".

References

1. Cazenave, T., Dickstein, F., Weissler, F.B.: An equation whose Fujita critical exponent is not given by scaling. Nonlinear Anal. **68**, 862–874 (2008)
2. Chiarello, F.A., Girardi, G., Lucente, S.: Fujita modified exponent for scale invariant damped semilinear wave equations. J. Evol. Equations **21**, 2735–2748 (2021). https://doi.org/10.1007/s00028-021-00705-2
3. Cui, S.: Local and global existence of solutions to semilinear parabolic initial value problems. Nonlinear Anal. **43**, 293–323 (2001)
4. D'Abbicco, M.: Small data solutions for semilinear wave equations with effective damping. Discrete Contin. Dyn. Syst., 183–191 (2013)
5. D'Abbicco, M.: A wave equation with structural damping and nonlinear memory. Nonlinear Differential Equations Appl. **21**(5), 751–773 (2014). ISSN: 1021-9722. https://doi.org/10.1007/s00030-014-0265-2
6. D'Abbicco, M.: The influence of a nonlinear memory on the damped wave equation. Nonlinear Anal. **95**, 130–145 (2014). https://doi.org/10.1016/j.na.2013.09.006
7. D'Abbicco, M.: A new critical exponent for the heat and damped wave equations with nonlinear memory and not integrable data. In: Cicognani, M., Del Santo, D., Parmeggiani, A., Reissig, M. (eds.) Anomalies in Partial Differential Equations. Springer INdAM Series, vol. 43. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-61346-4-9
8. D'Abbicco, M., Girardi, G.: A structurally damped σ-evolution equation with nonlinear memory. Math. Methods Appl. Sci. (2020). https://doi.org/10.1002/mma.6633
9. D'Abbicco, M., Lucente, S.: A modified test function method for damped wave equations. Adv. Nonlinear Stud. **13**, 867–892 (2013)
10. D'Abbicco, M., Girardi, G., Reissig, M.: A scale of critical exponents for semilinear waves with time-dependent damping and mass terms. Nonlinear Anal. **179**, 15–40 (2019). https://doi.org/10.1016/j.na.2018.08.006
11. D'Abbicco, M., Lucente, S., Reissig, M.: Semilinear wave equations with effective damping. Chinese Ann. Math. **34B**(3), 345–380 (2013). https://doi.org/10.1007/s11401-013-0773-0
12. do Nascimento, W.N., Palmieri, A., Reissig, M.: Semi-linear wave models with power nonlinearity and scale invariant time-dependent mass and dissipation. Math. Nachr. **290**, 1779–1805 (2017)
13. Girardi, G.: Semilinear damped Klein-Gordon models with time-dependent coefficients. In: D'Abbicco, M., Ebert, M., Georgiev, V., Ozawa, T. (eds.) New Tools for Nonlinear PDEs and Application. Trends in Mathematics. Birkhäuser, Cham (2019). https://doi.org/10.1007/978-3-030-10937-0-7
14. Girardi, G.: Small data solutions for semilinear waves with time-dependent damping and mass terms. In: Boggiatto, P., et al. (eds.) Advances in Microlocal and Time-Frequency Analysis. Applied and Numerical Harmonic Analysis. Birkhäuser, Cham (2020). https://doi.org/10.1007/978-3-030-36138-9-14
15. Girardi, G., Wirth, J.: Decay estimates for a Klein–Gordon model with time-periodic coefficients. In: Cicognani, M., Del Santo, D., Parmeggiani, A., Reissig, M. (eds.) Anomalies in Partial Differential Equations. Springer INdAM Series, vol. 43. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-61346-4-14
16. Ikehata, R., Ohta, M.: Critical exponents for semilinear dissipative wave equations in $\mathbb{R}^N$. J. Math. Anal. Appl. **269**, 87–97 (2002)
17. Ikehata, R., Mayaoka, Y., Nakatake, T.: Decay estimates of solutions for dissipative wave equations in $\mathbb{R}^N$ with lower power nonlinearities. J. Math. Soc. Jpn. **56**(2), 365–373 (2004)
18. Li, T.T., Zhou, Y.: Breakdown of solutions to $\Box u + u_t = |u|^{1+\alpha}$. Discrete Contin. Dyn. Syst. **1**, 503–520
19. Matsumura, A.: On the asymptotic behavior of solutions of semi-linear wave equations. Publ. RIMS. **12**, 169–189 (1976)

20. Narazaki, T.: $L^p - L^q$ estimates for damped wave equations and their applications to semi-linear problem. J. Math. Soc. Jpn. **56**(2), 585–626 (2004)
21. Nishihara, K.: $L^p - L^q$ estimates for damped wave equation in 3-dimensional space and their application. Math. Z. **244**, 631–649 (2003)
22. Reissig, M.: $L^p - L^q$ decay estimates for wave equations with time-dependent coefficients. J. Nonlin. Math. Phys. **11/4**, 534–548 (2004)
23. Todorova, G., Yordanov, B.: Critical exponent for a nonlinear wave equation with damping. J. Differential Equations **174**, 464–489 (2001)
24. Wirth, J.: Wave equations with time-dependent dissipation I. Non-effective dissipation. J. Differential Equations **222/2**, 487–514 (2006)
25. Wirth, J.: Wave equations with time-dependent dissipation II. Effective dissipation. J. Differential Equations **232/1**, 74–103 (2007)
26. Wirth, J.: Scattering and modified scattering for abstract wave equations with time-dependent dissipation. Adv. Differential Equations **12**(10), 1115–1133 (2007)
27. Wirth, J.: On the influence of time-periodic dissipation on energy and dispersive estimates. Hiroshima Math. J. **38**(3), 397–410 (2008). https://doi.org/10.32917/hmj/1233152777
28. Wirth, J.: Energy inequalities and dispersive estimates for wave equations with time-dependent coefficients. Rend. Istit. Mat. Univ. Trieste **42**(suppl.), 205–219 (2010)
29. Zhang, Q.S.: A blow-up result for a nonlinear wave equation with damping: the critical case. C. R. Acad. Sci. Paris Sér. I Math. **333**, 109–114 (2001)

Intrinsic Polynomial Squeezing for Balakrishnan-Taylor Beam Models

Eduardo H. Gomes Tavares, Marcio A. Jorge Silva, Vando Narciso, and André Vicente

Abstract We explore the energy decay properties related to a model in extensible beams with the so-called *energy damping*. We investigate the influence of the nonlocal damping coefficient in the stability of the model. We prove, for the first time, that the corresponding energy functional is squeezed by polynomial-like functions involving the power of the damping coefficient, which arises intrinsically from the Balakrishnan-Taylor beam models. As a consequence, it is shown that such models with nonlocal energy damping are never exponentially stable in its essence.

1 Introduction

In 1989 Balakrishnan and Taylor [3] derived some prototypes of vibrating extensible beams with the so-called *energy damping*. Accordingly, the following one dimensional beam equation is proposed

$$\partial_{tt}u - 2\zeta\sqrt{\lambda}\partial_{xx}u + \lambda\partial_{xxxx}u - \alpha\left[\int_{-L}^{L}\left(\lambda|\partial_{xx}u|^2 + |\partial_t u|^2\right)dx\right]^q \partial_{xxt}u = 0, \quad (1)$$

where $u = u(x,t)$ represents the transversal deflection of a beam with length $2L > 0$ in the rest position, $\alpha > 0$ is a damping coefficient, ζ is a constant appearing in Krylov-Bogoliubov's approximation, $\lambda > 0$ is related to mode frequency and

E. H. Gomes Tavares (✉) · M. A. Jorge Silva
State University of Londrina, Londrina, PR, Brazil
e-mail: marcioajs@uel.br

V. Narciso
State University of Mato Grosso do Sul, Dourados, MS, Brazil
e-mail: vnarciso@uems.br

A. Vicente
Western Paraná State University, Cascavel, PR, Brazil
e-mail: andre.vicente@unioeste.br

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_47

spectral density of external forces, and $q = 2(n+\beta)+1$ with $n \in \mathbb{N}$ and $0 \leq \beta < \frac{1}{2}$. We still refer to [3, Sect. 4] for several other beam equations taking into account nonlocal energy damping coefficients, as well as [2, 4, 6, 7, 12, 17, 18] for associated models. A normalized n-dimensional equation corresponding to (1) can be seen as follows

$$\partial_{tt}u - \kappa\Delta u + \Delta^2 u - \alpha\left[\int_\Omega \left(|\Delta u|^2 + |\partial_t u|^2\right) dx\right]^q \Delta\partial_t u = 0, \tag{2}$$

where we denote $\lambda = 1$ and $\kappa = 2\zeta$; Ω may represent an open bounded of $\mathbb{R}^n$; and the symbols Δ and Δ^2 stand for the usual Laplacian and Bi-harmonic operators, respectively. Additionally, in order to see the problem within the frictional context of dampers, we rely on materials whose viscosity can be essentially seen as friction between moving solids. In this way, besides reflecting on a more challenging model (at least) from the stability point of view, one may metaphysically supersede the viscous damping in (2) by a nonlocal frictional one so that we cast the model

$$\partial_{tt}u - \kappa\Delta u + \Delta^2 u + \alpha\left[\int_\Omega \left(|\Delta u|^2 + |\partial_t u|^2\right) dx\right]^q \partial_t u = 0. \tag{3}$$

The main goal of this paper is to explore the influence of the nonlocal damping coefficient in the stability of problem (3). Unlike the existing literature on extensible beams with full viscous or frictional damping, we are going to see for the first time that the feature of the *energy* damping coefficient

$$\mathcal{E}_q(t) := \mathcal{E}_q(u, u_t)(t) = \left[\int_\Omega \left(|\Delta u(t)|^2 + |\partial_t u(t)|^2\right) dx\right]^q, \quad q > 0, \tag{4}$$

not only prevents exponential decay, but also gives us a polynomial range in terms of q whose energy is squeezed and goes to zero polynomially when time goes to infinity. More precisely, by noting that the corresponding energy functional is given by

$$E_\kappa(t) := E_\kappa(u, u_t)(t) = \int_\Omega \left(|\Delta u(t)|^2 + |\partial_t u(t)|^2 + \kappa|\nabla u(t)|^2\right) dx, \quad \kappa \geq 0, \tag{5}$$

then it belongs to an area of variation between upper and lower polynomial limits as follows

$$c_0\, t^{-\frac{1}{q}} \lesssim E_\kappa(t) \lesssim C_0\, t^{-\frac{1}{q}}, \quad t \to +\infty, \tag{6}$$

for some constants $0 < c_0 \leq C_0$ depending on the initial energy $E_\kappa(0)$, $\kappa \geq 0$. Indeed, such a claim corresponds to an intrinsic polynomial range of (uniform) stability and will follow as a consequence of a more general result that is rigorous stated in Theorem 2. See also Corollary 1. In particular, we can conclude that (3) is

not exponentially stable when dealing with weak initial data, that is, with solution in the standard energy space. See Corollary 2.

In conclusion, Theorem 2 truly reveals the stability of the associated energy $E_\kappa(t)$, which leads us to the concrete conclusions provided by Corollaries 1 and 2, being pioneering results on the subject. Due to technicalities in the well-posedness process, we shall work with $q \geq 1/2$. In Sect. 2 we prepare all notations and initial results. Then, all precise details on the stability results shall be given in Sect. 3.

1.1 Previous Literature, Comparisons and Highlights

In what follows, we are going to highlight that our approach and results are different or else provide generalized results, besides keeping more physical consistency in working exactly with (4) instead of modified versions of it. Indeed, there are at least three mathematical ways of attacking the energy damping coefficient (4) along Eq. (3) (or (2)), namely:

1. Keeping the potential energy in (4), but neglecting the kinetic one;
2. Keeping the kinetic energy in (4), but neglecting the potential one;
3. Keeping both potential and kinetic energies, but considering them under the action of a strictly (or not) positive function $M(\cdot)$ as a non-degenerate (or possibility degenerate) damping coefficient.

In the first case, equation (3) becomes to

$$\partial_{tt}u - \kappa\Delta u + \Delta^2 u + \alpha\left[\int_\Omega |\Delta u|^2 dx\right]^q \partial_t u = 0 \quad \text{in} \quad \Omega \times (0,\infty). \tag{7}$$

This is, for sure, the most challenging case once the damping coefficient becomes now to a real degenerate coefficient. In [5, Theorem 3.1], working on a bounded domain Ω with clamped boundary condition, it is proved the following with $q = 1$ in (7): *for every* $R > 0$, *there exist constants* $C_R = C(R) > 0$ *and* $\gamma_R = \gamma(R) > 0$ *depending on* R *such that*

$$E_\kappa(t) \leq C_R\, E_\kappa(0)\, e^{-\gamma_R t}, \quad t > 0, \tag{8}$$

only holds for every regular solution u *of* (3) *with initial data* (u_0, u_1) *satisfying*

$$\|(u_0, u_1)\|_{(H^4(\Omega)\cap H_0^2(\Omega))\times H_0^2(\Omega)} \leq R. \tag{9}$$

We stress that (8) only represents a *local stability result* since it holds on every ball with radius $R > 0$ in the strong topology $(H^4(\Omega) \cap H_0^2(\Omega)) \times H_0^2(\Omega)$, but they are not independent of the initial data. Moreover, as observed by the authors in [5], the drawback of (8) and (9) is that it could not be proved in the weak topology

$H_0^2(\Omega) \times L^2(\Omega)$, even taking initial data uniformly bounded in $H_0^2(\Omega) \times L^2(\Omega)$. Although we recognized that our results for (3) can not be fairly compared to such a result, we do can conclude by means of the upper and lower polynomial bounds (6) that the estimate (8) will never be reached for weak initial data given in $H_0^2(\Omega) \times L^2(\Omega)$. Therefore, our results act as complementary conclusions to [5] by clarifying such drawback raised therein, and yet giving a different point of view of stability by means of (6) and its consequences concerning problem (3).

In the second case, Eq. (3) falls into

$$\partial_{tt}u - \kappa\Delta u + \Delta^2 u + \alpha\left[\int_\Omega |\partial_t u|^2 dx\right]^q \partial_t u = 0 \quad \text{in} \quad \Omega \times (0,\infty). \tag{10}$$

Unlike the first case, here we have an easier setting because the kinetic damping coefficient provides a kind of monotonous (polynomial) damping whose computations to achieve (6) remain unchanged (and with less calculations). This means that all results highlighted previously still hold for this particular case. In addition, they clarify what is precisely the stability result related to problems addressed in [19, 20], which in turn represent particular models of abstract damping given by Aloui et al. [1, Section 8]. In other words, in terms of stability, our methodology provides a way to show the existence of absorbing sets with polynomial rate (and not faster than polynomial rate depending on q) when dealing with generalized problems relate to (10), subject that is not addressed in [19, 20].

Finally, in the third case let us see Eqs. (2) and (3) as follows

$$\partial_{tt}u - \kappa\Delta u + \Delta^2 u + M\left(\int_\Omega \left(|\Delta u|^2 + |\partial_t u|^2\right) dx\right) A\partial_t u = 0 \ \text{ in } \ \Omega \times (0,\infty), \tag{11}$$

where operator A represents the Laplacian operator $A = -\Delta$ or else the identity one $A = I$. Thus, here we clearly have two subcases, namely, when $M(\cdot) \geq 0$ is a non-degenerate or possibly degenerate function. For instance, when $M(s) = \alpha s^q$, $s \geq 0$, and $A = -\Delta$, then we go back to problem (2). For this (degenerate) nonlocal strong damping situation with $q \geq 1$, it is considered in [11, Theorem 3.1] an upper polynomial stability for the corresponding energy, which also involves a standard nonlinear source term. Nonetheless, we call the attention to the following prediction result provided in [11, Theorem 4.1] for (2) addressed on a bounded domain Ω with clamped boundary condition and $q \geq 1$: *By taking finite initial energy* $0 < E_\kappa(0) < \infty$, *then* $E_\kappa(t)$ *given in* (5) *satisfies*

$$E_\kappa(t) \leq 3E_\kappa(0)e^{-\delta\int_0^t \|u(s)\|^{2q} ds}, \quad t > 0, \tag{12}$$

where $\delta = \delta(\frac{1}{E_\kappa(0)}) > 0$ *is a constant proportional to* $1/E_\kappa(0)$.

Although the estimate (12) provides a new result with an *exponential face*, it does not mean any kind of stability result. Indeed, it is only a peculiar estimate

indicating that prevents exponential decay patterns as remarked in [11, Section 4]. In addition, it is worth pointing out that our computations to reach the stability result for problem (3) can be easily adjusted to (2), even for $q \geq 1/2$ thanks to a inequality provided in [1, Lemma 2.2]. Therefore, through the polynomial range (6) we provide here a much more accurate stability result than the estimate expressed by (12), by concluding indeed that both problems (2) and (3) are never exponentially stable in the topology of the energy space.

On the other hand, in the non-degenerate case $M(s) > 0$, $s \geq 0$, but still taking $A = -\Delta$, a generalized version of (11) has been recently approached by Sun and Yang [16] in a context of *strong attractors*, that is, the existence of attractors in the topology of more regular space than the weak phase space. In this occasion, the C^1-regularity for $M > 0$ brings out the non-degeneracy of the damping coefficient, which in turn allowed them to reach interesting results on well-posedness, regularity and long-time behavior of solutions over more regular spaces. Such assumption of positiveness for the damping coefficient has been also addressed by other authors for related problems, see e.g. [8–10]. From our point of view, in spite of representing a nice case, the latter does not portray the current situation of this paper so that we do not provide more detailed comparisons with such a non-degenerate problems, but we refer to [5, 8–11, 16] for a nice survey on this kind of non-degenerate damping coefficients. Additionally, we note that the suitable case of non-degenerate damping coefficient $M(s) > 0$, $s \geq 0$, and $A = I$ in (11) has not been considered in the literature so far and shall be concerned in another work by the authors in the future.

At light of the above statements, one sees e.g. when $M(s) = \alpha s^q$, $s \geq 0$, and $A = I$, then problem (11) falls into (3), being a problem not yet addressed in the literature that brings out a new branch of studies for such a nonlocal (possibly degenerate) damped problems, and also justifies all new stability results previously specified.

2 The Problem and Well-Posedness

Let us consider again the beam model with energy damping

$$\partial_{tt}u+\Delta^2 u-\kappa\,\Delta u+\alpha\left[\int_\Omega \left(|\partial_t u|^2+|\Delta u|^2\right)dy\right]^q \partial_t u = 0 \text{ in } \Omega\times(0,\infty), \tag{13}$$

with clamped boundary condition

$$u = \frac{\partial u}{\partial \nu} = 0 \text{ on } \partial\Omega\times[0,\infty), \tag{14}$$

and initial data

$$u(x,0) = u_0(x), \quad \partial_t u(x,0) = u_1(x), \quad x \in \Omega. \tag{15}$$

To address problem (13)–(15), we introduce the Hilbert phase space (still called *energy space*)

$$\mathcal{H} := H_0^2(\Omega) \times L^2(\Omega),$$

equipped with the inner product $\langle z^1, z^2 \rangle_{\mathcal{H}} := \langle \Delta u^1, \Delta u^2 \rangle + \langle v^1, v^2 \rangle$ for $z^i = (u^i, v^i) \in \mathcal{H}$, $i = 1, 2$, and norm $\|z\|_{\mathcal{H}} = \left(\|\Delta u\|^2 + \|v\|^2\right)^{1/2}$, for $z = (u, v) \in \mathcal{H}$, where $\langle u, v \rangle := \displaystyle\int_\Omega uv\, dx$, $\|u\|^2 := \langle u, u \rangle$ and $\|z\|_{\mathcal{H}}^2 := \langle z, z \rangle_{\mathcal{H}}$.

In order to establish the well-posedness of (13)–(15), we define the vector-valued function $z(t) := (u(t), v(t))$, $t \geq 0$, with $v = \partial_t u$. Then we can rewrite system (13)–(15) as the following first order abstract problem

$$\begin{cases} \partial_t z = \mathcal{A}z + \mathcal{M}(z), \quad t > 0, \\ z(0) = (u_0, u_1) := z_0, \end{cases} \tag{16}$$

where $\mathcal{A} : \mathcal{D}(\mathcal{A}) \subset \mathcal{H} \to \mathcal{H}$ is the linear operator given by

$$\mathcal{A}z = (v, -\Delta^2 u), \quad \mathcal{D}(\mathcal{A}) := H^4(\Omega) \cap H_0^2(\Omega), \tag{17}$$

and $\mathcal{M} : \mathcal{H} \to \mathcal{H}$ is the nonlinear operator

$$\mathcal{M}(z) = (0, \kappa \Delta u - \alpha \|z\|_{\mathcal{H}}^{2q} v), \quad z = (u, v) \in \mathcal{H}. \tag{18}$$

Therefore, the existence and uniqueness of solution to the system (13)–(15) relies on the study of problem (16). Accordingly, we have the following well-posedness result.

Theorem 1 *Let $\kappa, \alpha \geq 0$ and $q \geq \frac{1}{2}$ be given constants. If $z_0 \in \mathcal{H}$, then (16) has a unique mild solution z in the class $z \in C([0, \infty), \mathcal{H})$.*

In addition, if $z_0 \in \mathcal{D}(\mathcal{A})$, then z is a regular solution lying in the class

$$z \in C([0, \infty), \mathcal{D}(\mathcal{A})) \cap C^1([0, \infty), \mathcal{H}).$$

Proof To show the local version of the first statement, it is enough to prove that $\mathcal{A}$ given in (17) is the infinitesimal generator of a C_0-semigroup of contractions $e^{\mathcal{A}t}$ (which is very standard) and $\mathcal{M}$ set in (18) is locally Lipschitz on $\mathcal{H}$ which will be done next. Indeed, let $r > 0$ and $z^1, z^2 \in \mathcal{H}$ such that $\max\{\|z^1\|_{\mathcal{H}}, \|z^2\|_{\mathcal{H}}\} \leq r$. We note that

$$\begin{aligned} \left\| \|z^1\|_{\mathcal{H}}^{2q} v^1 - \|z^2\|_{\mathcal{H}}^{2q} v^2 \right\| &\leq \left[\|z^1\|_{\mathcal{H}}^{2q} + \|z^2\|_{\mathcal{H}}^{2q} \right] \|v^1 - v^2\| \\ &\quad + \left| \|z^1\|_{\mathcal{H}}^{2q} - \|z^2\|_{\mathcal{H}}^{2q} \right| \|v^1 + v^2\|. \end{aligned} \tag{19}$$

The first term on the right side of (19) can be estimated by

$$\left[\|z^1\|_{\mathcal{H}}^{2q} + \|z^2\|_{\mathcal{H}}^{2q}\right]\|v^1 - v^2\| \leq 2r^{2q}\|z^1 - z^2\|_{\mathcal{H}}.$$

Now, from a suitable inequality provided in [1][1] we estimate the second term as follows

$$\left|\|z^1\|_{\mathcal{H}}^{2q} - \|z^2\|_{\mathcal{H}}^{2q}\right|\|v^1 + v^2\| \leq 4qr^{2q}\|z^1 - z^2\|_{\mathcal{H}}.$$

Plugging the two last estimates in (19), we obtain

$$\left\|\|z^1\|_{\mathcal{H}}^{2q}v^1 - \|z^2\|_{\mathcal{H}}^{2q}v^2\right\|_{\mathcal{H}} \leq 2(2q+1)r^{2q}\|z^1 - z^2\|_{\mathcal{H}}.$$

Thus,

$$\|\mathcal{M}(z^1) - \mathcal{M}(z^2)\|_{\mathcal{H}} \leq \left(\kappa + 2(2q+1)\alpha r^{2q}\right)\|z^1 - z^2\|_{\mathcal{H}},$$

and $\mathcal{M}$ is locally Lipschitz in $\mathcal{H}$.

Hence, according to Pazy [15, Chapter 6], if $z_0 \in \mathcal{H}$ ($z_0 \in D(\mathcal{A})$), there exists a time $t_{\max} \in (0, +\infty]$ such that (16) has a unique mild (regular) solution

$$z \in C([0, t_{\max}), \mathcal{H}) \ (z \in C([0, t_{\max}), D(\mathcal{A})) \cap C^1([0, t_{\max}), \mathcal{H})).$$

Moreover, such time $t_{\max}$ satisfies either the conditions $t_{\max} = +\infty$ or else $t_{\max} < +\infty$ with

$$\lim_{t \to t_{\max}^-} \|z(t)\|_{\mathcal{H}} = +\infty. \tag{21}$$

In order to show that $t_{\max} = +\infty$, we consider $z_0 \in D(\mathcal{A})$ and the corresponding regular solution z of (16). Taking the inner product in $\mathcal{H}$ of (16) with z, we obtain

$$\frac{1}{2}\frac{d}{dt}\left[\|z(t)\|_{\mathcal{H}}^2 + \kappa\|\nabla u(t)\|^2\right] + \alpha\|z(t)\|_{\mathcal{H}}^{2q}\|\partial_t u(t)\|^2 = 0 \quad t \in [0, t_{\max}). \tag{22}$$

Integrating (22) over $(0, t)$, $t \in [0, t_{\max})$, we get

$$\|z(t)\|_{\mathcal{H}} \leq (1 + c'\kappa)^{1/2}\|z_0\|_{\mathcal{H}}, \quad t \in [0, t_{\max}).$$

[1] See [1, Lemma 2.2]: *Let X be a normed space with norm $\|\cdot\|_X$. Then, for any $s \geq 1$ we have*

$$\left|\|u\|_X^s - \|v\|_X^s\right| \leq s\max\{\|u\|_X, \|v\|_X\}^{s-1}\|u - v\|_X, \quad \forall\, u, v \in X. \tag{20}$$

Here, the constant $c' > 0$ comes from the embedding $H_0^2(\Omega) \hookrightarrow H_0^1(\Omega)$. The last estimate contradicts (21). Hence, $t_{max} = +\infty$. Using a limit process, one can conclude the same result for mild solutions.

The proof of Theorem 1 is then complete.

3 Lower-Upper Polynomial Energy's Bounds

By means of the notations introduced in Sect. 2, we recall that the energy functional corresponding to problem (13)–(15) can be expressed by

$$E_\kappa(t) = \frac{1}{2}\left[\|(u(t), \partial_t u(t))\|_{\mathcal{H}}^2 + \kappa\|\nabla u(t)\|^2\right], \quad t \geq 0. \tag{23}$$

Our main stability result reveals that $E_\kappa(t)$ is squeezed by decreasing polynomial functions as follows.

Theorem 2 *Under the assumptions of Theorem 1, there exists an increasing function $\mathcal{J} : \mathbb{R}^+ \to \mathbb{R}^+$ such that the energy $E_\kappa(t)$ satisfies*

$$\left[2^{q+1}\alpha q t + \left[E_\kappa(0)\right]^{-q}\right]^{-1/q} \leq E_\kappa(t) \leq \left[\frac{q}{\mathcal{J}(E_\kappa(0))}(t-1)^+ + \left[E_\kappa(0)\right]^{-q}\right]^{-1/q}, \tag{24}$$

for all $t > 0$, where we use the standard notation $s^+ := (s + |s|)/2$.

Proof Taking the scalar product in $L^2(\Omega)$ of (13) with $\partial_t u$, we obtain

$$\frac{d}{dt}E_\kappa(t) = -\alpha||(u(t), \partial_t u(t))||_{\mathcal{H}}^{2q}\|\partial_t u(t)\|^2, \quad t > 0. \tag{25}$$

Let us prove the lower and upper estimates in (24) in the sequel.

Lower Bound We first note that

$$||(u(t), \partial_t u(t))||_{\mathcal{H}}^{2q}\|\partial_t u(t)\|^2 \leq 2^{q+1}\,[\,E_\kappa(t)\,]^{q+1},$$

and replacing it in (25), we get

$$\frac{d}{dt}E_\kappa(t) \geq -2^{q+1}\alpha\,[\,E_\kappa(t)\,]^{q+1}, \quad t > 0. \tag{26}$$

Thus, integrating (26) and proceeding a straightforward computation, we reach the first inequality in (24).

Upper Bound Now, we are going to prove the second inequality of (24). To do so, we provide some proper estimates and then apply a Nakao's result (cf. [13, 14]).

We start by noting that

$$||(u(t), \partial_t u(t)||_{\mathcal{H}}^{2q} \|\partial_t u(t)\|^2 \geq \|\partial_t u(t)\|^{2(q+1)}, \tag{27}$$

and replacing (27) in (25), we get

$$\frac{d}{dt} E_\kappa(t) + \alpha \|\partial_t u(t)\|^{2(q+1)} \leq 0, \quad t > 0, \tag{28}$$

which implies that $E_\kappa(t)$ is non-increasing with $E_\kappa(t) \leq E_\kappa(0)$ for every $t > 0$. Also, integrating (28) from t to $t+1$, we obtain

$$\alpha \int_t^{t+1} \|\partial_t u(s)\|^{2(q+1)}\, ds \leq E_\kappa(t) - E_\kappa(t+1) := [\, D(t)\,]^2. \tag{29}$$

Using Hölder's inequality with $\frac{q}{q+1} + \frac{1}{q+1} = 1$ and (29), we infer

$$\int_t^{t+1} \|\partial_t u(s)\|^2 ds \leq \frac{1}{\alpha^{\frac{1}{q+1}}} [\, D(t)\,]^{\frac{2}{q+1}}. \tag{30}$$

From the Mean Value Theorem for integrals, there exist $t_1 \in [t, t+\frac{1}{4}]$ and $t_2 \in [t+\frac{3}{4}, t+1]$ such that

$$\|\partial_t u(t_i)\|^2 \leq 4 \int_t^{t+1} \|\partial_t u(s)\|^2 ds \leq \frac{4}{\alpha^{\frac{1}{q+1}}} [\, D(t)\,]^{\frac{2}{q+1}}, \ i = 1, 2. \tag{31}$$

On the other hand, taking the scalar product in $L^2(\Omega)$ of (13) with u and integrating the result over $[t_1, t_2]$, we have

$$\begin{aligned} \int_{t_1}^{t_2} E_\kappa(s)\, ds = \int_{t_1}^{t_2} \|\partial_t u(s)\|^2\, ds + \frac{1}{2} [(\partial_t u(t_1), u(t_1)) - (\partial_t u(t_2), u(t_2))] \\ - \frac{\alpha}{2} \int_{t_1}^{t_2} ||(u(s), \partial_t u(s))||_{\mathcal{H}}^{2q} (\partial_t u(s), u(s))\, ds. \end{aligned} \tag{32}$$

Let us estimate the terms in the right side of (32). Firstly, we note that through Hölder's inequality, (31) and Young's inequality, we obtain

$$\begin{aligned}|(\partial_t u(t_1), u(t_1)) - (\partial_t u(t_2), u(t_2))| &\leq d \sum_{i=1}^{2} \|\partial_t u(t_i)\| \|\Delta u(t_i)\| \\ &\leq \frac{8d}{\alpha^{\frac{1}{2(q+1)}}} [\, D(t)\,]^{\frac{1}{q+1}} \sup_{t_1 \leq s \leq t_2} [E_\kappa(s)]^{1/2} \\ &\leq \frac{128\, d^2}{\alpha^{\frac{1}{q+1}}} [\, D(t)\,]^{\frac{2}{q+1}} + \frac{1}{8} \sup_{t_1 \leq s \leq t_2} E_\kappa(s),\end{aligned}$$

where the constant $d > 0$ comes from the embedding $H_0^2(\Omega) \hookrightarrow L^2(\Omega)$. Additionally, using that $E_\kappa(t) \leq E_\kappa(0)$, we have

$$\|(u(t), \partial_t u(t))\|_{\mathcal{H}}^{2q} \leq 2^q\, [\, E_\kappa(t)\,]^q \leq 2^q\, [\, E_\kappa(0)\,]^q\,.$$

From this and (30) we also get

$$\begin{aligned}\left| \int_{t_1}^{t_2} ||(u(s), \partial_t u(s))||_{\mathcal{H}}^{2q} (\partial_t u(s), u(s))\, ds \right| &\leq \frac{2^{2q+3} d^2\, [\, E_\kappa(0)\,]^{2q}}{\alpha^{-\frac{q}{q+1}}} [\, D(t)\,]^{\frac{2}{q+1}} \\ &\quad + \frac{1}{8\alpha} \sup_{t_1 \leq s \leq t_2} E_\kappa(s).\end{aligned}$$

Regarding again (30) and replacing the above estimates in (32), we obtain

$$\int_{t_1}^{t_2} E_\kappa(s)\, ds \leq \mathcal{K}(E_\kappa(0))\, [\, D(t)\,]^{\frac{2}{q+1}} + \frac{1}{8} \sup_{t_1 \leq s \leq t_2} E_\kappa(s), \tag{33}$$

where we set the function $\mathcal{K}$ as

$$\mathcal{K}(s) := \left[\frac{64\, d^2 + 1}{\alpha^{\frac{1}{q+1}}} + 2^{(q+1)} d^2 \alpha^{\frac{2q+1}{q+1}} s^{2q} \right] > 0.$$

Using once more the Mean Value Theorem for integrals and the fact that $E_\kappa(t)$ is non-increasing, there exists $\zeta \in [t_1, t_2]$ such that

$$\int_{t_1}^{t_2} E_\kappa(s)\, ds = E_\kappa(\zeta)(t_2 - t_1) \geq \frac{1}{2} E_\kappa(t+1),$$

and then

$$\sup_{t\le s\le t+1} E_\kappa(s) = E_\kappa(t) = E_\kappa(t+1) + [\,D(t)\,]^2 \le 2\int_{t_1}^{t_2} E_\kappa(s)\,ds + [\,D(t)\,]^2.$$

Thus, from this and (33), we arrive at

$$\begin{aligned}\sup_{t\le s\le t+1} E_\kappa(s) &\le [\,D(t)\,]^2 + 2\int_{t_1}^{t_2} E_\kappa(s)ds \\ &\le [\,D(t)\,]^2 + 2\mathcal{K}(E_\kappa(0))\,[\,D(t)\,]^{\frac{2}{q+1}} + \frac{1}{4}\sup_{t\le s\le t+1} E_\kappa(s),\end{aligned}$$

and since $0 < \frac{2}{q+1} \le 2$, we obtain

$$\sup_{t\le s\le t+1} E_\kappa(s) \le \frac{4}{3}[\,D(t)\,]^{\frac{2}{q+1}}\left[\,[\,D(t)\,]^{\frac{2q}{q+1}} + 2\mathcal{K}(E_\kappa(0))\right]. \tag{34}$$

Observing that $[\,D(t)\,]^{\frac{2q}{q+1}} \le [E_\kappa(t) + E_\kappa(t+1)]^{\frac{q}{q+1}} \le 2^{\frac{q}{q+1}}\,[E_\kappa(0)\,]^{\frac{q}{q+1}}$, and denoting by

$$\mathcal{J}(s) := \left(\frac{4}{3}\right)^{q+1}\left[(2s)^{\frac{q}{q+1}} + 2\mathcal{K}(s)\right]^{q+1} > 0, \tag{35}$$

and also recalling the definition of $[D(t)]^2$ in (29), we obtain from (34) that

$$\sup_{t\le s\le t+1} [E_\kappa(s)]^{q+1} \le \mathcal{J}(E_\kappa(0))\,[\,E_\kappa(t) - E_\kappa(t+1)\,].$$

Hence, applying e.g. Lemma 2.1 of [14] with $E_\kappa = \phi, \mathcal{J}(E_\kappa(0)) = C_0$, and $K = 0$, we conclude $E_\kappa(t) \le \left[\frac{q}{\mathcal{J}(E_\kappa(0))}(t-1)^+ + \frac{1}{\left[E_\kappa(0)\right]^q}\right]^{-1/q}$, which ends the proof of the second inequality in (24).

The proof of Theorem 2 is therefore complete.

Remark 1 It is worth point out that we always have

$$\left[\,2^{2q+1}\alpha q t + \left[E_\kappa(0)\right]^{-q}\,\right]^{-1/q} \le \left[\frac{q}{\mathcal{J}(E_\kappa(0))}(t-1)^+ + \left[E_\kappa(0)\right]^{-q}\right]^{-1/q}, \tag{36}$$

so that it makes sense to express $E_\kappa(t)$ between the inequalities in (24). Indeed, from the definition $\mathcal{J}$ in (35) one easily sees that $\mathcal{J}(E_\kappa(0)) \ge \frac{1}{2^{2q+1}\alpha}$, from where one concludes (36) promptly.

Corollary 1 (Polynomial Range of Decay) *Under the assumptions of Theorem 2, the energy functional $E_\kappa(t)$ defined in (23) decays squeezed as follows*

$$c_0\, t^{-\frac{1}{q}} \lesssim E_\kappa(t) \lesssim C_0\, t^{-\frac{1}{q}} \quad as \quad t \to +\infty, \tag{37}$$

for some constants $0 < c_0 \leq C_0$ depending on the initial energy $E_\kappa(0)$.

In other words, $E_\kappa(t)$ decays polynomially at rate $t^{-1/q}$ $(q \geq 1/2)$ as $t \to +\infty$. □

Corollary 2 (Non-exponential Stability) *Under the assumptions of Theorem 2, the energy $E_\kappa(t)$ set in (23) never decays exponentially as e^{-at} $(a > 0)$ as $t \to +\infty$.* □

References

1. Aloui, F., Ben Hassen, I., Haraux, A.: Compactness of trajectories to some nonlinear second order evolution equations and applications. J. Math. Pures Appl. **100**(3), 295–326 (2013)
2. Balakrishnan, A.V.: A theory of nonlinear damping in flexible structures. In: Stabilization of Flexible Structures, pp. 1–12 (1988)
3. Balakrishnan, A.V., Taylor, L.W.: Distributed parameter nonlinear damping models for flight structures. In: Proceedings Daming 89, Flight Dynamics Lab and Air Force Wright Aeronautical Labs, WPAFB (1989)
4. Bass, R.W., Zes, D.: Spillover, nonlinearity, and flexible structures. In: L.W. Taylor (ed.) The Fourth NASA Workshop on Computational Control of Flexible Aerospace Systems, NASA Conference Publication 10065, pp. 1–14 (1991)
5. Cavalcanti, M.M., Domingos Cavalcanti, V.N., Jorge Silva, M.A., Narciso, V.: Stability for extensible beams with a single degenerate nonlocal damping of Balakrishnan-Taylor type. J. Differential Equations **290**, 197–222 (2021)
6. Dowell, E.H.: Aeroelasticity of Plates and Shells. Noordhoff Int. Publishing Co., Groninger, NL (1975)
7. Hughes, T.J., Marsden, J.E.: Mathematical Foundation of Elasticity. Prentice-Hall, Englewood Cliffs (1983)
8. Jorge Silva, M.A., Narciso, V.: Long-time behavior for a plate equation with nonlocal weak damping. Differential Integral Equations **27**(9–10), 931–948 (2014)
9. Jorge Silva, M.A., Narciso, V.: Attractors and their properties for a class of nonlocal extensible beams. Discrete Contin. Dyn. Syst. **35**(3), 985–1008 (2015)
10. Jorge Silva, M.A., Narciso, V.: Long-time dynamics for a class of extensible beams with nonlocal nonlinear damping. Evol. Equations Control Theory **6**(3), 437–470 (2017)
11. Jorge Silva, M.A., Narciso, V., Vicente, A.: On a beam model related to flight structures with nonlocal energy damping. Discrete Contin. Dyn. Syst. Ser. B **24**, 3281–3298 (2019)
12. Mu, C., Ma, J.: On a system of nonlinear wave equations with Balakrishnan-Taylor damping. Z. Angew. Math. Phys. **65**, 91–113 (2014)
13. Nakao, M.: Convergence of solutions of the wave equation with a nonlinear dissipative term to the steady state. Mem. Fac. Sci. Kyushu Univ. Ser. A **30**, 257–265 (1976)
14. Nakao, M.: A difference inequality and its application to nonlinear evolution equations. J. Math. Soc. Jpn. **30**(4), 747–762 (1978)
15. Pazy, A.: Semigroups of Linear Operators and Applications to Partial Differential Equations, vol. 44. Springer (1983)

16. Sun, Y., Yang, Z.: Strong attractors and their robustness for an extensible beam model with energy damping. Discrete Contin. Dyn. Syst. (2021). https://doi.org/10.3934/dcdsb.2021175
17. You, Y.: Inertial manifolds and stabilization of nonlinear beam equations with Balakrishnan-Taylor damping. Abstracts Appl. Anal. **1**(1), 83–102 (1996)
18. Zhang, W.: Nonlinear damping model: response to random excitation. In: 5th Annual NASA Spacecraft Control Laboratory Experiment (SCOLE) Workshop, pp. 27–38 (1988)
19. Zhao, C., Zhao, C., Zhong, C.: The global attractor for a class of extensible beams with nonlocal weak damping. Discrete Contin. Dyn. Syst. B **25**, 935–955 (2020)
20. Zhao, C., Ma, S., Zhong, C.: Long-time behavior for a class of extensible beams with nonlocal weak damping and critical nonlinearity. J. Math. Phys. **61**, 032701 (2020)

On the Wave-Like Energy Estimates of Klein-Gordon Type Equations with Time Dependent Potential

Kazunori Goto and Fumihiko Hirosawa

Abstract We consider the conditions for the time dependent potential in which the energy of the Cauchy problem of Klein-Gordon type equation asymptotically behaves like the energy of the wave equation. The conclusion of this paper is that the condition is not always given by the order of the potential itself, but should be given by "generalized zero mean condition", which is represented by the integral of the potential. We also introduce "generalized modified energy conservation" in order to describe the appropriate energy for our problem.

1 Introduction

Let us consider the following Cauchy problem for Klein-Gordon type equation with time dependent potential:

$$\begin{cases} \left(\partial_t^2 - \Delta + M(t)\right) u(t,x) = 0, & (t,x) \in (0,\infty) \times \mathbb{R}^n, \\ u(0,x) = u_0(x), \quad \partial_t u(0,x) = u_1(x), & x \in \mathbb{R}^n, \end{cases} \tag{1}$$

where Δ denotes the Laplace operator in $\mathbb{R}^n$ and the potential M is real valued but not necessarily a definite sign. It may be natural that M is positive from the point of view of the physical model, but we study it as a mathematical model and remove the restriction.

It is well known that the energy conservation holds if M is a non-negative constant, and in the case of general M, the following property of generalized energy conservation of Klein-Gordon type is proved in [1, 2]:

$$q(t)^2 E_{KG}(u;p)(0) \lesssim E_{KG}(u;p)(t) \lesssim E_{KG}(u;p)(0) \tag{2}$$

K. Goto · F. Hirosawa (✉)
Yamaguchi University, Yamaguchi, Japan
e-mail: b003vbv@yamaguchi-u.ac.jp; hirosawa@yamaguchi-u.ac.jp

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_48

for positive decreasing functions p and q under appropriate conditions to M, where

$$E_{KG}(u;p) := \|\nabla u(t,\cdot)\|_{L^2}^2 + \|\partial_t u(t,\cdot)\|_{L^2}^2 + p(t)\|u(t,\cdot)\|_{L^2}^2.$$

More precisely, if $M = \mu^2(1+t)^{-2\nu}$ with $\mu > 0$ and $0 \le \nu \le 1$, then p and q are given by $p = (1+t)^{-\nu_1}$ with $\nu_1 < 2$ and $q = (1+t)^{-\nu_2}$, respectively, where ν_1 and ν_2 are determined by μ and ν. On the other hand, if $\nu > 1$, that is, $\sqrt{|M|} \in L^1([0,\infty))$, then the solution has more wave-like property. In [7], the following model is studied as a perturbation problem of [2]:

$$M = \mu^2(1+t)^{-2} + \delta(t). \tag{3}$$

A conclusion of [7] is that the same estimate from above in (2) as in the case $\delta = 0$ is valid under some suitable assumptions to $\delta(t)$ which permit $\limsup_{t\to\infty}(1+t)^2\delta(t) = \infty$. The main purposes of this paper is to determine the conditions for $\delta(t)$ of (3) with $\mu = 0$ that the *generalized modified energy conservation* of wave type defined later, is established. From another point of view, we will determine *generalized zero mean condition* for M that (1) has wave-like property in spite of $\sqrt{|M|} \notin L^1([0,\infty))$.

2 Main Theorem

For $b \in C^0([0,\infty))$ satisfying $\lim_{t\to\infty} b(t) = 0$ and large T, we define the modified energy of the wave type $E(u;b)$ and *generalized modified energy conservation* by

$$E(u;b)(t) := \|\nabla u(t,\cdot)\|_{L^2}^2 + \|\partial_t u(t,\cdot) - b(t)u(t,\cdot)\|_{L^2}^2$$

and

$$E(u;b)(t) \simeq E(u;b)(T) \quad (t \ge T). \tag{4}$$

For M, we introduce the following properties with parameters α, β and γ:

(M1) For $\alpha \le 1$:

$$\int_0^t \left| \int_s^\infty \int_\sigma^\infty M(\tau)\, d\tau\, d\sigma \right| ds \lesssim (1+t)^\alpha \quad (\alpha \ge 0) \tag{5}$$

$$\int_t^\infty \left| \int_s^\infty \int_\sigma^\infty M(\tau)\, d\tau\, d\sigma \right| ds \lesssim (1+t)^\alpha \quad (\alpha \le 0). \tag{6}$$

(M2) For $\beta < 1$:

$$|M(t)| \lesssim (1+t)^{-2\beta}. \tag{7}$$

(M3) For $\gamma > 0$:

$$\left| \int_t^\infty M(s)\, ds \right| \lesssim (1+t)^{-\gamma}, \tag{8}$$

$$\int_t^\infty \left(\int_s^\infty M(\sigma)\, d\sigma \right)^2 ds \lesssim (1+t)^{-\gamma} \tag{9}$$

and

$$\int_0^\infty \int_t^\infty \left(\int_s^\infty M(\sigma)\, d\sigma \right)^2 ds\, dt < \infty. \tag{10}$$

Remark 1

(i) The following estimate is implicitly assumed in (5):

$$\left| \int_0^\infty \int_t^\infty M(s)\, ds\, dt \right| < \infty. \tag{11}$$

(ii) If (9) holds for $\gamma > 1$, then (5) with $\alpha = -\gamma + 2$ is trivial . Moreover, if $\gamma > 2$, then (6) with $\alpha = -\gamma + 2$ is trivial.

Theorem 1 *Let $u_0 \in H^1$ and $u_1 \in L^2$. If (M1), (M2) and (M3) are valid for*

$$\gamma \geq \beta \begin{cases} \geq (\alpha+1)/2 & \text{for } \alpha \neq 0, \\ > 1/2 & \text{for } \alpha = 0, \end{cases} \tag{12}$$

and the following estimate holds:

$$\sup_{t \geq 0} \left\{ (1+t)^\alpha \int_t^\infty \int_s^\infty \left(\int_\sigma^\infty M(\tau)\, d\tau \right)^2 d\sigma\, ds \right\} < \infty, \tag{13}$$

then there exist $T > 0$ and $b \in C^0([0,\infty))$ satisfying $b(t) \lesssim (1+t)^{-\gamma}$ such that (4) is established. Moreover, the following estimate is established for any $t \geq 0$:

$$E(u;b)(t) \lesssim E_{KG}(u;1)(0). \tag{14}$$

In [7], $\beta \geq (-\gamma+3)/2$ is assumed instead of (12) without assuming (M1), that is, only the trivial case $\alpha = -\gamma+2$ in Remark 1 (ii), is considered. The following M is an example of the non-trivial case $\alpha < -\gamma+2$.

Example

Let $M(t) := \frac{d}{dt}(\sin((1+t)^\kappa)(1+t)^{-2\beta-\kappa+1})$ with $\beta \leq 1/2$ and $\kappa > 2(1-\beta)$. Noting the estimates $|M(t)| \lesssim (1+t)^{-2\beta}$, $\int_t^\infty M(s)\,ds = \sin((1+t)^\kappa)(1+t)^{-2\beta-\kappa+1}$ and $|\int_t^\infty \int_s^\infty M(\sigma)\,d\sigma\,ds| \lesssim (1+t)^{-2\beta-2\kappa+2}$, (M1), (M2) and (M3) are valid for $\gamma = 2\beta+\kappa-1$ and $\alpha = -2\beta-2\kappa+3 = -\gamma-\kappa+2$, it follows that $\alpha < -\gamma+2$. Moreover, (12) and (13) are valid by $\gamma > 1 > \beta = (\alpha+1)/2+2\beta+\kappa-2 > (\alpha+1)/2$ and $\alpha-2\gamma+2 = -6\beta-4\kappa+7 < 2\beta-1 \leq 0$.

The conditions (M1)–(M3) seem to be artificial, but they can be actually natural from the viewpoint of previous studies. (M1) and (M2) correspond to *stabilization property* and C^2*-property* with *very fast oscillation*, respectively, which were introduced in [4, 5] for the energy estimate of the wave equation with time dependent propagation speed. Equations (9) and (11) are corresponding to *generalized zero mean condition*, which was introduced in [8]. Moreover, (10) is considered to be related to the classification of scale invariant potential for the Klein-Gordon type equation in [2].

3 Proof of the Theorem

The proof of the theorem is based on the methods introduced in [3, 6, 7] that the Klein-Gordon type equation is reduced to a dissipative wave equation or a wave equation with time dependent propagation speed. Then, solutions of the equations are estimated in a particular zones of time-frequency space by the method introduced in [5, 8] after the Fourier transformation with respect to spatial variables.

3.1 Reduction to a Dissipative Wave Equation

For $t \geq T$ with a large T, we reduce the Klein-Gordon type equation of (1) to the dissipative wave equation $(\partial_t^2 - \Delta + 2b(t)\partial_t)w = 0$ by the transformation

$$w(t,x) := \exp\left(\int_t^\infty b(s)\,ds\right) u(t,x),$$

where b is a solution of the following Riccati equation

$$b'(t) + b(t)^2 + M(t) = 0. \tag{15}$$

Let us derive the representation of a particular solution of (15). We define $\{q_k(t)\}_{k=1}^{\infty}$ and $\{Q_k(t)\}_{k=1}^{\infty}$ on $[T, \infty)$ by

$$q_1(t) := M(t), \quad q_k(t) := \sum_{j=1}^{k-1} Q_j(t) Q_{k-j}(t) \quad (k = 2, 3, \ldots)$$

and

$$Q_k(t) := -\int_t^{\infty} q_k(s)\, ds \quad (k = 1, 2, \ldots).$$

Lemma 1 *A particular solution of* (15) *is represented by* $b(t) := \sum_{k=1}^{\infty} Q_k(t)$.

Proof The proof is straightforward calculation. □

The following lemmas ensure the convergence of $b(t)$ on $[T, \infty)$ for large T.

Lemma 2 $Q_2(t) \le 0$ *and the following estimate is established for any* $k \ge 2$:

$$|Q_k(t)| \le 4^{k-1}(-Q_2(t))\phi(t)^{\frac{k-2}{2}}, \quad \phi(t) := -\int_t^{\infty} Q_2(s)\, ds. \tag{16}$$

Proof $Q_2(t) \le 0$ is trivial from the definition, and $\lim_{t\to\infty} Q_2(t) = 0$ by (9). Equation (16) is trivial for $k = 2$. If (16) is valid for $k = 3, \ldots, l$, then by Cauchy-Schwarz inequality, integration by parts, noting $\phi'(t) = Q_2(t)$ and $\frac{d}{dt} Q_2(t) = q_2(t) \ge 0$, we have

$$\begin{aligned}
\left| \int_t^{\infty} Q_1(s) Q_l(s)\, ds \right| &\le 4^{l-1} \left(\int_t^{\infty} Q_1(s)^2\, ds \right)^{\frac{1}{2}} \left(\int_t^{\infty} Q_2(s)^2 \phi(s)^{l-2}\, ds \right)^{\frac{1}{2}} \\
&= \frac{4^{l-1}}{(l-1)^{\frac{1}{2}}} (-Q_2(t))^{\frac{1}{2}} \left(\int_t^{\infty} Q_2(s) \frac{d}{ds} \phi(s)^{l-1}\, ds \right)^{\frac{1}{2}} \\
&\le 4^{l-1}(-Q_2(t))\phi(t)^{\frac{l-1}{2}}.
\end{aligned}$$

Moreover, for $2 \le j \le l$ we have

$$\begin{aligned}
\left| \int_t^{\infty} Q_j(s) Q_{l+1-j}(s)\, ds \right| &\le 4^{l-1} \int_t^{\infty} Q_2(s)^2 \phi(s)^{\frac{l-3}{2}}\, ds \\
&= \frac{4^l}{2(l-1)} \int_t^{\infty} Q_2(s) \frac{d}{ds} \phi(s)^{\frac{l-1}{2}}\, ds \\
&\le \frac{4^l (-Q_2(t)) \phi(t)^{\frac{l-1}{2}}}{2(l-1)}.
\end{aligned}$$

Therefore, we obtain

$$|Q_{l+1}(t)| \le 2\left|\int_t^\infty Q_1(\sigma)Q_l(\sigma)\,d\sigma\right| + \sum_{j=2}^{l-1}\left|\int_t^\infty Q_j(\sigma)Q_{l+1-j}(\sigma)\,d\sigma\right|$$
$$= 4^l(-Q_2(t))\phi(t)^{\frac{l-1}{2}}\left(\frac{1}{2} + \sum_{j=2}^{l-1}\frac{1}{2(l-1)}\right) \le 4^l(-Q_2(t))\phi(t)^{\frac{l-1}{2}},$$

it follows that (16) is also valid for any $k \le l+1$. Thus (16) is valid for any $k \ge 2$. □

Lemma 3 *There exist positive constants T, $b_0 = b_0(T)$, $b_1 = b_1(T)$ and $b_2 = b_2(T)$ such that the following estimates are established for any $t \ge T$:*

$$\sum_{k=2}^{\infty}|Q_k(t)| \le \frac{3}{2}|Q_2(t)|, \tag{17}$$

$$\left|\int_t^\infty b(s)\,ds\right| \le b_0, \tag{18}$$

$$|b(t)| \le b_1(1+t)^{-\gamma} \tag{19}$$

and

$$|b'(t)| \le b_2(1+t)^{-2\beta}. \tag{20}$$

Proof By (8), (10) and (11) there exists $T > 0$ such that

$$\left|\int_t^\infty Q_1(s)\,ds\right| \le 1 \quad \text{and} \quad \phi(t) \le \frac{1}{6^4} \tag{21}$$

for any $t \ge T$. Then, by Lemma 2, we have

$$\left|\sum_{k=3}^{\infty}Q_k(t)\right| \le |Q_2(t)|\sum_{k=3}^{\infty}4^{k-1}\phi(t)^{\frac{k-2}{2}} \le \frac{1}{2}|Q_2(t)|,$$

which gives (17). By (10), (11) and (17), we have

$$\left|\int_t^\infty b(s)\,ds\right| \le \left|\int_t^\infty Q_1(s)\,ds\right| + \sum_{k=2}^{\infty}\int_t^\infty |Q_k(s)|\,ds \le b_0. \tag{22}$$

By (8), (9) and (17), we have

$$|b(t)| \le |Q_1(t)| + \frac{3}{2}|Q_2(t)| \le b_1(1+t)^{-\gamma}.$$

By (7), (12), (15) and (19), we have

$$|b'(t)| \le |M(t)| + b_1^2(1+t)^{-2\gamma} \le b_2(1+t)^{-2\beta}.$$

Thus the proof is concluded. □

Lemma 3 ensures that the solution of (1) is represented by the solution of the following dissipative wave equation:

$$(\partial_t^2 - \Delta + 2b(t)\partial_t)w(t,x) = 0 \tag{23}$$

for $t \ge T$. By carrying out partial Fourier transformation with respect to spatial variables and denoting the Fourier image of $w(t,x)$ as $\hat{w}(t,\xi)$, (23) is represented as follows:

$$(\partial_t^2 + |\xi|^2 + 2b(t)\partial_t)\hat{w}(t,\xi) = 0. \tag{24}$$

Moreover, (24) is represented by the following first order system:

$$\partial_t W = AW, \quad A := \begin{pmatrix} -2b(t) & i|\xi| \\ i|\xi| & 0 \end{pmatrix}, \quad W := \begin{pmatrix} \partial_t w \\ i|\xi| w \end{pmatrix}. \tag{25}$$

We estimate the solution of (25) in different ways in the following two zones of the time-frequency space $[T,\infty) \times \mathbb{R}^n$:

$$\begin{cases} Z_H := \{(t,\xi) \in [T,\infty) \times \mathbb{R}^n \ ; \ (1+t)^\alpha |\xi| \ge N\}, \\ Z_\Psi := \{(t,\xi) \in [T,\infty) \times \mathbb{R}^n \ ; \ (1+t)^\alpha |\xi| \le N\}, \end{cases}$$

where N is a positive constant which will be chosen later. Denoting

$$t_\xi := \max\left\{T, (N|\xi|^{-1})^{\frac{1}{\alpha}} - 1\right\}$$

for $\alpha \ne 0$ and $|\xi| > 0$, we see that $Z_H = \{t \ge t_\xi\}$ and $Z_\Psi = \{T \le t \le t_\xi\}$ for $\alpha > 0$, and that $Z_H = \{T \le t \le t_\xi\}$ and $Z_\Psi = \{t \ge t_\xi\}$ for $\alpha < 0$.

3.2 Estimate in Z_H

Proposition 1 *There exist positive constants N and K_1 such that the following estimates are established in Z_H:*

$$\begin{cases} K_1^{-1}|W(t_\xi,\xi)| \le |W(t,\xi)| \le K_1|W(t_\xi,\xi)| & (\alpha > 0), \\ K_1^{-1}|W(T,\xi)| \le |W(t,\xi)| \le K_1|W(T,\xi)| & (\alpha \le 0). \end{cases}$$

Proof Let $(t,\xi) \in Z_H$. Setting $N \ge 2b_1$, by (12) and (19) we have $|b(t)| \le b_1(1+t)^{-\gamma} \le b_1(1+t)^{-\alpha} \le b_1 N^{-1}|\xi| \le |\xi|/2$. Since the eigenvalues and the respective eigenvectors of A are given by $\{\lambda, \overline{\lambda}\}$ and $\{{}^t(1, i\delta), {}^t(-i\delta, 1)\}$, where $\lambda = -b(t) - i\sqrt{|\xi|^2 - b^2}$ and $\delta = \lambda|\xi|^{-1}$, and noting the inequalities

$$\sqrt{3} \le |1-\delta^2| = 2\sqrt{1 - b^2|\xi|^{-2}} \le 2, \tag{26}$$

A is diagonalized as $M^{-1}AM = \mathrm{diag}(\lambda, \overline{\lambda}) =: \Lambda$ by the diagonalizer

$$M := \begin{pmatrix} 1 & -i\delta \\ i\delta & 1 \end{pmatrix}.$$

Denoting $W_1 := M^{-1}W$, (25) is rewritten as follows:

$$\partial_t W_1 = \left(\Lambda - M^{-1}(\partial_t M)\right) W_1 = (\Lambda_1 + R_1)\, W_1,$$

where

$$\Lambda_1 := \left(-b - \frac{\partial_t \log\left(1-\delta^2\right)}{2}\right) I - i\sqrt{|\xi|^2 - b^2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

and

$$R_1 := -i\frac{b'|\xi|^{-1}}{2\left(1 - b^2|\xi|^{-2}\right)} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Then, by (26) we have

$$\partial_t |W_1|^2 \lessgtr -\left(2b + \Re\left(\partial_t \log\left(1-\delta^2\right)\right) \pm \frac{4}{3}|b'||\xi|^{-1}\right)|W_1|^2.$$

Noting that (12) and (20) conclude the following estimates:

$$|\xi|^{-1}\int_{t_\xi}^{\infty} |b'(s)|\, ds < \infty \ (\alpha \ge 0) \quad \text{and} \quad |\xi|^{-1}\int_{T}^{t_\xi} |b'(s)|\, ds < \infty \ (\alpha < 0),$$

by Lemma 3, (26) and Gronwall's inequality, we have $|W_1(t,\xi)| \simeq |W_1(t_\xi,\xi)|$ for $\alpha > 0$ and $|W_1(t,\xi)| \simeq |W_1(T,\xi)|$ for $\alpha \le 0$ in Z_H. Finally, noting that $|\delta|^2 = 1$ and (26) gives $\sqrt{1/2}|W| \le |W_1| \le \sqrt{2/3}|W|$, we conclude the proof. □

3.3 Estimate in Z_Ψ

Proposition 2 *There exist a positive constant K_2 such that the following estimates are established in Z_Ψ:*

$$\begin{cases} K_2^{-1}|W(T,\xi)| \le |W(t,\xi)| \le K_2|W(T,\xi)| & (\alpha > 0), \\ K_2^{-1}|W(t_\xi,\xi)| \le |W(t,\xi)| \le K_2|W(t_\xi,\xi)| & (\alpha \le 0). \end{cases}$$

Proof Let us introduce the change of variable from $t \in [T,\infty)$ to $\theta \in [0,\infty)$ by

$$\theta := \int_T^t \exp\left(-2\int_T^s b(\sigma)\,d\sigma\right) ds.$$

Here we note that $\theta(t)$ is strictly increasing and satisfying $e^{-2b_0}t \le \theta(t) \le e^{2b_0}t$ by (18). We define $a(\tau)$ and $\eta(\tau)$ by

$$a(\tau) := \exp\left(2\int_T^{\theta^{-1}(\tau)} b(s)\,ds\right)$$

and

$$\eta(\tau) := \exp\left(2\int_T^\infty Q_1(s)\,ds + 2\int_T^{\theta^{-1}(\tau)} \sum_{k=2}^\infty Q_k(s)\,ds\right).$$

Here we remark that $e^{-2b_0} \le a(\tau)$, $\eta(\tau) \le e^{2b_0}$ and $\eta'(\tau) \ge 0$ are valid by Lemma 2, (17) and (22). By mean value theorem, (5) and (21), there exist constants $a_0 > 0$ and $0 < \kappa < 1$ such that the following estimates are established:

$$\begin{aligned} \int_{\theta(T)}^{\theta(t)} |a(\sigma)-\eta(\sigma)|\,d\sigma &= \int_T^t \left|1-\exp\left(2\int_s^\infty Q_1(\sigma)\,d\sigma\right)\right| ds \\ &= \int_T^t \left|2\int_s^\infty Q_1(\sigma)\,d\sigma \exp\left(2\kappa\int_s^\infty Q_1(\sigma)\,d\sigma\right)\right| ds \\ &\le 2e^{2\kappa}\int_0^t \left|\int_s^\infty Q_1(\sigma)\,d\sigma\right| ds \le a_0(1+t)^\alpha. \end{aligned}$$

Moreover, if (6) holds for $\alpha \le 0$, then we have $\int_{\theta(t)}^\infty |a(\sigma)-\eta(\sigma)|\,d\sigma \le a_0(1+t)^\alpha$.

By the change of variables $t \to \theta$ and denoting $y(\theta(t), \xi) = \hat{w}(t, \xi)$, (25) is represented by

$$\partial_\theta Y = BY, \quad Y := \begin{pmatrix} \partial_\theta y + i\eta|\xi|y \\ \partial_\theta y - i\eta|\xi|y \end{pmatrix}$$

and

$$B := \frac{i|\xi|\left(a^2+\eta^2\right)}{2\eta}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} + \frac{\eta'}{2\eta}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} + \frac{i|\xi|\left(a^2-\eta^2\right)}{2\eta}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Then, by Lemma 3, we have

$$\begin{aligned} \partial_\theta |Y|^2 =& \frac{\eta'}{\eta}\left(|Y_1|^2 - 2\Re\left(\overline{Y_1}Y_2\right) + |Y_2|^2\right) + \frac{|\xi|\,(a+\eta)\,(a-\eta)}{\eta} 2\Re\left(iY_1\overline{Y_2}\right) \\ \lesseqgtr& \left(\frac{2\eta'}{\eta} \pm 2e^{4b_0}|\xi||a-\eta|\right)|Y|^2. \end{aligned}$$

Therefore, noting the following estimates:

$$\begin{cases} |\xi| \int_{\theta(T)}^{\theta(t)} |a(s) - \eta(s)|\, ds \le a_0|\xi|(1+t)^\alpha \le a_0 N & (\alpha \ge 0), \\ |\xi| \int_{\theta(t_\xi)}^{\theta(t)} |a(s) - \eta(s)|\, ds \le a_0|\xi|(1+t_\xi)^\alpha = a_0 N & (\alpha < 0), \end{cases}$$

by Gronwall's lemma, we have

$$|Y(\theta(t),\xi)|^2 \lesseqgtr \left(\frac{\eta(\theta(t))}{\eta(\theta(T))}\right)^2 \exp\left(\pm 2a_0 N e^{4b_0}\right)|Y(\theta(T),\xi)|^2 \simeq |Y(\theta(T),\xi)|^2$$

for $\alpha > 0$ and $T \le t \le t_\xi$, and

$$|Y(\theta(t),\xi)|^2 \lesseqgtr \left(\frac{\eta(\theta(t))}{\eta(\theta(t_\xi))}\right)^2 \exp\left(\pm 2a_0 N e^{4b_0}\right)|Y(\theta(t_\xi),\xi)|^2 \simeq |Y(\theta(t_\xi),\xi)|^2$$

for $\alpha \le 0$ and $t \ge t_\xi$. Noting the equalities $|Y|^2 = 2|\partial_\theta y|^2 + \eta^2|\xi|^2|y|$ and $a\partial_\theta y = \partial_t \hat{w}$, we have $2e^{-4b_0}|W| \le |Y|^2 \le 2e^{4b_0}|W|$, and thus we conclude the proof. □

3.4 Completion of the Proof

If $\alpha > 0$, then by Propositions 1 and 2, we have

$$\begin{cases} K_2^{-1}|W(T,\xi)| \le |W(t,\xi)| \le K_2|W(T,\xi)| & (T \le t \le t_\xi), \\ |W(t,\xi)| \begin{cases} \le K_1|W(t_\xi,\xi)| \le K_1K_2|W(T,\xi)| \\ \ge K_1^{-1}|W(t_\xi,\xi)| \ge K_1^{-1}K_2^{-1}|W(T,\xi)| \end{cases} & (t \ge t_\xi). \end{cases}$$

On the other hand, if $\alpha \le 0$, then we have

$$\begin{cases} K_1^{-1}|W(T,\xi)| \le |W(t,\xi)| \le K_1|W(T,\xi)| & (T \le t \le t_\xi), \\ |W(t,\xi)| \le \begin{cases} \le K_2|W(t_\xi,\xi)| \le K_1K_2|W(T,\xi)| \\ \ge K_2^{-1}|W(t_\xi,\xi)| \ge K_2^{-1}K_2^{-1}|W(T,\xi)| \end{cases} & (t \ge t_\xi). \end{cases}$$

Consequently, since the estimate $E(u;b)(t) = \exp\left(-2\int_t^\infty b(s)\,ds\right)\|W(t,\cdot)\|_{L^2}^2 \simeq \|W(t,\cdot)\|_{L^2}^2$ holds by (18) and Parseval's equality, we have (4).

In order to prove (14), introduce the following proposition:

Proposition 3 *For any $T > 0$, there exists a positive constant $K_0 = K_0(T)$ such that the following estimate is established on $[0,T]$:*

$$E_{KG}(u;1)(t) \le K_0 E_{KG}(u;1)(0).$$

Proof We extend $b(t)$ on $[0,T)$ as $b \in C^0([0,\infty))$ and $|b|$ is monotone decreasing. By Cauchy-Schwarz inequality, we have

$$\frac{d}{dt}E_{KG}(u;1)(t) = (1-M(t))\Re\left(u(t,\cdot),\partial_t u(t,\cdot)\right)_{L^2} \le |1-M(t)|E_{KG}(u;1)(t).$$

Therefore, by (7) and Gronwall's inequality, we have

$$\begin{aligned} E(u;b)(t) &\le \|u(t,\cdot)\|_{L^2}^2 + 2\|\partial_t u(t,\cdot)\|^2 + 2b(T)^2\|u(t,\cdot)\|_{L^2}^2 \\ &\simeq E_{KG}(u;1)(t) \le \exp\left(T \sup_{0\le t\le T}\{|1-M(s)|\}\right) E_{KG}(u;1)(0). \end{aligned}$$

for any $t \in [0,T]$. □

Thus (14) is proved by combining Proposition 3 and (4).

Acknowledgments This work was supported by JSPS KAKENHI Grant Number 18K03372.

References

1. Böhme, C.: Decay rates and scattering states for wave models with time-dependent potential. Ph.D. Thesis, TU Bergakademie Freiberg, Germany, 2011
2. Böhme, C., Reissig, M.: A scale-invariant Klein-Gordon model with time-dependent potential. Ann. Univ. Ferrara Sez. VII Sci. Mat. **58**, 229–250 (2012)
3. Ebert, M.R., Kapp, R.A., Nascimento, W.N., Reissig, M.: Klein-Gordon type wave equation models with non-effective time-dependent potential. In: Dubatovskaya, M.V., Rogosin, S.V. (eds.) AMADE2012, vol. 60, pp. 143–161. Cambridge Scientific Publishers, Cambridge (2014)
4. Ebert, M.R., Fitriana, L., Hirosawa, F.: On the energy estimates of the wave equation with time dependent propagation speed asymptotically monotone functions. J. Math. Anal. Appl. **432**, 654–677 (2015)
5. Hirosawa, F.: On the asymptotic behavior of the energy for the wave equations with time depending coefficients. Math. Ann. **339**, 819–839 (2007)
6. Hirosawa, F.: On the energy estimate for Klein-Gordon-type equations with time-dependent singular mass, in Trends in Mathematics. Analysis, Probability, Applications, and Computation, Proceedings of the 11th ISAAC Congress, Växjö (Sweden) 2017, pp. 325–335. Birkhäuser, Basel (2019). https://doi.org/10.1007/978-3-030-04459-6_31
7. Hirosawa, F., Nascimento, W.N.: Energy estimates for the Cauchy problem of Klein-Gordon-type equations with non-effective and very fast oscillating time-dependent potential. Ann. Math. Pura. Appl. (4) **197**, 817–841 (2018)
8. Hirosawa, F.,Wirth, J.: C^m-theory of damped wave equation with stabilisation. J. Math. Anal. Appl. **343**, 1022–1035 (2008)

Non-Linear Evolution Equations with Non-Local Coefficients and Zero-Neumann Condition: One Dimensional Case

Akisato Kubo and Hiroki Hoshino

Abstract In this paper, we investigate the global existence in time and asymptotic behaviour of solutions of non-linear evolution equations with strong dissipation and non-local coefficients in one spacial dimension, arising in mathematical models of cell migration. We consider the initial boundary value problem with zero-Neumann condition for the equation, applying the argument of the singular integral operator to the non-local term, and we obtain the L^2-estimate of it which is necessary for the energy estimates of our problems. Finally we can prove the desired result by the standard argument of the interation scheme of our problem.

1 Introduction

Let us consider the following non-linear evolution equations with non-local term, for $w := w(x,t)$ with $(x,t) \in \Omega \times (0,T)$

$$\text{(NE)} \begin{cases} w_{tt} = D\Delta w_t + \nabla \cdot (\alpha(w_t)e^{-w}\chi[w]) + \mu(1-w_t)w_t, & \text{in } \Omega \times (0,T) \quad (1.1) \\ \dfrac{\partial}{\partial \nu} w = 0 & \text{on } \partial\Omega \times (0,T) \quad (1.2) \\ w(x,0) = w_0(x),\ w_t(x,0) = w_1(x) & \text{in } \Omega \quad (1.3) \end{cases}$$

where D, μ are positive constants, $\alpha(\cdot)$ is an sufficiently smooth function, Ω is a bounded domain in $\mathbb{R}^n$ with smooth boundary $\partial\Omega$ and ν is the outer unit normal vector on $\partial\Omega$, $\chi[w] := \chi[w](x,t)$ is a non-local term. In this paper we study the one-dimensional case of (NE).

A. Kubo (✉) · H. Hoshino
Fujita Health University, Toyoake, Japan
e-mail: akikubo@fujita-hu.ac.jp; hhoshino@fujita-hu.ac.jp

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_49

Let $\tilde{w}$ be an extension function of w satisfying $w = \tilde{w}$ in Ω, $\tilde{w} = 0$ for $|r| > |\Omega|$. Let us recall that

$$\|\tilde{w}\|_m \leq C\|w\|_{m,\Omega} \tag{1.4}$$

(cf. Mizohata [15]; Chap. 3), where $\|\cdot\|_{m.\Omega}$ is the Sobolev norm of order m defined in Ω, which will be specified later soon, and it is written by $\|\cdot\|_m$ simply when $\Omega = \mathbb{R}$. The definition of $\chi[w]$ for $n = 1$ is given as follows: for a step function with respect to r:

$$\chi_{\pm}(x,t) = \begin{cases} \chi(x,t)\ (r > 0) \\ -\chi(x,t)\ (r < 0), \end{cases}$$

and a smooth function $\chi(x,t)$ in $\Omega \times (0,T)$,

$$\chi[w](x,t) = \text{v.p.}\chi_{\pm}(x,t)\frac{1}{r} * \tilde{w}_{xt}(r,t)$$

$$:= \lim_{\epsilon \to 0} \int_{|r| \geq \epsilon} \chi_{\pm}(x,t)\frac{1}{r}\tilde{w}_{xt}(x-r,t)dr.$$

In this paper, we investigate (NE) for $n = 1$ and our purpose is to establish the existence theorem of time global solutions to (NE). Our difficulty to deal with (NE) lies in the discontinuity of $\chi_{\pm}(x,t)$ at $r = 0$. We apply the L^2-estimate of the singular integral operator to the non-local term to derive the estimates of (NE), which play an important role to obtain our desired results.

Now let us introduce the function spaces used in this paper. Firstly, $H^l(\Omega)$ denotes the Sobolev space $W^{l,2}(\Omega)$ of order l on Ω. For functions $h(x,t)$ and $k(x,t)$ defined in $\Omega \times [0,\infty)$, putting $(h,k)_\Omega(t) = \displaystyle\int_\Omega h(x,t)k(x,t)dx$, $\|h\|_\Omega^2 = (h,h)_\Omega(t)$, then we define the norm of $H^l(\Omega)$ by $\|h\|_{l,\Omega}^2(t) = \displaystyle\sum_{|\beta| \leq l} \|\partial_x^\beta h(\cdot,t)\|_\Omega^2(t)$, and also denote $\|\cdot\|_0$ and $\|\cdot\|_{0,\Omega}$ simply by $\|\cdot\|$ and $\|\cdot\|_\Omega$ respectively, where $\partial_x = (\partial_{x_1}, \cdots, \partial_{x_n}) = \left(\dfrac{\partial}{\partial x_1}, \cdots, \dfrac{\partial}{\partial x_n}\right)$ and β is a multi-index for $\beta = (\beta_1, \cdots, \beta_n)$.

Secondly let define $W^l(\Omega)$ as follows, which is a subspace of $H^l(\Omega)$. The eigenvalues of $-\Delta$ with the homogeneous Neumann boundary conditions are denoted by $\{\lambda_i | i = 0, 1, 2, \cdots\}$, which are arranged as $0 = \lambda_0 < \lambda_1 \leq \cdots \to +\infty$. Let $\varphi_i = \varphi_i(x)$ indicate the L^2 normalized eigenfunction corresponding to λ_i. Then we put for $h(x), k(x) \in H^l(\Omega)$, and a non-negative integer l, $(h,k)_{l,\Omega} = (h,k)_\Omega + (\nabla^l h, \nabla^l k)_\Omega$, $|h|_{l,\Omega}^2 = (h,h)_{l,\Omega}$. We set $W^l(\Omega)$ as a closure of the subset spanned by $\{\varphi_1, \varphi_2, \cdots, \varphi_n, \cdots\}$ in $H^l(\Omega)$. Taking $\lambda_1 > 0$ into account, it is noticed that we have $\int_\Omega h(x)dx = 0$ for $h(x) \in W^l(\Omega)$, which enables us to use Poincare's inequality. We know the equivalence of norms $|\cdot|_{l,\Omega}$, $\|\cdot\|_{l,\Omega}$.

1.1 Known Results and Reduction Process to (NE)

Recently in [6] Gerisch and Chaplain proposed a non-local model of a cell migration (see also [5]): for $n = n(x,t)$, $f = f(x,t)$, $m = m(x,t)$, and positive constants D_1, D_3, γ, α, λ, μ_1 and μ_2,

$$\text{(CG)} \begin{cases} \partial_t n = \nabla \cdot [D_1 \nabla n - n\mathcal{A}\{\underline{u}(t,\cdot)\}] + \mu_1 n(1-n-f), & (1.5) \\ \partial_t f = -\gamma m f + \mu_2(1-n-f), & (1.6) \\ \partial_t m = \nabla \cdot [D_3 \nabla m] + \alpha n - \lambda m, & (1.7) \end{cases}$$

with initial data and zero-Neumann condition, where

$$\mathcal{A}\{\underline{u}(t,\cdot)\}(x) = \frac{1}{R}\int_{-R}^{R} \frac{1}{R}\Omega(r)\sigma(\underline{u}(t,x-r))dr, \qquad \underline{u}(x,t) = (n,f,m)(x,t)$$

is the non-local term for "sensing radius" $R > 0$, which detects the local environment of the cell, a stp funstion $\Omega(r)$:

$$\Omega(r) = \begin{cases} c \text{ for } r > 0 \\ -c \text{ for } r < 0, \end{cases}$$

and a smooth function $\sigma(\,\cdot\,)$. In the same reduction way as used in [7–10, 12, 13], (NE) is reduced from (CG). Following [4–6] it is seen that as $R \to 0$ the non-local model (CG) is reduced to a corresponding local model, for example, a cell migration model (CL) proposed by Chaplain and Lolas [3], for positive constants d, μ, γ_1, γ_2, d_3, α_1, λ_1

$$\text{(CL)} \begin{cases} \partial_t n = d\Delta n - \gamma_1 \nabla \cdot \left(n \nabla \sigma(\underline{u})\right) + \mu n(1-n-f), & (1.8) \\ \partial_t f = -\gamma_2 m f, \\ \partial_t m = \nabla \cdot [d_3 \nabla m] + \alpha_1 n - \lambda_1 m. \end{cases}$$

with initial data and zero-Neumann condition. Mathematical analysis of (CL) is studied in [7, 8, 13] and related chemotaxis models to (CL) are considered in [1, 2, 11, 14, 16, 17].

For a constant $\epsilon > 0$ putting $R = \epsilon$, the non-local term is written by the form

$$\mathcal{A}\{\underline{u}(t,\cdot)\}(x) = \int_{\epsilon > |r|} \Omega(r)\frac{1}{\epsilon^2}\sigma(\underline{u}(t,x-r))dr, \tag{1.9}$$

taking Taylor expansion of $\sigma(\underline{u}(t, x-r))$ at $r=0$,

$$= \frac{1}{\epsilon^2}\sum_{k=0}^{K}\frac{d^k}{dx^k}\sigma(\underline{u}(t,x))A_k(\epsilon) + R_K$$

where $A_k(\epsilon) = \int_{-\epsilon}^{\epsilon}\Omega(r)\frac{(-r)^k}{k!}dr$ and R_K is a remainder term. We see that if k is even,

$$\frac{1}{\epsilon^2}A_k(\epsilon) = 0.$$

Following to [6], as $\epsilon \to 0$ it holds that for a constant A

$$\mathcal{A}\{\underline{u}(t,\cdot)\}(x) \to \frac{d}{dx}\sigma(\underline{u}(t,x))A. \tag{1.10}$$

Due to (1.10) they justify $\mathcal{A}\{\underline{u}(t,\cdot)\}(x)$ as a generalization of the differential operator of $\partial_x\sigma(\underline{u}(t,x))$. Further in order to consider the non-local term in more details we divide it into two parts taking account of (1.9) for any fixed constant $\epsilon > 0$ and it is respected as

$$\mathcal{A}\{\underline{u}(t,\cdot)\}(x) = \Big(\int_{R\geq|r|>\epsilon} + \int_{\epsilon>|r|}\Big)\Omega(r)\frac{1}{\epsilon^2}\sigma(\underline{u}(t,x-r))dr. \tag{1.11}$$

In the first term of (1.11) the integrated function should not depend on ϵ if we apply our mathematical analysis. Hence by changing ϵ to the variable $r \in \mathbb{R}$ in the term: $\frac{1}{\epsilon^2}$ of the first term of (1.11), as $\epsilon \to 0$, the non-local term is rewritten by

$$\mathcal{A}\{\underline{u}(t,\cdot)\}(x) \simeq \lim_{\epsilon\to 0}\int_{R\geq|r|>\epsilon}\frac{1}{r^2}c_{\pm}\sigma(\underline{u}(t,x-r))dr + A\frac{d}{dx}\sigma(\underline{u}). \tag{1.12}$$

Hence it is enough to focus our arguments only on the first term of (1.12) because the second term is just a local term, in which case (CG) can be considered as the same type of problem of (CL), assuming $\sigma(\underline{u}) = f$.

Further let us study a simpler case of the first term of (1.12) for $\sigma(\underline{u}) = u$. From integration by parts it follows that

$$\text{v.p.}c_{\pm}\frac{1}{r^2} * u(r,t) = \lim_{\epsilon\to 0}\int_{|r|\geq\epsilon} c_{\pm}\frac{1}{r^2}u(x-r,t)dr$$

$$= \lim_{\epsilon\to 0}\int_{|r|\geq\epsilon} c_{\pm}\frac{r}{r^2}\frac{d}{dr}u(x-r,t)dr - \lim_{\epsilon\to 0}\Big\{c_{+}\frac{\epsilon}{\epsilon^2}u(x-\epsilon) + c_{-}\frac{\epsilon}{\epsilon^2}u(x+\epsilon)\Big\}$$

$$= -\text{v.p.}c_{\pm}\frac{r}{r^2} * (\frac{d}{dx}u)(r,t) + A_1(\frac{d}{dx}u)(x,t) \tag{1.13}$$

where A_1 is a constant. In fact, by using the argument from (1.9) to (1.10) the boundary terms can be expressed by

$$\sum_{k=0}^{K} - \left(c_+ \frac{r}{r^2} \frac{(-r)^k}{k!} \frac{d^k}{dx^k} u(x,t)\right)|_{r=\epsilon} + \left(c_- \frac{r}{r^2} \frac{(-r)^k}{k!} \frac{d^k}{dx^k} u(x,t)\right)|_{r=-\epsilon}$$

$$+ \text{the remainder term}$$

which tend to $A_1 \frac{d}{dx} u(x,t)$ as $\epsilon \to 0$. Thus the first term of (1.13) leads us to the definition of $\chi[w]$.

In [6] Gerisch and Chaplain investigate and explorate the model by computational simulations. Mathematical analysis of the model is given by Chaplain, Lachowicz, et al. [4] for a more abstract form than (CG).

However in their non-local term such discontinuity at $r = 0$ as in $\Omega(r)$ of (CG) is not considered. In this sense they do not to deal with this critical point of the problem and the regularity theorem of their problem was not obtained.

To overcome these difficulties, we introduce the argument of a singular integral operator to the non-local term and consider (NE) in **$\mathbf{L^2}$-framework**. Finally we obtain the existence theorem and asymptotic behaviour of solutions of (NE) in the analogous way as used in the previous papers [7–10, 12, 13]. In this paper, we consider (NE) for $\mu = 0, \alpha(w_t) = w_t$ in one spacial dimension for the simplicity.

Remark 1 We already have dealt with local cases corresponding to (NE), which is given by replacing $\chi[w]$ with $\chi(x,t)\nabla w$ in the non-local term, and we call (LE) for the replaced one below. The problem (LE) is reduced from (CL) and using results of (LE) the existence theorem of (CL) is shown (cf. [7–10, 12, 13]). In the same line we will be able to deal with (CG) by using results of (NE). The same type of problems of (LE) is studied in [7–14, 16–18].

2 Existence and Asymptotic Profile of (NE)

For $g(w)(x,t) = e^{-w} w_t \chi[w]$, we set

$$P[w] = w_{tt} - D\Delta w_t - \nabla \cdot g(w).$$

We seek the solution in the form of $w(x,t) = a + bt + v(x,t)$ for positive parameters a and b. Then (NE) is rewritten by

$$(\mathrm{NE})_{ab} \begin{cases} P_{a,b}[v] = v_{tt} - D\Delta v_t - \nabla \cdot (g_{a,b}(v)) = 0 \\ \partial_\nu v|_{\partial\Omega} = 0 \\ v(x,0) = v_0(x) = w_0(0) - a, \\ v_t(x,0) = v_1(x) = w_1(x) - b \end{cases}$$

where we denote $g_{a,b}(v) = e^{-a-bt-v}(b+v_t)\chi_{(a,b)}[v]$ with $\chi_{(a,b)}[v] = \chi[a+bt+v]$. We will seek the time global solution of $(\mathrm{NE})_{ab}$. It is noticed that $\chi_{(a,b)}[v]$ is represented as follows.

$$\chi_{(a,b)}[v] = \text{v.p.}\chi_{\pm}(x,t)\frac{1}{r} * v_{xt} = \lim_{\epsilon\to 0}\int_{|r|\geq\epsilon} \chi_{\pm}(x,t)\frac{1}{r}\widetilde{v_{xt}}(x-r,t)dr.$$

Here and hearafter below we often use notations ∇ and Δ instead of ∂_x and ∂_x^2 respectively for the readability.

2.1 Singular Integral Operator

Let us introduce the singular integral operator H as follows for a bounded function $h_j(x)$:

$$Hu(x) = (2\pi i)^2 \sum_{i=1}^{n} h_j(x) R_j u(x),$$

$$R_j u(x) = \text{v.p.}\frac{x_j}{|x|^{n+1}} * u(x), \quad \text{for } x \in \mathbb{R}^n,$$

where R_j is Riesz operator. We have the following well known L^2-estimate of the singular integral operators (see Mizohata [15]; Chap. 6).

Proposition 1 *For $u \in L^2(\mathbb{R}^n)$ there exists a constant $C > 0$ such that it follows that*

$$\|Hu(x)\| \leq C\|u\|.$$

2.2 Estimate of the Non-Local Term

For $u \in C((0,T); L^2(\mathbb{R}))$ and a constant $c > 0$ we have

$$\text{v.p.}c_{\pm}\frac{1}{r} * u(r,t) = \lim_{\epsilon\to 0}\int_{|r|\geq\epsilon} c_{\pm}\frac{r}{r^2}u(x-r,t)dr = \text{v.p.}\frac{r}{r^2} * c_{\pm}u(r,t) \tag{2.1}$$

Taking account of (2.1) and Proposition 1 we have

$$\|\text{v.p.}c_{\pm}\frac{1}{r} * u(r,t)\| \leq C\|\text{v.p.}\frac{r}{r^2} * c_{\pm}u(r,t))\| \leq C\|u\|. \tag{2.2}$$

Lemma 1 *For $u \in C((0,T); L^2(\mathbb{R}))$ there exists a constant $C > 0$ such that (2.2) holds.*

Let us estimate the non-local term $\chi[u]$ for $u_t(x,t) \in C((0,T); H^1(\Omega))$ as follows. By using an extension $\tilde{u}$ of u we have

$$\|\text{v.p.}\chi_{\pm}(r,t)\frac{1}{r} * u_{xt}(r,t)\|_{\Omega} \le \|\lim_{\epsilon\to 0}\int_{|r|\ge\epsilon} \chi_{\pm}(x,t)\frac{r}{r^2}\tilde{u}_{xt}(x-r,t)dr\|$$
$$\le C\|\tilde{u}_t(x)\|_1 \le C\|u_t\|_{1,\Omega}$$

in the same procedure from (2.1) to (2.2) to this term. Then we obtain the following result.

Lemma 2 *For $u_t \in C((0,T); H^1(\Omega))$ there exists a constant $c > 0$ such that we obtain*

$$\|\chi[u](x,t)\|_{\Omega} \le C\|u_t\|_{1,\Omega}.$$

2.3 Estimates of Non-Local Problem $(NE)_{ab}$

We assume the regularity and the boundedness conditions for $m \ge [n/2]+1$

$$v_t \in L^2([0,\infty); W^m(\Omega)) \text{ and } (v_t, e^{-a-bt-v}) \in B_{\mathrm{r}+}, \tag{2.3}$$

where $B_{\mathrm{r}+}$ is an upper semicircle of radius r at 0 in $\mathbb{R}^2$. We first prepare a result required to derive energy estimates of $(NE)_{ab}$. The following lemma is shown by the integration by parts with respect to t (see [7–10, 12, 13]).

Lemma 3 *Assume that $v = v(x,t)$ satisfies the condition (2.3) with $m > M \ge [n/2]+1$. For $0 < b' < b$ and $i = 1, 2, \cdots, n$, it holds that*

$$\|e^{-b't}v_{x_i}\|^2_{M,\Omega}(t) + \int_0^t \|e^{-b's}v_{x_i}\|^2_{M,\Omega}(s)ds$$
$$\le C\left(\int_0^t \|e^{-b's}v_{x_is}\|^2_{M,\Omega}(s)ds + \|v_{x_i}\|^2_{M,\Omega}(0)\right).$$

Proof For any $\varepsilon > 0$, integration by parts for the second term leads us to

$$\|e^{-b't}v_{x_i}\|^2_{\Omega}(t) + 2b'\int_0^t e^{-2b's}\|v_{x_i}\|^2_{\Omega}(s)ds$$
$$\le C\left(\frac{1}{\varepsilon}\int_0^t e^{-2b's}\|v_{x_is}\|^2_{\Omega}(s)ds + \varepsilon\int_0^t e^{-2b's}\|v_{x_i}\|^2_{\Omega}(s)ds + \|v_{x_i}\|^2_{\Omega}(0)\right).$$

Taking ε sufficiently small, finally we have the desired result. □

Lemma 4 (Basic estimate of (NE)$_{ab}$) *We have a basic energy estimate of $(NE)_{ab}$ under the regularity and boundedness conditions (2.3) on $v = v(x,t)$,*

$$\|v_t\|_\Omega^2(t) + \int_0^t D\|\nabla v_s\|_\Omega^2 ds \le CE_a[v](0), \tag{2.4}$$

for sufficiently large a where $E_a[v] = \|v_t\|_\Omega^2 + e^{-2a}\|\nabla v\|_\Omega^2$.

Proof We get by the integration by parts

$$\begin{aligned} 2(P_{a,b}[v], v_t)_\Omega &= 2(\partial_t^2 v - D\Delta v_t - \nabla\cdot(g_{a,b}(v)), v_t)_\Omega \\ &= \frac{\partial}{\partial t}\|v_t\|_\Omega^2 + 2D\|\nabla v_t\|_\Omega^2 + 2(g_{a,b}(v), \nabla v_t)_\Omega = 0. \end{aligned} \tag{2.5}$$

From Lemmas 2 and 3 it follows that for any $\varepsilon > 0$ and a constant $0 < b' < b$,

$$\begin{aligned} \int_0^t (e^{-a-bt-v}(b+v_s)\chi_{(a,b)}[v_s], \nabla v_s)_\Omega ds \le C(\varepsilon^{-1}\int_0^t (e^{-2a-2b's}\nabla v_s, \nabla v_s)_\Omega ds \\ +\varepsilon\int_0^t \|\nabla v_s\|_\Omega^2 ds). \end{aligned}$$

By integrating the equality (2.5) over $(0,t)$ and using the above estimate, we get by the same way as in previous papers [7–10, 12, 13]

$$\begin{aligned} &\|v_t\|_\Omega^2(t) + \int_0^t 2D\|\nabla v_s\|_\Omega^2(s) ds \\ &\le C(E_a[v](0) + \varepsilon^{-1}\int_0^t (e^{-2(a+b's)}\nabla v_s, \nabla v_s)_\Omega ds + \varepsilon\int_0^t \|\nabla v_s\|_\Omega^2 ds). \end{aligned} \tag{2.6}$$

Since the last term of the right hand side of (2.6) is negligible for sufficiently small ε, we have

$$\|v_t\|_\Omega^2(t) + \int_0^t D\|\nabla v_s\|_\Omega^2(s) ds \le CE_a[v](0) + C_\varepsilon e^{-a}\int_0^t \|\nabla v_s\|_\Omega^2 ds. \tag{2.7}$$

□

Lemma 5 (Higher Order Estimates for (NE)$_{ab}$) *Under the regularity and boundedness conditions (2.3) on $v = v(x,t)$ with $m > M \ge [n/2]+1$, we have the higher order energy estimate of $(NE)_{ab}$ for sufficiently large a:*

$$\sum_{j=1}^{M+1}\{\|\nabla^{j-1}v_t\|_\Omega^2(t) + \int_0^t D\|\nabla^j v_s\|_\Omega^2(s) ds\} \le CE_{a,M}[v](0), \tag{2.8}$$

where we denote for any non-negative integer k, $E_{a,k}[v](t) = E_a[\nabla^k v]$.

Proof Suppose that the estimate (2.8) holds for $M = k - 1 \geq 0$. Considering $\nabla^k v$ instead of v in (2.4), in the same way as in Lemma 4 we can obtain (2.8) for $M = k$. In fact, in order to show it, it is enough to prove that the following estimate holds for a parameter $\kappa > 0$

$$\left(\nabla^k\left(e^{-a-bt-v}(b+v_t)\chi_{n(a,b)}[v]\right) - e^{-a-bt-v}(b+v_t)\chi_{n(a,b)}[\nabla^k v], \nabla^k v_t\right)_\Omega$$

$$\leq C(\kappa^{-1}\sum_{j=1}^{k}(e^{-a-b't}\nabla^j v_t, \nabla^j v_t)_\Omega + \kappa|v_t|^2_{k,\Omega} + E_{a,k-1}[v](0)). \tag{2.9}$$

By using Lemma 4 and taking a and κ sufficiently large and small respectively, the first and second terms of (2.9) can be neglected. Hence we obtain (2.8). □

3 Existence of the Solution to $(\mathrm{NE})_{ab}$

We obtain the following result of the global existence in time and some properties of the solution to $(\mathrm{NE})_{ab}$.

Theorem 1 *Let* $(v_0(x), v_1(x)) \in W^{m+1}(\Omega) \times W^m(\Omega)$ *for* $v_0(x) = w_0(x) - a$ *and* $v_1(x) = w_1(x) - b$ *and* $m \geq [n/2] + 3$. *For sufficiently large a, there exists the solution:*

$$w(x,t) = a + bt + v(x,t) \in \bigcap_{i=0}^{1} C^i([0,\infty); H^{m-i}(\Omega))$$

to $(NE)_{ab}$, *moreover for* $\overline{w}_1 = |\Omega|^{-1}\int_\Omega w_1(x)dx$ *it holds that*

$$\lim_{t\to\infty}\|w_t(x,t) - \overline{w}_1\|_{m-1,\Omega} = 0. \tag{3.1}$$

Proof The proof will be shown in the same manner as in [7–10, 12, 13]. We give an iteration scheme and derive the energy estimate of it.

$$(NE)_{(i+1)}\begin{cases} P_i[v_{i+1}] = \partial_t^2 v_{i+1} - \partial_t\Delta v_{i+1} \\ \qquad -\nabla\cdot(e^{-a-bt}(b+v_{it})\chi_{(a,b)}[v_i]e^{-v_i}) = 0 \\ \partial_\nu v_{i+1}|_{\partial\Omega} = 0, \\ v_{i+1}(x,0) = v_0(x), \quad v_{i+1t}(x,0) = v_1(x) \end{cases}$$

where $v_i = \sum_{j=1}^{\infty} f_{ij}(t)\varphi_j(x)$, $v_0(x) = \sum_{j=1}^{\infty} h_j\varphi_j(x)$, $v_1(x) = \sum_{j=1}^{\infty} h'_j\varphi_j(x)$. Taking $E_{a,M}[v](0)$ sufficiently small in the energy estimate, we see that (2.8) guarantees the estimate with a uniform upper bound of each problem $(NE)_{(i+1)}$ for large enough r in $B_{\mathrm{r}+}$ and $i = 1, 2, \cdots$.

We determine $f_{ij}(t)$ by Galerkin method and by applying (2.8) to the following system of ordinary deferential equations with initial data, for $j = 1, 2, \cdots$, we obtain the global smooth solution in time, which satisfies (2.3),

$$\begin{cases} (P_i[v_{i+1}], \varphi_j) = 0, \\ f_{i+1j}(0) = h_{i+1}, \quad f_{i+1jt}(0) = h'_{i+1}. \end{cases}$$

Eventually the energy estimate enables us to get the solution of $(\mathrm{NE})_{ab}$ by considering $P_i[v_{i+1}] - P_{i-1}[v_i]$ and the standard argument of convergence for $v_{i+1} - v_i = u_i$. In fact, we consider the following problem for $l(t) = a + bt$.

$$(NE)_{(i+1)-(i)} \begin{cases} P_i[v_{i+1}] - P_{i-1}[v_i] = \partial_t^2 u_{i+1} - \partial_t \Delta u_{i+1} \\ \qquad -\nabla \cdot (e^{-l(t)}(b + v_{it})\chi_{(a,b)}[u_i]e^{-v_i}) \\ \qquad -\nabla \cdot \left(e^{-l(t)}((b + v_{it})\chi_{(a,b)}[v_{i-1}]e^{-v_i} \right. \\ \qquad \left. -(b + v_{i-1t})\chi_{(a,b)}[v_{i-1}]e^{-v_{i-1}})\right) = 0, \\ u_{i+1}(0, x) = u_{i+1t}(0, x) = 0. \end{cases}$$

In order to obtain the estimate of $(NE)_{(i+1)-(i)}$ it is enough to deal with the last term of $P_i - P_{i-1}$ as follows: for $\theta > 0$

$$2\int_0^t \Big(\nabla^{M-1}\nabla \cdot \Big(e^{-l(s)}((b$$

$$+v_{it})e^{-v_i} - (b + v_{i-1t})e^{-v_{i-1}})\chi[v_{i-1}]\Big), \nabla^{M-1}u_{is}\Big)_\Omega ds$$

$$\leq \frac{C_a}{\theta}\int_0^t e^{-2bs}\sum_{j=0}^{1}\|\partial_s^j u_{i-1}\|^2_{M-1,\Omega}ds + \theta\int_0^t \|\nabla u_{is}\|^2_{M-1,\Omega}ds. \tag{3.2}$$

Then in the same way as derived the energy estimate (2.8), we have by using (3.2) for sufficiently large a and small θ,

$$\|u_{it}\|^2_{M-1,\Omega}(t) + D\int_0^t \|\nabla u_{is}\|^2_{M-1,\Omega}ds \leq C_a\|u_{i-1t}\|^2_{M-1,\Omega}, \tag{3.3}$$

where C_a depends on $\sup_t \|\partial_t v_i\|^2_{M,\Omega}$, $\sup_t \|\partial_t v_{i-1}\|^2_{M,\Omega}$ and e^{-a}, $C_a \to 0$ as $a \to \infty$ and for sufficiently small θ we can neglect the last term of (3.2). Then we take a so large that $C_a < 1$. By the standard argument of the iteration scheme there exists the solution $v(x, t)$ of $(\mathrm{NE})_{ab}$ such that $\{v_i\}$ converges strongly to v satisfying

$$\lim_{i\to\infty} v_i = v \text{ in } \bigcap_{i=0}^{1} C^i([0, \infty); H^{m-i}(\Omega)), \quad m \geq [n/2] + 3. \tag{3.4}$$

The proof of (3.1) is shown in the same way as in [9]. □

Concluding Remark We can consider (NE) for $\mu \neq 0, \alpha(w_t) \not\equiv w_t, n \geq 1$ if an appropriate generalization $\chi_n[w]$ of $\chi[w]$ would be given for $n \geq 1$, for example, $\chi_n[w] = (2\pi i)^2 \sum_{i=1}^n \text{v.p.}\chi_{i\pm}(x,t) \frac{\xi_i}{|\xi|^2} * w_{x_i t}(\xi, t)$ for $\xi \in \Omega \subset \mathbb{R}^n$. Also we will be able to show the existence and asymptotic behaviour of the solution to (CG) by using Theorem 1 for $n = 1$ and its generalization for $n \geq 2$. Such results with the full proof of them will be published somewhere soon.

Acknowledgments This work was supported in part by the Grants-in-Aid for Scientific Research (C) 16540176, 19540200, 22540208, 25400148 and 16K05214 from Japan Society for the Promotion of Science.

References

1. Anderson, A.R.A., Chaplain, M.A.J.: Continuous and discrete mathematical models of tumour-induced angiogenesis. Bull. Math. Biol. **60**(5), 857–899 (1998)
2. Anderson, A.R.A., Chaplain, M.A.J.: Mathematical modelling of tissue invasion. In: Preziosi, L. (ed.) Cancer Modelling and Simulation, pp.269–297. Chapman Hall/CRC, Boca Raton (2003)
3. Chaplain, M.A.J., Lolas, G.: Mathematical modeling of cancer invasion of tissue: dynamic heterogeneity. Netw. Heterogeneous Media **1**(3), 399–439 (2006)
4. Chaplain, M.A.J., Lachowicz, M., Szymanska, Z., Wrzosek, D.: Mathematical modeling of cancer invasion: the importance of cell-cell adhesion and cell-matrix adhesion. Math. Mod. Meth. Appl. Sci. **21**(4), 719–743 (2011)
5. Domschke, P., Trucu, D., Gerisch, A., Chaplain, M.A.J.: Mathematical modelling of cancer invasion: implications of cell adhesion variability for tumour infiltrative growth patterns. J. Theor. Biol. **361**, 41–60 (2014)
6. Gerisch, A., Chaplain, M.A.J.: Mathematical modeling of cancer cell invasion of tissue: local and non-local models and the effect of adhesion. J. Theor. Biol. **250**, 684–704 (2008)
7. Kubo, A., Hoshino, H.: Nonlinear evolution equation with strong dissipation and proliferation. Current Trends in Analysis and its Applications, pp. 233–241. Birkhauser/Springer (2015)
8. Kubo, A., Hoshino, H.: Nonlinear evolution equations and their application to chemotaxis models. Analysis, Probability, Applications, and Computation, pp. 333–347. Birkhauser/Springer (2017)
9. Kubo, A., Suzuki, T.: Asymptotic behavior of the solution to a parabolic ODE system modeling tumour growth. Differ. Integr. Equ. **17**(7–8), 721–736 (2004)
10. Kubo, A., Suzuki, T.: Mathematical models of tumour angiogenesis. J. Comput. Appl. Math. **204**(1), 48–55 (2007)
11. Kubo, A., Tello, J.I.: Mathematical analysis of a model of chemotaxis with competition terms. Differ. Intergr. Equ. **29**(5–6), 233–241 (2016)
12. Kubo, A., Suzuki, T., Hoshino, H.: Asymptotic behavior of the solution to a parabolic ODE system. Math. Sci. Appl. **22**, 121–135 (2005)
13. Kubo, A., Hoshino, H., Kimura, K.: Global existence and asymptotic behaviour of solutions for nonlinear evolution equations related to tumour invasion, Dynamical Systems. Differential Equations and Applications, AIMS Proceedings, pp. 733–744 (2015)
14. Levine, H.A., Sleeman, B.D.: A system of reaction and diffusion equations arising in the theory of reinforced random walks. SIAM J. Appl. math. **57**(3), 683–730 (1997)

15. Mizohata, S.: The Theory of Partial Differential Equations. Cambridge University Press, London (1973)
16. Othmer, H.G., Stevens, A.: Aggregation, blowup, and collapse: the ABCs of taxis in reinforced random walks. SIAM J. Appl. Math. **57**(4), 1044–1081 (1997)
17. Sleeman, B.D., Levine, H.A.: Partial differential equations of chemotaxis and angiogenesis. Math. Mech. Appl. Sci. **24**(6), 405–426 (2001)
18. Yang, Y., Chen, H., Liu, W.: On existence and non-existence of global solutions to a system of reaction-diffusion equations modeling chemotaxis. SIAM J. Math. Anal. **33**(4), 763–785 (1997)

Nonlinear Perturbed BLMP Equation

Sandra Lucente

Abstract In the present paper we consider the non-existence of weak solutions related to a class of differential equations connected with the Airy operator. More precisely, we deal with a positive semilinear perturbation of important PDEs involved in fluido-dynamic: Airy, KdV and BLMP equation.

1 Introduction

Many quasilinear PDEs describe physical phenomena and take the form

$$P(t, x, \partial_t, D_x)u = f(D_x u, \partial_t u)$$

with $t > 0$ as time variable and $x \in \mathbb{R}^N$ as space variable. The operator P is a linear operator while f gives the nonlinear part of the equation. In the present paper we want to add forcing terms of type $|u|^q$ and $|Du|^{2q}$ in BLMP (Boiti Leon Manna Pempinelli) equation and show that for some $q > 1$ and suitable initial data the corresponding weak solutions do not exist.

In Sect. 2 we motivate the choice of BLMP equation that can be considered a generalization of KdV equation. In both cases the linear operator is of Airy type. More precisely, in Sect. 2.1 we discuss two Liouville semilinear problems related to Airy operator. In Sect. 2.2 we consider the correspondent initial value problems. In Sect. 2.3 we consider a perturbed KdV equation.

In Sect. 3 we state and prove the main result on the non-existence of weak solutions for BLMP perturbed equation. From such proof one derives, as corollary, all the proofs of theorems stated in Sect. 2.

S. Lucente (✉)
Department of Physics, University of Bari, Bari, Italy
e-mail: sandra.lucente@uniba.it

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_50

2 From Airy to BLMP Equations

Let be $N = 1$. Let us start with the *Airy* operator $\partial_t + \partial_x^3$ and its most famous nonlinear related equation:

$$\partial_t u + \partial_x^3 u = g(u)\partial_x u\,. \tag{1}$$

We can write the right side as $\partial_x(G(u))$ being G a primitive of g. For $G(s) = 3s^2$ this is the classical *Korteweg de Vries* equation introduced in 1895. It is completely integrable and describes for example the evolution of long, one-dimensional waves.

Let $N = 2$. Many variants of the previous couple operator/equation have been studied. For example one can consider

$$\partial_x(\partial_t + \partial_x^3)u + \lambda\partial_y^2 u = \partial_x(g(u)\partial_x u) \tag{2}$$

obtained formally as a linear second order perturbation of (1), after deriving it with respect to x. In particular any solution of (1) gives a solution of this equation independent of y. For $g(u) = u$ this is a *Kadomtsev–Petviashvili* type equation, studied since 1970, see [6]. It gives a model of waves in ferromagnetic media.

In 1986 Boiti, Leon, Manna, Pempinelli [1] proposed a 2D-variant for KdV:

$$\partial_y(\partial_t + \partial_x^3)u = 3\partial_x u\partial_y\partial_x u + 3\partial_x^2 u\partial_y u\,. \tag{3}$$

The linear operator is a derivation of Airy's one. Concerning the nonlinear term, it can be seen as nonlinear perturbation of $\partial_y(G(\partial_x u))$, with $G(s) = 3/2s^2$, by an extra-term which involves $\partial_x^2 u$. Formally taking $x = y$ and $v = \partial_x u$, from (3) we come back to (1). We mention that in this case the nonlinear part does not depend on the function u, but only on its derivatives. It describes the interaction of two different waves along the two axes. The 3D-version was given in 2012 in [4]:

$$(\partial_y + \partial_z)(\partial_t + \partial_x^3)u = 3\partial_x u(\partial_y + \partial_z)\partial_x u + 3\partial_x^2 u(\partial_y + \partial_z)u\,. \tag{4}$$

Such kind of equations appear in oceanography and plasma physics. The 4D case has been recently studied in [8]. In order to introduce the ND version, we take $x \in \mathbb{R}$ as special direction, in which Airy operator applied, and $\xi \in \mathbb{R}^{N-1}$. We denote by

$$S(\nabla_\xi) = \partial_{\xi_1} + \cdots + \partial_{\xi_{N-1}}$$

the divergence of the $N-1$ vector with any component equal to an assigned scalar function. Hence we consider

$$(\partial_t + \partial_x^3)S(\nabla_\xi)u(t,x,\xi) = 3(\partial_x u S(\nabla_\xi)\partial_x u + \partial_x^2 u S(\nabla_\xi)u)\,. \tag{5}$$

In the present paper we want to study nonexistence of weak solution with extra sources which growth polynomially in u and $\nabla_{x,\xi} u$.

2.1 Liouville Problems for Semilinear Airy Equations

The idea to split space-variables into two sets, appears in many papers. Concerning the non-existence of weak solutions, we have to mention for example [2]. Let us rewrite such result for $(t, x, \xi) \in \mathbb{R}^{N+1} = \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{N-1}$. The dual variables, in the sense of Fourier transform, are denoted by $(\tau, \tilde{x}, \eta) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{N-1}$.

Let L be a linear differential operator of order $m \geq 1$ of the form

$$L(t, x, \xi, D_t, D_x, D_\xi) = \sum_{1 \leq |(\alpha,\beta,\gamma)| \leq m} l_{\alpha,\beta,\gamma}(t, x, \xi) D_t^\alpha D_x^\beta D_\xi^\gamma , \tag{6}$$

with multi-index $(\alpha, \beta, \gamma) \in \mathbb{N} \times \mathbb{N} \times \mathbb{N}^{N-1}$ and symbol

$$L(t, x, \xi, \tau, \tilde{x}, \eta) = \sum_{1 \leq |(\alpha,\beta,\gamma)| \leq m} (-1)^{|\alpha|+|\beta|+|\gamma|} l_{\alpha,\beta}(t, x, \xi) \tau^\alpha \tilde{x}^\beta \eta^\gamma .$$

We emphasize that no 0-order term is present in the operator L.

Definition 1 The m-th order operator L is called *quasi-homogeneous* if there exist $\delta_1, \delta_2, \delta_3 > 0$ such that for any $\lambda > 0$, (t, x, ξ), $(\tau, \tilde{x}, \eta) \in \mathbb{R}^{N+1}$, it holds

$$L(\lambda^{-\delta_1} t, \lambda^{-\delta_2} x, \lambda^{-\delta_3} \xi, \lambda^{\delta_1} \tau, \lambda^{\delta_2} \tilde{x}, \lambda^{\delta_3} \eta) = \lambda^m L(t, x, \xi, \tau, \tilde{x}, \eta).$$

By *quasi-homogeneous dimension* we mean the quantity

$$Q = \delta_1 + \delta_2 + \delta_3 (N - 1) .$$

We recall that the adjoint of a linear operator L, denoted by L^*, satisfies

$$\int_{\mathbb{R}^{N+1}} (Lf) g \, dt \, dx \, d\xi = \int_{\mathbb{R}^{N+1}} f L^* g \, dt \, dx \, d\xi, \tag{7}$$

for any $f \in D(L)$, $g \in D(L^*)$. Clearly, $D(L)$ and $D(L^*)$ depend on the regularity of the coefficients $l_{\alpha,\beta,\gamma}$. For simplicity here we consider smooth coefficients.

Definition 2 Let $q > 1$. A function $u \in L^q_{loc}(\mathbb{R}^{N+1})$ is called *weak solution* of $Lu = |u|^q$ if, for any $\varphi \in C_c^\infty(\mathbb{R}^{N+1}, \mathbb{R}_+)$, it holds

$$\int_{\mathbb{R}^{N+1}} |u|^q \varphi \, dt \, dx \, d\xi = \int_{\mathbb{R}^{N+1}} u L^* \varphi \, dt \, dx \, d\xi . \tag{8}$$

Here and in the sequel $C_c^\infty(D)$ denotes the space of functions belonging to $C^\infty(D)$ with compact support in a domain D.

The main result of [2], based on *test function method*, can be written as follows.

Theorem 1 *Let $q > 1$. Consider a m-th order quasi-homogeneous operator L in form (6). Assume*

$$D_t^\alpha D_x^\beta D_\xi^\gamma l_{\alpha,\beta,\gamma}(t,x,\xi) = 0\,, \tag{9}$$

for any $(\alpha,\beta,\gamma) \in \mathbb{N}\times\mathbb{N}\times\mathbb{N}^{N-1}$ such that $1 \le |\alpha|+|\beta|+|\gamma| \le m$. Let Q be the quasi-homogeneous dimension of L. If

$$(Q-m)q \le Q \tag{10}$$

then $Lu = |u|^q$ has no nontrivial weak solutions.

In the case of Airy operator, we have $m = 3$, symbol $\tau+\tilde{x}^3$ and quasi-homogeneous dimension $Q = 4$. Indeed we can take $\delta_1 = 3, \delta_2 = 1$ so that $(\lambda^3\tau) + (\lambda\tilde{x})^3 = \lambda^3(\tau+\tilde{x}^3)$. Hence the equation

$$(\partial_t + \partial_x^3)u(t,x) = |u|^q$$

has no nontrivial weak solution (in the sense of Definition 2), for any

$$1 < q \le 4\,.$$

Now we consider the semilinear case associated to the left side of BLMP equation:

$$(\partial_t + \partial_x^3)S(\nabla_\xi)u(t,x,\xi) = |u|^q \quad q > 1, \tag{11}$$

where $\xi \in \mathbb{R}^{N-1}$ with $N \ge 2$. We see that the operator has order $m = 4$ and it is still quasi-homogeneous, indeed $((\lambda^3\tau) + (\lambda\tilde{x})^3)(\lambda\eta) = \lambda^4(\tau+\tilde{x}^3)\eta$. More precisely $(\delta_1,\delta_2,\delta_3) = (3,1,1)$ and $Q = 3+1+1\times(N-1) = 3+N$. As a conclusion, Eq. (11) has no nontrivial weak solution (in the sense of Definition 2), provided $(3+N-4)q \le 3+N$. Explicitly, we require

$$1 < q \le \frac{N+3}{N-1}\,, \quad N \ge 2.$$

Taking $N = 1$ we do not arrive to the condition $1 < q \le 4$, this reveals that the presence of $S(\nabla_\xi)$ deeply changes the critical exponents, acting both on Q and m. This is a first hint to see that BLMP equation will give different results with respect to Airy's one. Passing to initial value problems, some other differences with respect to KdV will also appear.

2.2 Initial Value Problems for Semilinear Airy Equations

In [3] many generalizations of the test function method appear. For example one can treat initial value problems. Let us adapt such results to our situation.

Definition 3 Let $i = 1, 2$. Let P_i be a linear differential operator of order $k_i \geq 1$:

$$P_i(x, \xi, D_x, D_\xi) = \sum_{1 \leq |(\beta,\gamma)| \leq k_i} p^{(i)}_{\beta,\gamma}(x, \xi) D_x^\beta D_\xi^\gamma , \tag{12}$$

where the multi-index $(\beta, \gamma) \in \mathbb{N} \times \mathbb{N}^{N-1}$. Consider the Cauchy Problem

$$\begin{cases} (\partial_t P_1 + P_2)u(t, x, \xi) = |u(t, x, \xi)|^q , & t \geq 0, x \in \mathbb{R}, \xi \in \mathbb{R}^{N-1}, \ q > 1 \\ u(0, x, \xi) = u_0(x, \xi) . \end{cases} \tag{13}$$

Denoted by $L = \partial_t P_1 + P_2$, we say that $u \in L^q_{\text{loc}}([0, +\infty) \times \mathbb{R}^N)$ is a weak solution to (13) if for any $\eta \in C_c^\infty([0, \infty), \mathbb{R}_+)$, $\phi \in C_c^\infty(\mathbb{R}, \mathbb{R}_+)$ and $\psi \in C_c^\infty(\mathbb{R}^{N-1}, \mathbb{R}_+)$ one has

$$\begin{aligned} &\int_0^\infty \int_{\mathbb{R}^N} |u(t, x, \xi)|^q \, \eta(t)\phi(x)\psi(\xi) \, dx \, dt \, d\xi = \\ &- \int_{\mathbb{R}^N} u_0(x)\eta(0) P_1^*(\phi(x)\psi(\xi)) \, d\xi \\ &+ \int_0^\infty \int_{\mathbb{R}^N} u(t, x, \xi) L^*(\eta(t)\phi(x)\psi(\xi)) \, dx \, dt \, d\xi . \end{aligned}$$

We will discuss two different cases according to the following assumption is satisfied:

$$D_x^\beta D_\xi^\gamma p^{(1)}_{\beta,\gamma}(x, \xi) = 0 , \quad 1 \leq |\beta| + |\gamma| \leq k_1, \tag{14}$$

with $(\beta, \gamma) \in \mathbb{N} \times \mathbb{N}^{N-1}$.

Theorem 2 *Assume that $L = \partial_t P_1 + P_2$ is a quasi-homogeneous operator of order m and quasi-homogeneous dimension Q. Suppose (9) holds. If*

$$(Q - m)q \leq Q$$

then (13) has no global weak solution provided either $u_0 \in L^1(\mathbb{R}^N)$ and (14) holds or

$$u_0 \in D(P_1)\,, \quad P_1 u_0 \in L^1(\mathbb{R}^N) \quad \text{and} \int_{\mathbb{R}^N} P_1 u_0(x,\xi)\,dx\,d\xi > 0\,.$$

We are stating that if $u(t,x,\xi)$ solves (13), according Definition 3, then there exists $T_* > 0$ such that $[0,T_*)\times\mathbb{R}^N$ is the maximal domain for u. The proof of Theorem 2 will follow from the proof of Theorem 4.

Let us apply such theorem to conclude that Airy equation

$$\begin{cases} (\partial_t + \partial_x^3)u(t,x) = |u(t,x)|^q\,, & t \geq 0, x \in \mathbb{R}, \\ u(0,x) = u_0(x)\,. \end{cases}$$

has no global weak solution (in sense of Definition 3) once $1 < q \leq 4$ provided that $\int_{\mathbb{R}} u_0(x)\,dx > 0$.

Similarly semilinear equation associated to BLMP linear part

$$\begin{cases} (\partial_t + \partial_x^3)S(\nabla_\xi)u(t,x,\xi) = |u(t,x,\xi)|^q\,, & t \geq 0, x \in \mathbb{R}, \xi \in \mathbb{R}^{N-1} \\ u(0,x,\xi) = u_0(x,\xi)\,. \end{cases}$$

has no global weak solutions once $1 < q \leq \frac{N+3}{N-1}$ provided that $u_0 \in L^1(\mathbb{R}^N)$. Another difference between Airy and BLMP operator appears: the initial condition does not require positivity. To our knowledge this result is new.

2.3 *Perturbation of KdV Equation*

Starting from [7], the test function method technique has been applied for quasilinear equations. Here we consider a perturbation of KdV equation:

$$\begin{cases} (\partial_t + \partial_x^3)u = \partial_x(|u|^2) + |u|^q + \beta|u|^{2q}\,, & u = u(t,x)\,, \quad t \geq 0, x \in \mathbb{R}\,, \\ u(0,x) = u_0(x)\,. \end{cases} \tag{15}$$

We say that $u \in L^{2q}_{\text{loc}}([0,+\infty)\times\mathbb{R}^N)$ is a weak solution to (15) if for any $\eta \in C_c^\infty([0,\infty),\mathbb{R}_+)$ and for any $\phi \in C_c^\infty(\mathbb{R}^N,\mathbb{R}_+)$ it holds

$$\int_0^\infty \int_{\mathbb{R}^N} (|u|^q + \beta|u|^{2q})\,\eta(t)\phi(x)\,dx\,dt = -\int_{\mathbb{R}^N} u_0(x)\phi(x)\,dx$$
$$-\int_0^\infty \int_{\mathbb{R}^N} u\,(\partial_t + \partial_x^3)(\eta(t)\phi(x))\,dx\,dt - \int_0^\infty \int_{\mathbb{R}^N} \partial_x(|u|^2)\,\eta(t)\phi(x)\,dx\,dt\,.$$

Let us state the following result. For the proof see Remark 1.

Theorem 3 *Let $u_0 \in L^1(\mathbb{R})$ with $\int u_0\,dx > 0$. If $\beta \neq 0$ and $1 < q \leq 4/3$, then (15) has no global weak solution.*

3 Non Existence of Weak Solutions for Perturbed BLMP

We prove our main result for a perturbation of BLMP equation.

Theorem 4 *Let $N \geq 2$. Consider the initial value problem*

$$\begin{cases} (\partial_t + \partial_x^3) S(\nabla_\xi) u = 3(\partial_x u S(\nabla_\xi)\partial_x u + \partial_x^2 u S(\nabla_\xi) u) + |u|^q + |\nabla_{x,\xi} u|^{2q}\,, \\ u(0,x,\xi) = u_0(x,\xi)\,. \end{cases} \tag{16}$$

If $u_0 \in L^1(\mathbb{R}^N)$ and $1 < q \leq \frac{N+3}{N+2}$. then (16) has no global weak solution.

More precisely if $u \in L^q_{loc}([0,T)\times\mathbb{R}^N)$ and $\nabla_{x,\xi} u \in L^{2q}_{loc}([0,T)\times\mathbb{R}^N)$ solves

$$\begin{aligned} &\int_0^\infty \int_{\mathbb{R}^N} (|u|^q + |\nabla_{x,\xi} u|^{2q})\,\eta(t)\phi(x)\psi(\xi)\,dx\,d\xi\,dt = \\ &= -\int_{\mathbb{R}^N} u_0(x,\xi)\eta(0)\phi(x) S(\nabla_\xi)\psi(\xi)\,dx\,d\xi \\ &+ \int_0^\infty \int_{\mathbb{R}^N} u\,(\partial_t + \partial_x^3) S(\nabla_\xi)(\eta(t)\phi(x)\psi(\xi))\,dx\,d\xi\,dt \\ &- 3\int_0^\infty \int_{\mathbb{R}^N} (\partial_x u S(\nabla_\xi)\partial_x u + \partial_x^2 u S(\nabla_\xi) u)\,\eta(t)\phi(x)\psi(\xi)\,dx\,d\xi\,dt \end{aligned}$$

for any $\eta \in C_c^\infty([0,\infty),\mathbb{R}_+)$, $\phi \in C_c^\infty(\mathbb{R},\mathbb{R}_+)$ and $\psi \in C_c^\infty(\mathbb{R}^{N-1},\mathbb{R}_+)$, then $u(x,\xi,t)$ has a finite maximal time existence: $T < +\infty$.

Proof We will deal with a more general situation:

$$L = \partial_t P_1(x,\xi,\partial_x, D_\xi) + P_2(x,\xi,\partial_x, D_\xi)$$

a quasi-homogeneous operator of order m and quasi-homogeneous dimension $Q = \delta_1 + \delta_2 + \delta_3(N-1)$. Moreover we put

$$C_{l,n,r} := \left\{ (t,x,\xi) \in \mathbb{R}_+ \times \mathbb{R}^N : 0 \leq t < l\,, \quad |x| \leq n\,, \quad |\xi| \leq r \right\}.$$

$$C_{n,r} := \left\{ (x,\xi) \in \mathbb{R}^N :\ |x| \leq n\,, \quad |\xi| \leq r \right\}.$$

From Appendix A of [2], we recall that

$$L^* S^I_{\lambda^{\delta_1}} S^{II}_{\lambda^{\delta_2}} S^{III}_{\lambda^{\delta_3}} g = \lambda^m S^I_{\lambda^{\delta_1}} S^{II}_{\lambda^{\delta_2}} S^{III}_{\lambda^{\delta_3}} L^* g \quad \text{for } g \in D(L^*).$$

where $S^I_\lambda g(t,x,\xi) := g(\lambda t, x, \xi)$, $S^{II}_\lambda g(t,x,\xi) := g(t, \lambda x, \xi)$ and $S^{III}_\lambda g(t,x,\xi) := g(t,x,\lambda\xi)$. In what follows $\eta \in C_c^\infty([0,\infty), \mathbb{R}_+)$, $\phi \in C_c^\infty(\mathbb{R}, \mathbb{R}_+)$ and $\psi \in C_c^\infty(\mathbb{R}^{N-1}, \mathbb{R}_+)$. We choose $\eta = 1$ in $[0, -1/2]$, $\eta = 0$ for $t \geq 1$; $\phi = 1$ in $[-1/2, 1/2]$ with $\phi(\mathbb{R}) \subset [0,1]$ and $\operatorname{supp}\phi \subset [-1,1]$; $\psi = 1$ for $|\xi| \in [-1/2, 1/2]$ with $0 \leq \psi(\xi) \leq 1$ for any $\xi \in \mathbb{R}^{N-1}$ and $\operatorname{supp}\psi \subset \{|\xi| \leq 1\}$. Hence we put

$$\eta_R(t) = \eta\left(R^{-\delta_1} t\right), \quad \phi_R(x) = \phi\left(R^{-\delta_2} x\right), \quad \psi_R(\xi) = \psi\left(R^{-\delta_3}\xi\right).$$

We start from

$$\begin{aligned} I_R &:= \int_0^\infty \int_{\mathbb{R}^N} (|u|^q + \beta |\nabla_{x,\xi} u|^{2q})\, \eta_R(t)\phi_R(x)\psi_R(\xi)\, dx\, d\xi\, dt = \\ &= -\int_{\mathbb{R}^N} u_0(x) P_1^*(\phi_R(x)\psi_R(\xi))\, dx\, d\xi \\ &+ \int_0^\infty \int_{\mathbb{R}^N} u L^*(\eta_R(t)\phi_R(x)\psi_R(\xi))\, dx\, d\xi\, dt \\ &+ \alpha \int_0^\infty \int_{\mathbb{R}^N} (\partial_x u S(\nabla_\xi)\partial_x u + \partial_x^2 u S(\nabla_\xi) u)\, \eta_R(t)\phi_R(x)\psi_R(\xi)\, dx\, d\xi\, dt \\ &=: -D_R + L_R + N_R \end{aligned}$$

with $\alpha \in \mathbb{R}$ and $\beta > 0$ or $\alpha = 0 = \beta$. We see that

$$D_R = \int_{\operatorname{supp}(\phi_R \psi_R)} P_1(u_0(x))\phi_R(x)\psi_R(\xi)\, dx\, d\xi$$

If $\int_{\mathbb{R}^N} P_1(u_0(x))\, dx\, d\xi > 0$, then there exists $\bar{R} > 0$ such that $D_R > 0$ for any $R \geq \bar{R}$. If (14) holds, then we observe that P_1 is quasi-homogeneous of dimension $\delta_2 + \delta_3(N-1)$ and order $m - \delta_1$, so that

$$\begin{aligned} D_R &= R^{-m+\delta_1} \int_{C_{R^{\delta_2}, R^{\delta_3}} \setminus C_{R^{\delta_2}/2, R^{\delta_3}/2}} u_0(x.\xi) S^{II}_{R^{-\delta_2}} S^{III}_{R^{-\delta_3}} P_1^*(\phi(x)\psi(\xi))\, dx\, d\xi \\ &= R^{-m+\delta_1+\delta_2+\delta_3(N-1)} \int_{C_{1,1}\setminus C_{1/2,1/2}} u_0(R^{\delta_2}x.R^{\delta_3}\xi) P_1^*(\phi(x)\psi(\xi))\, dx\, d\xi \\ &= R^{-m+\delta_1+\delta_2+\delta_3(N-1)} \|P_1^*(\phi\psi)\|_\infty \|S^{II}_{R^{\delta_2}} S^{III}_{R^{\delta_3}} u_0\|_1 \lesssim R^{-m+\delta_1} \|u_0\|_1 . \end{aligned}$$

Since $u_0 \in L^1(\mathbb{R}^N)$, we find

$$|D_R| \to 0 \quad \text{if } -m + \delta_1 < 0, \tag{17}$$

Now we estimate

$$L_R \le \left(\iint_{C_{R^{\delta_1},R^{\delta_2},R^{\delta_3}} \setminus C_{R^{-\delta_1}/2,R^{-\delta_2}/2,R^{-\delta_3}/2}} |u|^q \eta_R(t)\phi_R(x)\psi_R(\xi)\,dx\,d\xi\,dt \right)^{1/q}$$
$$\times \left(\iint_{C_{R^{\delta_1},R^{\delta_2},R^{\delta_3}} \setminus C_{R^{\delta_1}/2,R^{\delta_2}/2,R^{\delta_3}/2}} \frac{|L^*\eta_R(t)\phi_R(x)\psi_R(\xi)|^{q'}}{|\eta_R(t)\phi_R(x)\psi_R(\xi)|^{(q'-1)}}\,dx\,d\xi\,dt \right)^{1/q'}$$
$$=: (I_R^\sharp)^{1/q} (\tilde{L}_R^\sharp)^{1/q'}.$$

Assuming (9), we can substitute $\eta\phi\psi$ with $(\eta\phi\psi)^\sigma$ with large $\sigma > mq'$ so that the function in the last integral is finite, see Lemma 2.1 in [3]. Moreover we have

$$\tilde{L}_R^\sharp \le R^{-mq'} \iint_{C_{R^{\delta_1},R^{\delta_2},R^{\delta_3}} \setminus C_{R^{\delta_1}/2,R^{\delta_2}/2,R^{\delta_3}/2}} S^I_{R^{\delta_1}} S^{II}_{R^{\delta_2}} S^{III}_{R^{\delta_3}} \frac{|L^*(\eta\phi\psi)|^{q'}}{|(\eta\phi\psi)|^{q'-1}}\,dx\,d\xi\,dt$$
$$= R^{-mq'+Q} \iint_{C_{1,1,1}\setminus C_{1/2,1/2,1/2}} \frac{|L^*(\eta\phi\psi)|^{q'}}{|(\eta\phi\psi)|^{q'-1}}\,dx\,d\xi\,dt.$$

In particular the last integral does not depend on $R > 1$. After Young inequality, we may conclude that there exists $C > 0$ such that

$$L_R \le I_R^\sharp/4q + C_q \tilde{L}_R^\sharp \le I_R^\sharp/4q + C R^{-mq'+Q}. \tag{18}$$

It remains to consider N_R. The main trick of this proof is the following relation that is the heart of BLMP equation from nonlinear point of view:

$$N_R = -\alpha \iint_{C_{R^{\delta_1},R^{\delta_2},R^{\delta_3}} \setminus C_{R^{\delta_1}/2,R^{\delta_2}/2,R^{\delta_3}/2}} \partial_x u S(\nabla_\xi) u \eta_R \partial_x \phi_R \psi_R\,dx\,d\xi\,dt$$

For $\alpha = \beta = 0$ this term has to be neglected.
Let $\beta > 0$. After Holder inequality we get

$$|N_R| \le \left(\iint_{C_{R^{\delta_1},R^{\delta_2},R^{\delta_3}} \setminus C_{R^{\delta_1}/2,R^{\delta_2}/2,R^{\delta_3}/2}} \beta |\nabla_{x,\xi} u|^{2q} \eta_R \phi_R \psi_R\,dx\,d\xi\,dt \right)^{1/q}$$
$$\times \frac{\alpha}{\beta^{1/q}} \left(\iint_{C_{R^{\delta_1},R^{\delta_2},R^{\delta_3}} \setminus C_{R^{\delta_1}/2,R^{\delta_2}/2,R^{\delta_3}/2}} \frac{|\partial_x \phi_R(x)|^{q'}}{|\phi_R(x)|^{q'-1}} \eta_R(t)\psi_R(\xi)\,dx\,d\xi\,dt \right)^{1/q'}$$
$$\le (I_R^\sharp)^{1/q} (N_R^\sharp)^{1/q'}.$$

Proceeding as before, we have

$$N_R \leq I_R^\sharp/4q + CR^{-\delta_2 q' + Q} \tag{19}$$

From (17), (18), (19) we can conclude that there exist η, ϕ, ψ such that

$$\int_0^\infty \int_{\mathbb{R}^N} (|u|^q + \beta|\nabla_{x,\xi} u|^{2q})\, \eta_R(t)\phi_R(x)\psi_R(\xi)\, dx\, d\xi\, dt \leq$$
$$\leq -D_R + C_1 R^{-mq' + Q} + C_2 R^{-\delta_2 q' + Q}\,.$$

For $1 < q < Q/(Q - \min\{m, \delta_2\})$, the right side goes to zero hence $u = 0$. For $u_0 \neq 0$, we get a maximal time existence for u. For BLMP Eq. (16) we get nonexistence of global weak solution below the critical exponent $q_0 = \frac{N+3}{N+2}$.

It remains to discuss the critical case. In (18) or (19) we have an exponent equal zero and a constant that does not depend on R, that is $L_R^\sharp$ and $N_R^\sharp$ are uniformly bounded from above. We may only conclude that $u \in L^q_{loc}([0, T) \times \mathbb{R}^N)$ and $\nabla_{x,\xi} u \in L^{2q}_{loc}([0, T) \times \mathbb{R}^N)$. From Lebesgue theorem this implies $I_R^\sharp \to 0$, so we can conclude the proof as before without using Young inequality. □

Remark 1 Taking $\alpha = \beta = 0$ in the proof of Theorem 4 we can deduce Theorem 2. Taking $u_0 = 0$ we can find the proof of Theorem 1 for the particular shape of $L = \partial_t P_1 + P_2$. Concerning Theorem 3 all the ingredients are similar to BLMP case, indeed after integration by parts $|u|^2$ instead of $|\nabla_{x,\xi} u|^2$ appears and it is compensated by $|u|^{2q}$. The main difference is the lack of $S(\nabla_\xi)$ in the linear part of KdV equation so that the positivity of the initial data is necessary.

4 Conclusion and Open Problems

Clearly this paper is only a starting point for studying nonlinear perturbations of BLMP equations. Let us list some open questions.

- With a different approach one can investigate the lifespan of the regular solutions of (16).
- One could study a more general perturbation such as

$$(\partial_t + \partial_x^3) S(\nabla_\xi) u = \alpha(\partial_x u S(\nabla_\xi)\partial_x u + \partial_x^2 u S(\nabla_\xi) u) + \beta_1 |u|^{q_1} + \beta_2 |\nabla_{x,\xi} u|^{q_2}$$

  and the interaction between q_1 and q_2.
- Similarly one can perturb generalized KdV equation:

$$(\partial_t + \partial_x^3) u = \partial_x (|u|^p) + \beta_1 |u|^{q_1} + \beta_2 |u|^{q_2}$$

  and investigate the interaction between p, q_1 and q_2.

- One can change the quasilinear part in such a way that after integration by parts one has

$$\int_{\mathbb{R}^N} (G_1(\partial_x u, \partial_x S(\nabla_\xi)u) + G_2(\partial_x^2 u, S(\nabla_\xi)u)\phi(x)\psi(\xi)\,dx\,d\xi$$
$$\leq \int |\nabla_{x,\xi} u|^p P(\partial_x, S(\nabla_\xi))(\phi(x)\psi(\xi))\,dx\,d\xi$$

for some $p > 1$ and linear operator P such that (14) holds. In BMLP case $p = 2$ and $P = \partial_x$. We believe that this kind of property is related to the possibility to write an equation in a bilinear form. For BLMP this happens by means of Hirota's differential operators. See for example [5].

Acknowledgments In this paper I wish to express my gratitude to Prof. M. Reissig for his efforts as President of the ISAAC in the period 2017–2021. I also thank him for being an important guide for many students, a special collaborator and a good friend.

References

1. Boiti, M., Leon, J., Manna, M., Pempinelli, F.: On the Spectral Transform of Korteweg-de Vries Equation in Two Spatial Dimensions. Inverse Problem **2**, 271–279 (1986)
2. D'Ambrosio, L., Lucente, S.: Nonlinear Liouville theorems for Grushin and Tricomi operators. J. Differ. Equ. **123**, 511–541 (2003)
3. D'Abbicco, M., Lucente, S.: A modified test function method for damped wave equations. Adv. Nonlinear Stud. **13**, 867–892 (2013)
4. Darvishi, M., Najafi, M., Kavitha, L., Venkatesh, M.: Stair and step soliton solutions of the integrable (2+1) and (3+1)-dimensional Boiti-Leon-Manna-Pempinelli equations. Commun. Theor. Phys. **58**, 785–794 (2012)
5. Liu, X., Dong, H., Zhang, Y., Chen, X.: The rational solutions and the interactions of the N-soliton solutions for Boiti-Leon-Manna-Pempinelli-like equation. J. Appl. Math. Phys. **5**, 700–714 (2017)
6. Manakov, S.V., Zakharov, V.E., Bordag, L.A., Its, A.R., Matveev, V.B.: Two-dimensional solitons of the Kadomtsev-Petviashvili equation and their interaction. Phys. Lett. A **63**, 205–206 (1977)
7. Mitidieri, E., Pohozaev, S.I.: Nonexistence of positive solutions for quasilinear elliptic problems on $\mathbb{R}^N$. Proc. Steklov Inst. Math. **227**, 186–216 (1999)
8. Xu, G., Wazwaz, A.: Integrability aspects and localized wave solutions for a new (4+1)-dimensional Boiti-Leon-Manna-Pempinelli equation. Nonlinear Dyn. **98**, 1379–1390 (2019)

Part XI
Wavelet Theory and Its Related Topics

Holomorphic Curves with Deficiencies and the Uniqueness Problem

Yoshihiro Aihara

Abstract In this note, we shall give an orverview of some results on holomorphic curves $f : \mathbf{C} \to \mathbf{P}^n(\mathbf{C})$ with deficiencies. We first recall theorems on the structure of the set of deficient divisors and give some uniqueness theorems for holomorphic curves. We also discuss several methods for constructing holomorphic curves with deficiencies.

1 Introduction

The aim of this note is to give an overview of some results on holomorphic curves with deficiencies. Let M be a smooth complex algebraic variety and $L \to M$ a very ample line bundle over M. We denote by $|L|$ the complete linear system of L. Let $f : \mathbf{C} \to M$ be a transcendental holomorphic curve. In [3], we gave the structure theorem for the set

$$\mathscr{D}_f = \{D \in |L|; \delta_f(D) > 0\}$$

of deficient divisors of f. The existence of holomorphic curves $f : \mathbf{C} \to M$ with $\mathscr{D}_f \neq \emptyset$ is a delicate matter. In the previous papers [1] and [2], we gave some uniqueness theorems for families of meromorphic maps $f : \mathbf{C}^m \to M$ with deficiencies. We studied how the existence of deficient divisors affects the uniqueness problem of meromorphic mappings. We shall give some unicity theorems for holomorphic curves with deficient hyprersurfaces in the case where $M = \mathbf{P}^n(\mathbf{C})$ (cf. [1]). We also consider methods for constructing holomorphic curves with deficiencies (cf. [3]). By making use of holomorphic curves of special exponential type, we shall construct holomorphic curves f of finite order with $\mathscr{D}_f \neq \emptyset$. In particular, the order ρ_f of f is a non-negative integer.

Y. Aihara (✉)
Fukushima University, Fukushima, Japan
e-mail: aihara@educ.fukushima-u.ac.jp

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_51

2 Preliminaries

We recall some known facts on Nevanlinna theory for holomorphic curves. For details, see [8] and [9]. In particular, for holomorphic line bundles and Chern classes, see the Section 1 of [8, Chapter 2].

Let z be the natural coordinate in $\mathbf{C}$, and set

$$\Delta(r) = \{z \in \mathbf{C};\ |z| < r\} \quad \text{and} \quad C(r) = \{z \in \mathbf{C};\ |z| = r\}.$$

For a (1,1)-current φ of order zero on $\mathbf{C}$ we set

$$N(r,\ \varphi) = \int_1^r \langle \varphi,\ \chi_{\Delta(t)} \rangle \frac{dt}{t},$$

where $\chi_{\Delta(r)}$ denotes the characteristic function of $\Delta(r)$.

Let E be an effective divisor on $\mathbf{C}$. We write $E = \sum_j k_j p_j$, where k_j are positive integer and $p_j \in \mathbf{C}$. For a positive integer l, we define by

$$N_l(r,\ E) = \int_1^r \sum_{p_j \in \Delta(t)} \min\{k_j, l\} \frac{dt}{t}$$

the l-truncated counting function of E.

Let M be a compact complex manifold and let $L \to M$ be a line bundle over M. We denote by $\Gamma(M,\ L)$ the space of all holomorphic sections of $L \to M$ and by $|L| = \mathbf{P}(\Gamma(M,\ L))$ the complete linear system of L. Denote by $||\cdot||$ a hermitian fiber norm in L and by ω its Chern form. Let $f : \mathbf{C} \to M$ be a holomorphic curve. We set

$$T_f(r,\ L) = N(r,\ f^*\omega)$$

and call it the characteristic function of f with respect to L. In the case where $M = \mathbf{P}^n(\mathbf{C})$ and L is the hyperplane bundle $\mathcal{O}_{\mathbf{P}^n}(1)$, we always take the Fubini-Study form ω_{FS}. We simply write $T_f(r)$ for $T_f(r, \mathcal{O}_{\mathbf{P}^n}(1))$. We notice

$$T_f(r, \mathcal{O}_{\mathbf{P}^n}(d)) = dT_f(r) + O(1).$$

for all positive integer d. If

$$\liminf_{r \to +\infty} \frac{T_f(r,\ L)}{\log r} = +\infty,$$

then f is said to be *transcendental*. We define *the order* ρ_f of $f : \mathbf{C} \to M$ by

$$\rho_f = \limsup_{r \to +\infty} \frac{\log T_f(r,\ L)}{\log r}.$$

We notice that the definition of ρ_f is independent of a choice of positive line bundles $L \to M$. Let $D = (\sigma) \in |L|$ with $||\sigma|| \leq 1$ on M. Assume that $f(\mathbf{C})$ is not contained in Supp D. We define the proximity function of D by

$$m_f(r,\ D) = \int_{C(r)} \log\left(\frac{1}{||\sigma(f(z))||}\right)\sigma(z).$$

Here σ is the invariant measure on $C(r)$ normalized so that $\sigma(C(1)) = 1$. Then we have the following first main theorem for holomorphic curves $\mathbf{C} \to M$.

Theorem 1 (First Main Theorem) *Let $L \to M$ be a line bundle over M and $f : \mathbf{C} \to M$ a non-constant holomorphic curve. Then*

$$T_f(r,\ L) = N(r,\ f^*D) + m_f(r,\ D) + O(1)$$

for $D \in |L|$ with $f(\mathbf{C}) \not\subseteq$ Supp D, where $O(1)$ stands for a bounded term as $r \to +\infty$.

Let f and D be as above. We define *Nevanlinna's deficiency* $\delta_f(D)$ by

$$\delta_f(D) = \liminf_{r\to+\infty} \frac{m_f(r,\ D)}{T_f(r,\ L)}.$$

It is clear that $0 \leq \delta_f(D) \leq 1$. Then we have a defect function δ_f defined on $|L|$. If $\delta_f(D) > 0$, then D is called *a deficient divisor in the sense of Nevanlinna*. Let $D_1, \ldots, D_q$ be smooth hypersurfaces of degree d in $\mathbf{P}^n(\mathbf{C})$. Set $Q = \{1, \ldots, q\}$.

Definition 1 We say that $D_1, \ldots, D_q$ are in general position if

$$\bigcap_{j\in R} \text{Supp } D_j = \emptyset \quad \text{for every subset } R \subseteq Q \text{ with } \sharp R = n+1.$$

In the case where $D_1, \ldots, D_q$ are hyperplanes, H. Cartan's second main theorem is well known. Now, assume that $D_1, \ldots, D_q$ are hypersurfaces of $d \geq 2$. For a positive real number s, we let $\lceil s \rceil$ denote the least positive integer not less than s. The following inequality of second main theorem type is due to An and Phuong [5].

Theorem 2 (An–Phuong) *Let $f : \mathbf{C} \to \mathbf{P}^n(\mathbf{C})$ be an algebraically non-degenerate holomorphic curve and let $D_1, \ldots, D_q$ be smooth hypersurfaces of degree d in general position in $\mathbf{P}^n(\mathbf{C})$. For $0 < \varepsilon < 1$, set $l = 2d\lceil 2^n(n+1)n(d+1)\varepsilon^{-1}\rceil^n$. Then*

$$(q-n-1)T_f(r, \mathscr{O}_{\mathbf{P}^n}(d)) \leq \sum_{j=1}^{q} N_l(r, f^*D_j) + \varepsilon T_f(r, \mathscr{O}_{\mathbf{P}^n}(d))$$

for all r except on $E \subseteq [1, \infty)$ with finite Lebesgue measure.

3 Deficiencies of the Base Loci of Linear Systems

We shall define the deficiency of the base locus of a linear system. We recall some basic facts in value distribution theory for coherent ideal sheaves (cf. [9, Chapter 2]). Let $f : \mathbf{C} \to M$ be a holomorphic curve and $\mathscr{I}$ a coherent ideal sheaf of the structure sheaf $\mathscr{O}_M$ of M. Let $\mathscr{U} = \{U_j\}$ be a finite open covering of M with a partition of unity $\{\eta_j\}$ subordinate to $\mathscr{U}$. We can assume that there exist finitely many sections $\sigma_{jk} \in \Gamma(U_j, \mathscr{I})$ such that every stalk $\mathscr{I}_p$ over $p \in U_j$ is generated by germs $(\sigma_{j1})_p, \ldots, (\sigma_{jl_j})_p$. Set

$$d_{\mathscr{I}}(p) = \left(\sum_j \eta_j(p) \sum_{k=1}^{l_j} |\sigma_{jk}(p)|^2 \right)^{1/2}.$$

We may assume that $d_{\mathscr{I}}(p) \leq 1$ for all $p \in M$. Set

$$\phi_{\mathscr{I}}(p) = -\log d_{\mathscr{I}}(p)$$

and call it the proximity potential for $\mathscr{I}$. It is easy to verify that $\phi_{\mathscr{I}}$ is well-defined up to addition by a bounded continuous function on M. We now define the proximity function $m_f(r, \mathscr{I})$ of f for $\mathscr{I}$, or equivalently, for the complex analytic subspace (may be non-reduced)

$$Y = (\mathrm{Supp}\,(\mathscr{O}_M/\mathscr{I}), \mathscr{O}_M/\mathscr{I})$$

by

$$m_f(r, \mathscr{I}) = \int_{C(r)} \phi_{\mathscr{I}}(f(z))\sigma(z),$$

provided that $f(\mathbf{C})$ is not contained in $\mathrm{Supp}\, Y$. For $z_0 \in f^{-1}(\mathrm{Supp}\ Y)$, we can choose an open neighborhood U of z_0 and a positive integer ν such that

$$f^*\mathscr{I} = ((z - z_0)^\nu) \quad \text{on } U.$$

Then we see

$$\log d_{\mathscr{I}}(f(z)) = \nu \log|z - z_0| + h_U(z) \quad \text{for } z \in U,$$

where h_U is a C^∞-function on U. Thus we have the counting function $N(r, f^*\mathscr{I})$ as above. Moreover, we set

$$\omega_{\mathscr{I},f} = -dd^c h_U \quad \text{on } U,$$

where $d^c = (\sqrt{-1}/4\pi)(\overline{\partial} - \partial)$. We obtain a well-defined smooth $(1,\ 1)$-form $\omega_{\mathscr{I},f}$ on $\mathbf{C}$. Define the characteristic function $T_f(r,\ \mathscr{I})$ of f for $\mathscr{I}$ by

$$T_f(r,\ \mathscr{I}) = \int_1^r \frac{dt}{t} \int_{\Delta(t)} \omega_{\mathscr{I},f}.$$

We have the first main theorem in value distribution theory for coherent ideal sheaves:

Theorem 3 (First Main Theorem) *Let $f : \mathbf{C} \to M$ and $\mathscr{I}$ be as above. Then*

$$T_f(r,\ \mathscr{I}) = N(r,\ f^*\mathscr{I}) + m_f(r,\ \mathscr{I}) + O(1).$$

Let $L \to M$ be an ample line bundle and $W \subseteq \Gamma(M,\ L)$ a linear subspace with $\dim W \geq 2$. Set $\Lambda = \mathbf{P}(W)$. The base locus $\mathrm{Bs}\,\Lambda$ of Λ is defined by

$$\mathrm{Bs}\,\Lambda = \bigcap_{D \in \Lambda} \mathrm{Supp}\, D.$$

We define a coherent ideal sheaf $\mathscr{I}_0$ in the following way. For each $p \in M$, the stalk $\mathscr{I}_{0,p}$ is generated by all germs $(\sigma)_p$ for $\sigma \in W$. Then $\mathscr{I}_0$ defines the base locus of Λ as a complex analytic subspace B_Λ, that is,

$$B_\Lambda = (\mathrm{Supp}\ (\mathscr{O}_M/\mathscr{I}_0),\ \mathscr{O}_M/\mathscr{I}_0).$$

Hence $\mathrm{Bs}\,\Lambda = \mathrm{Supp}\ (\mathscr{O}_M/\mathscr{I}_0)$. We define the *deficiency of* B_Λ for f by

$$\delta_f(B_\Lambda) = \liminf_{r \to +\infty} \frac{m_f(r,\ \mathscr{I}_0)}{T_f(r,\ L)}.$$

4 Structure Theorems for the Set of Deficient Divisors

In this section, we shall summarize theorems on the structure of the set of deficient divisors of holomorphic curves. Let $L \to M$ be an ample line bundle and $f : \mathbf{C} \to M$ a transcendental holomorphic curve. Let $\Lambda \subseteq |L|$ be a linear system. We say that f is *non-degenerate with respect to* Λ if $f(\mathbf{C})$ is not contained in $\mathrm{Supp}\, D$ for all $D \in \Lambda$. Set

$$\mathscr{D}_f = \{D \in \Lambda;\ \delta_f(D) > \delta_f(B_\Lambda)\}.$$

We call $\mathscr{D}_f$ *the set of deficient divisors* in Λ.

By making use of the generalized Crofton's formula due to R. Kobayashi ([9, Theorem 2.4.12]), we have the following proposition ([3, Proposition 4.1]).

Proposition 1 *Suppose that f is non-degenerate with respect to Λ. Then the set $\mathscr{D}_f$ is a null set in the sense of the Lebesgue measure on Λ. In particular $\delta_f(D) = \delta_f(B_\Lambda)$ for almost all $D \in \Lambda$.*

Definition 2 If $\rho_f < +\infty$, then f is said to be of finite type.

Then we have the structure theorem for the set $\mathscr{D}_f$ (see [3, §5]).

Theorem 4 *Suppose that f is of finite type and non-degenerate with respect to Λ. Then the set $\mathscr{D}_f$ of deficient divisors is a union of at most countably many linear systems included in Λ. Furthermore, the set $\delta_f(\Lambda)$ of values of deficiency of f is an at most countable subset $\{e_i\}$ of $[0, 1]$. For each e_i, there are linear systems $\Lambda_1(e_i), \ldots, \Lambda_s(e_i)$ included in Λ such that $e_i = \delta_f(B_{\Lambda_j(e_i)})$ for $j = 1, \ldots, s$.*

By the above theorem, there exists a family $\{\Lambda_j\}$ of at most countably many linear systems in Λ such that $\mathscr{D}_f = \bigcup_j \Lambda_j$. Define $\mathscr{L}_f = \{\Lambda_j\} \cup \{\Lambda\}$. We call $\mathscr{L}_f$ *the fundamental family of linear systems for* f.

Proposition 2 *If $\delta_f(D) > \delta_f(B_\Lambda)$ for a divisor D in Λ, then there exists a linear system $\Lambda(D) \in \mathscr{L}_f$ such that $\delta_f(D) = \delta_f(B_{\Lambda(D)})$.*

5 Unicity Theorems for Holomorphic Curves

In this section, we shall give unicity theorems for holomorphic curves f into $\mathbf{P}^n(\mathbf{C})$. In particular, we show that the existence of deficient divisors gives a strong effect on the unicity of holomorphic curves. Let $D_1, \ldots, D_q$ be smooth hypersurfaces of degree d in general position in $\mathbf{P}^n(\mathbf{C})$ and let $E_1, \ldots, E_q$ be divisors in $\mathbf{C}$ with $\operatorname{Supp} E_i \cap \operatorname{Supp} E_j = \emptyset \quad (i \neq j)$. By making use of Theorem 2 with $\varepsilon = 1/2$, Dulock and Ru [6] proved the uniqueness theorem as follows.

Theorem 5 (Dulock–Ru) *Let $f, g : \mathbf{C} \to \mathbf{P}^n(\mathbf{C})$ be algebraically non-degenerate holomorphic curves. Suppose that* $\operatorname{Supp} f^* D_j = \operatorname{Supp} g^* D_j = \operatorname{Supp} E_j$ *and $f = g$ on $\cup_j \operatorname{Supp} E_j$. Set $l_0 = 2d(2^{n-1}(n+1)n(d+1))^n$. If $q > (n+1)+(2ln/d)+(1/2)$, then f and g are identical.*

We first give a generalization of Theorem 5. For an effective divisor E on $\mathbf{C}$ and a positive integer k, we denote by $\operatorname{Supp}_k E$ the union of all irreducible components of E with the multiplicities at most k. In what follows we fix a transcendental algebraically non-degenerate holomorphic curve $f_0 : \mathbf{C} \to \mathbf{P}^n(\mathbf{C})$. Let $k_1, \ldots, k_q$ be positive integers and set $k_0 = \max\{k_1, \ldots, k_q\}$. Assume that

$$\operatorname{Supp}_{k_j} f_0^* D_j = \operatorname{Supp} E_j \quad \text{for all } j = 1, \ldots, q.$$

We denote by

$$\mathscr{F} = \mathscr{F}(f_0; \{k_j\}; (\mathbf{C}, \{E_j\}), (\mathbf{P}^n(\mathbf{C}), \{D_j\}))$$

the set of all algebraically non-degenerate transcendental holomorphic curves $f : \mathbf{C} \to \mathbf{P}^n(\mathbf{C})$ such that

$$\mathrm{Supp}_{k_j} f^* D_j = \mathrm{Supp}\, E_j, \quad j = 1, \ldots, q.$$

and

$$f = f_0 \quad \text{on} \quad \bigcup_{j=1}^{q} \mathrm{Supp}\, E_j.$$

Let $l_0 = 2d(2^{n-1}(n+1)n(d+1))^n$ and set

$$\kappa(\{D_j\}, \{E_j\}; \{k_j\}) = q - n - 2 - \sum_{j=1}^{q} \frac{l_0}{k_j + 1} - \frac{2l_0 k_0}{k_0 + 1}.$$

We have the following unicity theorem for the family $\mathscr{F}$ by Theorem 2 and methods similar to the argument in [1].

Theorem 6 *If* $\kappa(\{D_j\}, \{E_j\}; \{k_j\}) > n + 1$, *then* $\mathscr{F} = \{f_0\}$.

In the case where $\kappa(\{D_j\}, \{E_j\}; \{k_j\}) = n + 1$, we have the following unicity theorem.

Theorem 7 *Suppose that* $\kappa(\{D_j\}, \{E_j\}; \{k_j\}) = n + 1$. *If* $\delta_{f_0}(D_j) > 0$ *for at least one* D_j $(1 \le j \le q)$, *then* $\mathscr{F} = \{f_0\}$.

Remark 1 In the proofs of Theorems 6 and 7, we essentially use that the canonical bundle of $\mathbf{P}^n(\mathbf{C})$ is $\mathscr{O}_{\mathbf{P}^n}(-(n+1))$.

6 Methods for Constructing Holomorphic Curves with Deficiencies

In this section, we consider the case where $M = \mathbf{P}^n(\mathbf{C})$ and $L = \mathscr{O}_{\mathbf{P}^n}(d)$ for a positive integer d. The existence of $f : \mathbf{C} \to \mathbf{P}^n(\mathbf{C})$ with $\mathscr{D}_f \neq \emptyset$ is a delicate matter. In fact, Mori [7] showed that a family

$$\{f \in \mathrm{Hol}(\mathbf{C}, \mathbf{P}^n(\mathbf{C})); \delta_f(H) = 0 \text{ for all } H \in |\mathscr{O}_{\mathbf{P}^n}(1)|\}$$

of holomprphic curves is dense in $\mathrm{Hol}(\mathbf{C}, \mathbf{P}^n(\mathbf{C}))$ with respect to a certain kind of topology. However, for any $\Lambda \subseteq |\mathscr{O}_{\mathbf{P}^n}(d)|$, there exists an algebraically non-degenerate holomorphic curve $f : \mathbf{C} \to \mathbf{P}^n(\mathbf{C})$ with $\mathscr{D}_f \neq \emptyset$. In fact, we have the following theorem ([4, Theorem 3.2]).

Theorem 8 (Aihara–Mori) *Let $D \in |\mathscr{O}_{\mathbf{P}^n}(d)|$. There exists a constant $\lambda(D)$ with $0 < \lambda(D) \leq d$ depending only on D that satisfies the following property: Let α be a positive real constant such that*

$$0 < \alpha \leq \frac{\lambda(D)}{d}.$$

Then there exists an algebraically non-degenerate transcendental holomorphic curve $f : \mathbf{C} \to \mathbf{P}^n(\mathbf{C})$ such that $\delta_f(D) = \alpha$.

Remark 2 Let f be as in Theorem 8. By making use of Theorem 4 and the proof of Theorem 8 ([4, pp. 239–244]), we can show that the set $\delta_f(\Lambda)$ of values of δ_f is a finite set for all $\Lambda \subseteq |\mathscr{O}_{\mathbf{P}^n(d)}|$.

The proof of the above theorem is based on G. Valiron's theorem on entire functions of order zero. Hence the resulting holomorphic curves are of order zero. Let ρ is a positive integer. We can construct holomorphic curves $f : \mathbf{C} \to \mathbf{P}^n(\mathbf{C})$ with $\mathscr{D}_f \neq \emptyset$ and $\rho_f = \rho$ by another way.

We recall some basic results on holomorphic curves of special exponential type in $\mathbf{P}^n(\mathbf{C})$. For details, see [10]. Let λ be a positive integer and denote by $\mathbf{Z}^+$ the set of positive integers. We let $\mathscr{E}_\lambda$ denote the ring of holomorphic functions of the form

$$g = \sum_{j=1}^{k} \phi_j(z) \exp P_j(z),$$

where P_j are polynomials on $\mathbf{C}$ of degree at most λ and $\phi_j(z)$ are meromorphic functions on $\mathbf{C}$ such that

$$T_{\phi_j}(r) = o(r^\lambda).$$

We note that $g \in \mathscr{E}_\lambda$ must be holomorphic, although the ϕ_j may be meromorphic. For example, a holomorphic function

$$g(z) = \frac{1}{z} \exp z^\lambda - \frac{1}{z} \exp 0$$

is contained in $\mathscr{E}_\lambda$. Hence $\mathscr{E}_\lambda \neq \emptyset$ for all $\lambda \in \mathbf{Z}^+$. Set

$$\mathscr{E}_\lambda^{n+1}(\mathbf{C}) = \{(g_0, \ldots, g_n);\, g_0, \ldots, g_n \in \mathscr{E}_\lambda\}.$$

Definition 3 A holomorphic curve $f : \mathbf{C} \to \mathbf{P}^n(\mathbf{C})$ is called a special exponential curve of order λ if f has a reduced representation $\tilde{f}$ in $\mathscr{E}_\lambda^{n+1}(\mathbf{C})$ and the order ρ_f of f is λ. In particular, a holomoprphic curve $f : \mathbf{C} \to \mathbf{P}^n(\mathbf{C})$ defined by

$$\tilde{f} = (\exp a_0 z^\lambda, \ldots, \exp a_n z^\lambda) \quad (a_1, \ldots, a_n \in \mathbf{C})$$

is called a simple exponential curve of order λ.

We let $\mathscr{S}_\lambda$ denote the set of special exponential curves of order λ and set

$$\mathscr{S} = \bigcup_{\lambda \in \mathbf{Z}^+} \mathscr{S}_\lambda.$$

We shall compute the characteristic function $T_f(r)$ of a holomorphic curve f in $\mathscr{S}_\lambda$. In general, it is difficult to compute $T_f(r)$. We consider the case where holomorphic curves $f : \mathbf{C} \to \mathbf{P}^n(\mathbf{C})$ defined by

$$\tilde{f}(z) = (\exp P_0(z), \ldots, \exp P_n(z)),$$

where P_j are polynomials of degree λ. We regard the polynomial 0 as a polynomial of degree λ. Hence 0 may be contained in $\mathscr{P}$. Let $\mathscr{P} = \{P_0, \ldots, P_n\}$ be a finite collection of all polynomials. For a set $A \subseteq \mathbf{C}$, we let $\mathscr{C}(A)$ denote the circumference of the convex hul l $\hat{A}$ of A. Then Shiffman ([10]) proved the following lemma.

Lemma 1 *Set*

$$K(\mathscr{P}) = \int_{C(1)} \mathscr{C}(\mathscr{P}(z))\sigma(z),$$

where $\mathscr{P}(z) = \{P_j(z);\ P_j \in \mathscr{P}\}$. *Then*

$$T_f(r) = K(\mathscr{P})r^\lambda + O(1).$$

For a simple exponential curve f defined by

$$\tilde{f}(z) = (\exp a_0 z^\lambda, \ldots, \exp a_n z^\lambda),$$

we can give $T_f(r)$ explicitly. We denote by $\mathscr{C}_f$ the circumference of the convex polygon spanned by the set $\{a_0, \ldots, a_n\}$. If the convex polygon reduces to the segment with the end points with a_j and a_k, then we see $\mathscr{C}_f = 2|a_j - a_k|$. Then we have,

$$T_f(r) = \frac{\mathscr{C}_f}{2\pi} r^\lambda + O(1).$$

We first assume $d = 1$. Let H be a hyperplane in $\mathbf{P}^n(\mathbf{C})$ defined by

$$\sum_{j=0}^{n} \alpha_j \zeta_j = 0 \quad (\alpha_0, \ldots, \alpha_n \in \mathbf{C}),$$

where $\zeta = (\zeta_0 : \ldots : \zeta_n)$ is a homogeneous coordinate system in $\mathbf{P}^n(\mathbf{C})$. We define the set J_H of index by $J_H = \{j\,;\ \alpha_j \neq 0\}$. Let $\mathscr{C}_f(H)$ be the circumference of

the convex polygon spanned by the set $\{a_j\,;\ j \in J_H\}$. Then we have the following lemma [3, Lemma 6.6].

Lemma 2 *The deficiency of f for H is given by*

$$\delta_f(H) = 1 - \frac{\mathscr{C}_f(H)}{\mathscr{C}_f}.$$

We notice that the constant $\mathscr{C}_f(H)$ depends only on f and J_H.

Next, we consider the case where $d \geq 2$. Let $v_d : \mathbf{P}^n(\mathbf{C}) \to \mathbf{P}^m(\mathbf{C})$ be the Veronese map of degree d, where $m = \binom{n+d}{d} - 1$. Set $h = v_d \circ f$.Then the holomorphic curve

$$h : \mathbf{C} \to \mathbf{P}^m(\mathbf{C})$$

is also a simple exponential curve of order λ. Set $W = v_d(\mathbf{P}^n(\mathbf{C}))$. Then W is a smooth subvariety of $\mathbf{P}^m(\mathbf{C})$. For a hypersurface D of degree d in $\mathbf{P}^m(\mathbf{C})$, the image $v_d(D)$ is $W \cap H$ for a hyperplane H in $\mathbf{P}^m(\mathbf{C})$. Hence the method in the proof of Lemma 2 works for the case where $d \geq 2$. By making use of Theorem 4 and Lemma 2, we have the following theorem.

Theorem 9 *Let $\Lambda \subseteq |\mathscr{O}_{\mathbf{P}^n}(d)|$ and let λ be a positive integer. Then there is a transcendental holomorphic curve $f : \mathbf{C} \to \mathbf{P}^n(\mathbf{C})$ that is non-degenerate with respect to Λ and $\rho_f = \lambda$ such that the set $\delta_f(\Lambda)$ of values of δ_f is a finite set $\{e_1, \ldots, e_t\}$ with $0 < e_j \leq 1$. Furthermore, there are finitely many linear systems $\{\Lambda_1, \ldots, \Lambda_t\}$ such that*

$$\mathscr{D}_f = \bigcup_{j=1}^{t} \Lambda_j \qquad \text{and} \qquad \delta_f(D) = e_j \quad \text{for all} \quad D \in \Lambda_j \setminus \Big(\bigcup_{k \neq j} \Lambda_k\Big).$$

In Theorem 9, the set $\delta_f(\Lambda)$ is finite. It is a very difficult problem to construct a holomorphic curve f such that $\delta_f(\Lambda)$ is an infinite set.

Acknowledgments The author would like to thank the referees for their useful advice and valuable comments.

References

1. Aihara, Y.: Unicity theorems for meromorphic mappings with deficiencies. Complex Variavles Theory Appl. **42**, 259–268 (2000)
2. Aihara, Y.: Algebraic dependence of meromorphic mappings in value distribution theory. Nagoya Math. J. **169**, 145–178 (2003)
3. Aihara, Y.: Deficiencies of holomorphic curves in algebraic varieties. Tohoku Math. J. **64**, 287–315 (2012)

4. Aihara, Y., Mori, S.: Deficiencies of meromorphic mappings for hypersurfaces. J. Math. Soc. Jpn. **57**, 233–258 (2005)
5. An, T.T.H., Phuong, H.T.: An explicit estimate on multiplicity truncation in the second main theorem for holomorphic curves encountering hypersurfaces in general position in projective space. Houston J. Math. **35**, 775–786 (2009)
6. Dulock, M., Ru, M.: A uniqueness theorem for holomorphic curves sharing hypersurfaces. Complex Var. Elliptic Equ. **53**, 797–802 (2008)
7. Mori, S.: Defects of holomorphic curves into $\mathbf{P}^n(\mathbf{C})$ for rational moving targets and a space of holomorphic curves, Complex Variables Theory Appl. **43**, 363–379 (2001)
8. Noguchi, J., Ochiai, T.: Geometric Function Theory in Several Complex Variables. Translations of Mathematical Monographs, vol. 80. American Mathematical Society, Providence (1990)
9. Noguchi, J., Winkelmann, J.: Nevanlinna Theory in Several Complex Variables and Diophantine Approximation. Springer-Verlag, Tokyo-Heiderberg-New York-Dordrecht-London (2014)
10. Shiffman, B.: On holomorphic curves and meromorphic maps in projective spaces, Indiana Univ. Math. J. **28**, 627–641 (1979)

On Some Topics Related to the Gabor Wavelet Transform

Keiko Fujita

Abstract We have studied the Gabor wavelet transform, the windowed Fourier transform and the Fourier transform of analytic functional on the sphere. In the case of the sphere, the space of the square integrable functions on the sphere is a subspace of the space of analytic functionals. In this note, we will review our previous results and consider the relationship between the Gabor wavelet transform and the Fourier transform.

1 Introduction

For $z = (z_1, z_2, \cdots, z_{n+1}) \in \mathbf{C}^{n+1}$ and $w = (w_1, w_2, \cdots, w_{n+1}) \in \mathbf{C}^{n+1}$, we set

$$z \cdot w = z_1 w_1 + \cdots + z_{n+1} w_{n+1}, \quad z^2 = z \cdot z, \quad \|z\|^2 = z \cdot \overline{z}.$$

Let f be an integrable function on $\mathbf{R}^{n+1}$. The Fourier transform of f is defined by

$$\int_{R^{n+1}} e^{-ix\cdot\omega} \overline{f(x)} dx,$$

and the windowed Fourier transform with respect to the window function w is defined by

$$\int_{R^{n+1}} w(x-y) e^{-ix\cdot\omega} \overline{f(x)} dx.$$

Let $\omega_0 \in \mathbf{C}^{n+1} \setminus \{0\}$ be fixed. Put

$$G_{\omega_0}(x) = e^{-x^2/2} e^{-ix\cdot\omega_0}.$$

K. Fujita (✉)
University of Toyama, Toyama, Japan
e-mail: keiko@sci.u-toyama.ac.jp

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_52

For $a \in \mathbf{R}_+ = \{x : x > 0\}$, the Gabor wavelet transform of f is defined by

$$\frac{1}{a}\int_{\mathbf{R}^{n+1}} G_{\omega_0}\left(\frac{x-\tau}{a}\right)\overline{f(x)}dx.$$

In this note, we will treat the Gabor wavelet transform on the unit sphere S^n in $\mathbf{R}^{n+1}$; that is,

$$S^n = \{(x_1, x_2, \cdots, x_{n+1}) \in \mathbf{R}^{n+1}; x^2 = 1\}.$$

In Sect. 2, we will see the relationship between the Fourier transform and the Gabor wavelet transform of f. In Sect. 3, we will see the expansion formula of the Gabor wavelet transform of f. In Sect. 4, we remark on the inverse Gabor wavelet transformation.

2 Gabor Wavelet Transformation on the Sphere

In this section, first we will review the Fourier transformation on the sphere. Next, we will consider the windowed Fourier transformation on the sphere and consider the relationship between the the windowed Fourier transform and the Fourier transform. Then we will consider the Gabor wavelet transformation on the sphere and consider the relationship between the the Gabor wavelet transform and the Fourier transform.

2.1 Fourier Transformation on the Sphere

We denote by $d\Omega_{n+1}(x)$ the non-normalized invariant measure on S^n induced by the Lebesgue measure $dx = dx_1 dx_2 \cdots dx_{n+1}$ and by Ω_{n+1} the volume of S^n measured by this measure:

$$\Omega_{n+1} = \mathrm{vol}(S^n) = \int_{S^n} d\Omega_{n+1} = \frac{2\pi^{(n+1)/2}}{\Gamma((n+1)/2)},$$

where $\Gamma(\cdot)$ is the Gamma function.

For an integrable function f on S^n, we define the Fourier transformation $\mathcal{F}$ by

$$\mathcal{F} : f \mapsto (\mathcal{F}f)(w) = \int_{S^n} e^{-ix\cdot w}\overline{f(x)}d\omega_{n+1}(x), \tag{1}$$

where $d\omega_{n+1}$ is the normalized invariant measure on S^n: $d\omega_{n+1} = d\Omega_{n+1}/\Omega_{n+1}$.

For the square integrable functions f and g on S^n, we define a sesquilinear form $(f,\ g)_{S^2}$ by

$$(f,\ g)_{S^n} \equiv \int_{S^n} f(x)\overline{g(x)}d\omega_{n+1}(x).$$

Then $(\cdot\,,\ \cdot)_{S^n}$ gives an inner product. We denote by $L^2(S^n)$ the space of square integrable functions on S^n with the inner product $(\cdot\,,\ \cdot)_{S^n}$, and the norm $\|\cdot\|_{S^n}$ of f is given by $\|f\|_{S^n} = \sqrt{(f,\ f)_{S^n}}$.

2.2 Windowed Fourier Transformation on the Sphere

For $b > 0$, put

$$f_y(x) = \exp\left(-\frac{(x-y)^2}{2b^2}\right) = \exp\left(-\frac{x^2+y^2-2x\cdot y}{2b^2}\right),\quad y \in \mathbf{R}^{n+1}.$$

Then for $x \neq y$, $\lim_{b\to 0} f_y(x) = 0$ and $\lim_{b\to\pm\infty} f_y(x) = 1$. For $x, y \in S^n$, we have $1 - x\cdot y \geq 0$ and

$$\exp(-2/b^2) \leq f_y(x) = \exp(-(1-x\cdot y)/b^2) \leq 1. \tag{2}$$

We know

$$\int_{S^n} \exp(x\cdot\zeta)d\omega_{n+1}(x) = \tilde{j}_0\left(i\sqrt{\zeta^2}\right) = \sum_{l=0}^{\infty} \frac{\Gamma((n+1)/2)}{l!\Gamma(l+(n+1)/2)}\left(\frac{\zeta^2}{4}\right)^l.$$

(See Sect. 3.) Therefore for $y \in S^n$, we have

$$\exp(-2/b^2) \leq \int_{S^n} f_y(x)d\omega_{n+1} = \exp(-1/b^2)\tilde{j}_0\left(i/b^2\right) \leq 1 \tag{3}$$

by (2). By (3), we have $\exp(-t) \leq \tilde{j}_0(it) \leq \exp(t)$ for $t \in \mathbf{R}$.

Let $b \in \mathbf{R}\setminus\{0\}$ and put

$$w_b(x) = \exp(-x^2/(2b^2)).$$

Note that $W_b(x-\tau) = f_\tau(x)$. For $f \in L^2(S^n)$ and $\omega, \tau \in \mathbf{C}^{n+1}\setminus\{0\}$, we define the windowed Fourier transformation $\mathcal{W}_b\mathcal{F}$ with the window function $w_b(x)$ by

$$\mathcal{W}_b\mathcal{F}:\ f \mapsto (\mathcal{W}_b\mathcal{F}f)(\tau,\omega) = \int_{S^n} e^{-ix\cdot\omega}w_b(x-\tau)\overline{f(x)}d\omega_{n+1}(x). \tag{4}$$

For $x \in S^n$ we have

$$e^{-ix\cdot\omega} w_b(x-\tau) = e^{\frac{-1-\tau^2}{2b^2}} e^{-ix\cdot\frac{\omega+i\tau}{b^2}} = e^{\frac{-1-\tau^2}{2b^2}} e^{-ix\cdot\omega} e^{x\cdot\tau/b^2}.$$

Thus we have

$$(\mathcal{W}_b \mathcal{F} f)(\tau,\omega) = e^{-\frac{1+\tau^2}{2b^2}} (\mathcal{F} f)(\omega + i\tau/b^2) = e^{-\frac{1+\tau^2}{2b^2}} (\mathcal{F} g)(\omega), \tag{5}$$

where we put $g(x) = \overline{e^{x\cdot\tau/b^2}} f(x)$. Therefore, the windowed Fourier transform of $f \in L^2(S^n)$ can be expressed by means of the Fourier transform of f.

2.3 Gabor Wavelet Transformation on the Sphere

For $f \in L^2(S^n)$ and $a \in \mathbf{R}_+ = \{x : x > 0\}$, similar to the case of $\mathbf{R}^{n+1}$, we define the Gabor wavelet transformation $\mathcal{G}_{\omega_0}$ on S^n by

$$\mathcal{G}_{\omega_0} : f \mapsto (\mathcal{G}_{\omega_0} f)(\tau, a) = \frac{1}{a} \int_{S^n} G_{\omega_0}\left(\frac{x-\tau}{a}\right) \overline{f(x)} d\omega(x). \tag{6}$$

For $x \in S^n$ we have

$$G_{\omega_0}\left(\frac{x-\tau}{a}\right) = \frac{1}{a} e^{-\frac{1+\tau^2}{2a^2}} e^{i\tau\cdot\frac{\omega_0}{a}} e^{-ix\cdot(\frac{\omega_0}{a}+i\frac{\tau}{a^2})} = \frac{1}{a} e^{-\frac{1+\tau^2}{2a^2}} e^{i\tau\cdot\frac{\omega_0}{a}} e^{-ix\cdot\frac{\omega_0}{a}} e^{\frac{x\cdot\tau}{a^2}}.$$

Thus we have

$$(\mathcal{G}_{\omega_0} f)(\tau, a) = \frac{1}{a} e^{-\frac{1+\tau^2}{2a^2}} e^{i\tau\cdot\frac{\omega_0}{a}} (\mathcal{F} f)\left(\frac{\omega_0}{a} + i\frac{\tau}{a^2}\right) \tag{7}$$

$$= \frac{1}{a} e^{-\frac{1+\tau^2}{2a^2}} e^{i\tau\cdot\frac{\omega_0}{a}} (\mathcal{F} g)\left(\frac{\omega_0}{a}\right), \tag{8}$$

where we put $g(x) = \overline{\exp(x\cdot\tau/a^2)} f(x)$.

By (5), (7) and (8), if we can construct the inverse mapping of the Fourier transformation, we can find the inverse mappings of the windowed Fourier transformation, and of the Gabor wavelet transformation. Since we constructed the inverse mapping of the Fourier transformation, we can construct the inverse mappings of the windowed Fourier transformation, and of the Gabor wavelet transformation.

3 Expansion Formula of the Gabor Wavelet Transform

In this section, first we will review some notation to consider the image of the Gabor wavelet transform. Next, we will review the expansion formula.

3.1 Spherical Harmonics Expansion

We will express the images of (1), (4) and (6) by means of the infinite sum of the Legendre polynomials and the Bessel polynomials. We recall some notations.

Let $P_{k,n}(t)$ be the Legendre polynomial of degree k and of dimension $n+1$:

$$P_{k,n}(t) = \left(\frac{-1}{2}\right)^2 \frac{\Gamma(\frac{n}{2})}{\Gamma(k+\frac{n}{2})}(1-t^2)^{\frac{2-n}{2}} \frac{d^k}{dt^k}(1-t^2)^{k+\frac{n-2}{2}}.$$

We define the extended Legendre polynomial by

$$P_{k,n}(z,w) = (\sqrt{z^2})^k(\sqrt{w^2})^k P_{k,n}\left(\frac{z}{\sqrt{z^2}} \cdot \frac{w}{\sqrt{w^2}}\right), \quad z, w \in \mathbf{C}^{n+1}.$$

Then $P_{k,n}(z,w)$ is a homogeneous harmonic polynomial of degree k in z and in w. That is, $P_{k,n}(z,w) = P_{k,n}(w,z)$ and $\Delta_z P_{k,n}(z,w) \equiv (\frac{\partial^2}{\partial z_1^2} + \cdots + \frac{\partial^2}{\partial z_{n+1}^2})P_{k,n}(z,w) = 0$. We denote by $N(k,n)$ the dimension of the space of homogeneous harmonic polynomials of degree k. We know

$$N(k,n) = \frac{(2k+n-1)(k+n-2)!}{k!(n-1)!} \ .$$

By the orthogonality of the Legendre polynomials, we have

$$N(k,n)\int_{S^n} P_{k,n}(x,z)P_{l,n}(x,w)d\omega_{n+1}(x) = \delta_{kl}P_{k,n}(z,w) \ . \tag{9}$$

For $\nu \neq -1, -2, \cdots$, let

$$J_\nu(t) = \left(\frac{t}{2}\right)^\nu \sum_{l=0}^{\infty} \frac{1}{l!\Gamma(\nu+l+1)}\left(\frac{it}{2}\right)^{2l}$$

be the Bessel function of order ν. We put

$$\tilde{j}_k(t) = \Gamma(k + \frac{n+1}{2}) \left(\frac{t}{2}\right)^{-(k+\frac{n-1}{2})}$$

$$J_{k+\frac{n-1}{2}}(t) = \sum_{l=0}^{\infty} \frac{\Gamma(k + (n+1)/2)}{l!\Gamma(k + l + (n+1)/2)} \left(\frac{it}{2}\right)^{2l}.$$

Note that $\tilde{j}_k(0) = 1$ and $\tilde{j}_k(-t) = \tilde{j}_k(t)$. (See Lemma 5.13 in [6]).

By using the extended Legendre polynomials and the modified Bessel functions, the exponential function is represented as follows;

$$\exp(z \cdot w) = \sum_{k=0}^{\infty} \frac{\Gamma(\frac{n+1}{2})N(k,n)}{2^k \Gamma(k + \frac{n+1}{2})} \tilde{j}_k(i\sqrt{z^2}\sqrt{w^2}) P_{k,n}(z,w). \tag{10}$$

Therefore, by (9) and (10), for $\eta, \zeta \in \mathbf{C}^{n+1}$, we have

$$\int_{S^n} \exp(ix \cdot \eta) \exp(x \cdot \zeta) d\omega_{n+1}(x)$$
$$= \sum_{k=0}^{\infty} \frac{(\Gamma((n+1)/2))^2 N(k,n)}{2^{2k}\Gamma(k + \frac{n+1}{2})^2} \tilde{j}_k\left(\sqrt{\eta^2}\right) \tilde{j}_k\left(i\sqrt{\zeta^2}\right) P_{k,n}(\eta, \zeta).$$

On the other hands,

$$\int_{S^n} \exp(ix \cdot \eta) \exp(x \cdot \zeta) d\omega_{n+1}(x) = \int_{S^n} \exp(x \cdot (i\eta + \zeta))\, d\omega_{n+1}(x)$$
$$= \tilde{j}_0\left(i\sqrt{(\zeta + i\eta)^2}\right)$$
$$= \sum_{l=0}^{\infty} \frac{\Gamma((n+1)/2)}{l!\Gamma(l + (n+1)/2)} \left(\frac{(\zeta + i\eta)^2}{4}\right)^l$$
$$= \sum_{l=0}^{\infty} \frac{\Gamma((n+1)/2)}{l!\Gamma(l + (n+1)/2)} \left(\frac{\zeta^2 - \eta^2 + 2i\zeta \cdot \eta}{4}\right)^l.$$

For this calculation see [4] and [6], for example.

3.2 Expansion Formula

For $f \in L^2(S^n)$, define

$$f_k(z) = N(k,n) \int_{S^n} f(y) P_{k,n}(z,y) d\omega_{n+1}(y), \quad z \in \mathbf{C}^{n+1}.$$

Note that $\Delta_z f_k(z) = 0$. For $x \in S^n$ we have

$$f(x) = \sum_{k=0}^{\infty} f_k(x),$$

in the sense of $L^2(S^n)$. We remark that the infinite sum of the right hand side is a complex harmonic function and converges in the Lie ball

$$\tilde{B} = \left\{ z \in \mathbf{C}^{n+1} : \sqrt{\|z\|^2 + \sqrt{\|z\|^4 - |z^2|^2}} < 1 \right\}.$$

(See [6], for example.) Further we remark that

$$\sum_{k=0}^{\infty} N(k,n) P_{k,n}(z,w) = \frac{1 - z^2 w^2}{(1 - 2z \cdot w + z^2 w^2)^{(n+1)/2}}$$

is the Poisson kernel.

We consider the images of $f_z(x) = P_{k,n}(x, \overline{z})$, $z \in \mathbf{C}^{n+1}$ under the Fourier transformation, the windowed Fourier transformation and the Gabor wavelet transformation. Note that $\overline{P_{k,n}(z,w)} = P_{k,n}(\overline{z}, \overline{w})$. We recall $P_{k,n}(z,w) = P_{k,n}(w,z)$. By (9) and (10), we have

$$\begin{aligned}(\mathcal{F} f_z)(\omega) &= \int_{S^n} \sum_{l=0}^{\infty} \frac{\Gamma((n+1)/2) N(l,n)}{2^l \Gamma(l + \frac{n+1}{2})} \tilde{j}_l(i\sqrt{\omega^2}) P_{l,n}(x,\omega), \overline{P_{k,n}(x,z)} d\omega(x) \\ &= C_k \tilde{j}_k(i\sqrt{\omega^2}) P_{k,n}(\omega, z).\end{aligned}$$

where we put $C_k = \Gamma((n+1)/2)/(2^k \Gamma(k + (n+1)/2))$. Thus for $f \in L^2(S^n)$, we have

$$(\mathcal{F} f)(\omega) = \sum_{k=0}^{\infty} C_k \tilde{j}_k(i\sqrt{\omega^2}) f_k(\omega).$$

Therefore for $f \in L^2(S^n)$, we have

$$(\mathcal{W}_b\mathcal{F}f)(\tau,\omega) = e^{-\frac{1+\tau^2}{2b^2}} \sum_{k=0}^{\infty} C_k \tilde{j}_k \left(i\sqrt{(\omega + i\frac{\tau}{b^2})^2} \right) f_k \left(\omega + i\frac{\tau}{b^2} \right),$$

$$(\mathcal{G}_{\omega_0} f)(\tau,a) = \frac{e^{-\frac{1+\tau^2}{2a^2}} e^{i\tau\cdot\frac{\omega_0}{a}}}{a} \sum_{k=0}^{\infty} C_k \tilde{j}_k \left(i\sqrt{(\frac{\omega_0}{a} + i\frac{\tau}{a^2})^2} \right) f_k \left(\frac{\omega_0}{a} + i\frac{\tau}{a^2} \right).$$

4 Inverse Gabor Wavelet Transformation

In this section, to consider the inverse Gabor transformation on the sphere, we will consider the inverse Fourier transformation.

In [2], [4] and [5], we treated the inverse Gabor wavelet transformation on S^2.

For $z, w \in \mathbf{C}^{n+1}$, put

$$E(z,w) = \sum_{k=0}^{\infty} \frac{1}{2^k k! \tilde{j}_k(i\sqrt{z^2}\sqrt{w^2})} P_{k,n}(z,w).$$

For $0 < s < \infty$, let

$$K_\nu(s) = K_{-\nu}(s) = \int_0^\infty \exp(-s\cosh t)\cosh \nu t dt,$$

be the modified Bessel function. Put

$$\rho_r(s) = \begin{cases} \displaystyle\sum_{l=0}^{(n-1)/2} a_l r^{l+n+1} s^{l+1} K_k(rs), & n \text{ is odd}, \\ \displaystyle\sum_{l=0}^{n/2} a_l r^{l+n+1/2} s^{l+1} K_{l-1/2}(rs), & n \text{ is even}, \end{cases}$$

where the constants a_l are defined by

$$\int_0^\infty s^{2l+n-1}\rho_n(s)ds = \frac{N(l,n)l!\Gamma(l+(n+1)/2)2^{2k}}{\Gamma((n+1)/2)}, \quad l = 0,1,2,\cdots.$$

In [1] we define a measure $d\mu$ on $\mathbf{R}^{n+1}$ by

$$\int_{\mathbf{R}^{n+1}} f(x)d\mu(x) = \int_0^\infty \int_{S^n} f(s\omega)d\omega_{n+1}(\omega)s^{n-1}\rho(s)ds.$$

For $z \in S^n$, the mapping

$$F \mapsto \int_{\mathbf{R}^{n+1}} F(x)\overline{E(z,x)}d\mu(x) = \int_{\mathbf{R}^{n+1}} F(x)E(z,x)d\mu(x) \tag{11}$$

gives the inverse mapping of the Fourier transformation defined by (1).

By (5), (7) and (8), the inverse mapping of the Windowed Fourier transformation (4) and the inverse mapping of the Gabor wavelet transformation (6) can be constructed by using (11).

References

1. Fujita, K.: Hilbert spaces related to harmonic functions. Tôhoku Math. J. **48**, 149–163 (1996). http://doi.org/10.2748/tmj/1178225416
2. Fujita, K.: On an inverse transformation of the Gabor transformation on the sphere. In: Proceedings of the Seventh International Conference on Information, Information, pp. 45–48 (2015). ISBN 4-901329-08-1
3. Fujita, K.: Gabor transform of analytic functional on the sphere. Current Trends in Analysis and its Applications, Proceedings of the 9th ISAAC Congress, Kraków 2013, pp. 451–458. Springer International Publishing (2015). https://doi.org/10.1007/978-3-319-12577-0_50
4. Fujita, K.: Gabor transformation on the Sphere and its inverse transformation. New Trends in Analysis and Interdisciplinary Applications, Trends in Mathematics, Proceedings of the 10th ISAAC Congress, pp. 573–580. Springer International Publishing (2017). https://doi.org10.1007/978-3-319-48812-7_72
5. Fujita, K.: Some topics on the Gabor wavelet transformation. Current Trends in Analysis, its Applications and Computation, Proceedings of the 12th ISAAC Congress, Aveiro, Portugal, 2019. pp. 671–679 Springer Nature Switzerland AG. Part of Springer Nature (2021). https://doi.org/10.1007/978-3-030-87502-2
6. Morimoto, M.: Analytic Functionals on the Sphere. Translation of Mathematical Monographs, vol. 178. American Mathematical Society, Providence (1998)

On the Diameters and Radii of the Extended Sierpiński Graphs

Mai Fujita and Yoshiroh Machigashira

Abstract In this paper, the diameters and radii of the extended Sierpiński Graphs are discussed.

1 Introduction

Let G be a simple connected graph. We denote by $V(G)$ and $E(G)$ its vertex set and edge set, respectively. For any u and v in $V(G)$, the *distance* $d_G(u, v)$ between u and v, briefly $d(u, v)$, denotes the number of edges of a shortest path joining u with v. For a fixed $v \in V(G)$, we call the distance from v to the farthest vertex (or the vertices) the *eccentricity* of the vertex v and denote by $e_G(v)$, namely,

$$e_G(v) = \max_{u \in V(G)} d_G(u, v).$$

We define the *diameter* and *radius* of G by

$$\mathrm{diam}(G) = \max_{v \in V(G)} e_G(v) = \max_{u,v \in V(G)} d_G(u, v),$$

$$\mathrm{rad}(G) = \min_{v \in V(G)} e_G(v),$$

respectively.

In [1], the Sierpiński Graphs were introduced. Let $n, k \in \mathbb{N}$. We define the *Sierpiński Graphs* $S(n, k)$ as follows.

$$V\big(S(n, k)\big) = \{1, 2, \cdots, k\}^n.$$

M. Fujita · Y. Machigashira (✉)
Division of Math, Sciences, and Information Technology in Education, Osaka Kyoiku University, Kashiwara, Japan
e-mail: fujita-m81@cc.osaka-kyoiku.ac.jp; machi@cc.osaka-kyoiku.ac.jp

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_53

Instead of $v \in V(S(n,k))$ and $v = (v_1, v_2, \cdots, v_n)$, we simply write $v \in S(n,k)$ and $v = v_1 v_2 \cdots v_n$, respectively. Two different vertices $u = u_1 u_2 \cdots u_n$ and $v = v_1 v_2 \cdots v_n$ are adjacent if and only if there exists $\ell \in \{1, 2, \cdots, n\}$ such that

$$\begin{cases} u_j = v_j & \big(j \in \{1, \cdots, \ell - 1\}\big), \\ u_\ell \neq v_\ell, & \\ u_j = v_\ell \text{ and } u_\ell = v_j & \big(j \in \{\ell + 1, \cdots, n\}\big). \end{cases}$$

In [2], the diameters and radii of the Sierpiński Graphs were obtained.

Theorem 1 ([2, Corollary 2.2, Theorem 3.1]) *Let $n, k \in \mathbb{N}$. Then*

$$\operatorname{diam}(S(n,k)) = 2^n - 1, \tag{1}$$

$$\operatorname{rad}(S(n,k)) = \begin{cases} 2^n - 1 & (n < k), \\ 2^{n-1} + 2^{n-2} \cdots + 2^{n-(k-1)} & (n \geq k). \end{cases}$$

We call a vertex of $S(n,k)$ of the form i^n, $i \in \{1, 2, \cdots, k\}$, an *extreme vertex* of $S(n,k)$. Choosing i^n and $v = v_1 v_2 \cdots v_n$ satisfying $v_j \neq i$ for all $j \in \{1, 2, \cdots, k\}$, the diameter in $S(n,k)$ is attained. Especially, we can choose two different extreme vertices. Concerning about the radius, $n \geq k$, the eccentricities of vertices of the form $v = v_1 v_2 \cdots v_{k-1} v_k^{n-(k-1)}$ satisfying $\{v_1, v_2, \cdots, v_k\} = \{1, 2, \cdots, k\}$ attain the radius in $S(n,k)$ and the farthest vertices at that time are vertices v_k^n.

In [3, 4], the extended Sierpiński Graphs $S^+(n,k)$ and $S^{++}(n,k)$ were introduced. The *first typed extended Sierpiński Graphs* $S^+(n,k)$ are defined by adding a new vertex s, called the *special vertex* of $S^+(n,k)$, and all edges connecting s and all extreme vertices of $S(n,k)$, that is,

$$V\Big(S^+(n,k)\Big) = V\big(S(n,k)\big) \cup \{s\},$$

$$E\Big(S^+(n,k)\Big) = E\big(S(n,k)\big) \cup \big\{e_i : i \in \{1, 2, \cdots, k\}\big\},$$

where e_i are the edges linking s and the extreme vertices i^n. Also, we define the *second typed extended Sierpiński Graphs* $S^{++}(n,k)$ as follows.

$$V\Big(S^{++}(n,k)\Big) = V\big(S(n,k)\big) \cup V\big(S(n-1,k)\big),$$

$$E\Big(S^{++}(n,k)\Big) = E\big(S(n,k)\big) \cup E\big(S(n-1,k)\big) \cup \big\{\tilde{e_i} : i \in \{1, 2, \cdots, k\}\big\},$$

where $\tilde{e_i}$ are the edges connecting the extreme vertices $i^n \in S(n,k)$ and the extreme vertices $0i^{n-1} \in S(n-1,k)$, $v = 0v_2 \cdots v_n$ denote vertices in $S(n-1,k)$. In this paper, we investigate the diameters and radii of the first typed extended Sierpiński Graphs $S^+(n,k)$. From now on, we treat the case $k \geq 3$.

Theorem 2 *Let $n, k \in \mathbb{N}$ and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. Then*

$$\operatorname{diam}\left(S^+(n,3)\right) = \begin{cases} (5 \cdot 2^{n-1} - 1)/3 & (n = 2m, m \in \mathbb{N}), \\ (5 \cdot 2^{n-1} - 2)/3 & (n = 2m+1, m \in \mathbb{N}_0), \end{cases} \tag{2}$$

$$\operatorname{diam}\left(S^+(n,k)\right) = 2^n - 1 \qquad (k \geq 4), \tag{3}$$

$$\operatorname{rad}\left(S^+(n,k)\right) = 2^{n-1}. \tag{4}$$

Note that the special vertex s is the unique vertex which attains the radius of $S^+(n,k)$. This will be shown in the proof of (4). Also, we show the diameters and radii of the second typed extended Sierpiński Graphs $S^{++}(n,k)$.

Theorem 3 *Let $n, k \in \mathbb{N}$. Then*

$$\operatorname{diam}\left(S^{++}(n,k)\right) = 2^n - 1, \tag{5}$$

$$\operatorname{rad}\left(S^{++}(n,k)\right) = 2^{n-1} + 2^{n-2} + \cdots + 2^{n-k} \qquad (n \geq k). \tag{6}$$

In the case $k = 3$, for example, two vertices 1^n and $23^2 2^{n-3}$ no longer attain the diameter in $S^{++}(n,3)$.

2 Preliminaries

In this section, we collect lemmas which will be used later on.

Lemma 1 ([1, Lemma 4]) *Let $n, k \in \mathbb{N}$, $v = v_1 v_2 \cdots v_n \in S(n,k)$ and $i \in \{1, 2, \cdots, k\}$. Then*

$$d_{S(n,k)}(v, i^n) = \langle \rho_{v_1,i}\ \rho_{v_2,i}\ \cdots\ \rho_{v_n,i} \rangle_2, \tag{7}$$

where

$$\rho_{j,j'} := \begin{cases} 1 & (j \neq j'), \\ 0 & (j = j') \end{cases}$$

and the right hand side of (7) is a binary number, that is to say,

$$\langle a_1 a_2 \cdots a_n \rangle_2 = \sum_{k=1}^{n} a_k 2^{n-k}.$$

Lemma 2 ([2, Corollary 2.2 (i)]) *Let $n, k \in \mathbb{N}$, $n \geq 2$, $u = iu_2 \cdots u_n$, $v = iv_2 \cdots v_n \in S(n,k)$, where $i \in \{1, 2, \cdots, k\}$. Then*

$$d_{S(n,k)}(iu_2 \cdots u_n, iv_2 \cdots v_n) = d_{S(n-1,k)}(u_2 \cdots u_n, v_2 \cdots v_n).$$

3 Proof of Theorem 2

In this section, we prove Theorem 2. Firstly, we investigate (2). Let $u = u_1u_2 \cdots u_n$, $v = v_1v_2 \cdots v_n$ be arbitrary vertices of $S^+(n,3)$. We shall prove that $d_{S^+(n,3)}(u, v)$ is less than or equal to the right hand side of (2). We distinguish two cases.

Case 1: $u_1 = v_1$. It is sufficient to find a path linking u and v satisfying its length is less than or equal to the right hand side of (2). It follows immediately from Lemma 2 and (1) that the direct path is the desired one.

Case 2: $u_1 \neq v_1$. Without loss of generality, we may assume that $u_1 = 1$, $v_1 = 2$. Concerning paths connecting u and v, we can consider three cases. One is the direct path denoted by P_1, the second one is the path via the special vertex s denoted by P_2 and the third one is the path containing 13^{n-1} denoted by P_3. We denote by $L(P)$ the length of the path P. By Lemmas 2 and 1, we see that

$$L(P_1) + L(P_2) + L(P_3) = 5 \cdot 2^{n-1},$$

where we have used the fact that $\sum_{i=1}^{k} d_{S(n,k)}(v, i^n) = (k-1)(2^n - 1)$ for all vertices v of $S(n,k)$. Considering the remainder of dividing $5 \cdot 2^\ell$ by 3, $\ell \in \mathbb{N}$, it follows that at least one of $L(P_1)$, $L(P_2)$ and $L(P_3)$, therefore, $d_{S^+(n,3)}(u, v)$ is less than or equal to the right hand side of (2). Finally, we choose two vertices (u, v) with the shortest paths connecting u and v as follows.

$$\begin{cases} (u, v) = (131313 \cdots 13, 231313 \cdots 13) & \text{with } P_1, P_3 \quad (n = 2m, m \in \mathbb{N}), \\ (u, v) = (131313 \cdots 131, 231313 \cdots 131) & \text{with } P_1 \quad (n = 2m+1, m \in \mathbb{N}_0). \end{cases}$$

Thus, the diameters are attained and this completes the proof of (2).

Secondly, we consider (3). For all vertices u and v of $S^+(n,k)$, since $d_{S^+(n,k)}(u, v)$ denotes the number of edges of a shortest path in $S^+(n,k)$, clearly we have

$$d_{S^+(n,k)}(u, v) \leq d_{S(n,k)}(u, v) \leq 2^n - 1.$$

Choosing two vertices $(u, v) = (12^{n-1}, 34^{n-1})$ with the direct path, the diameter is attained and this proves (3).

Thirdly, we show (4). We consider the eccentricities of vertices. Note that

$$e_{S^+(n,k)}(s) = 2^{n-1}. \tag{8}$$

This is a consequence of the fact that the farthest vertices from s are the vertices $v = v_1v_2\cdots v_n$ satisfying $v_i \neq v_1$ for all $i \in \{2, \cdots, n\}$ with the path through v_1^n. We shall prove that for all vertices v of $S^+(n, k)$ except the special vertex s,

$$e_{S^+(n,k)}(v) > 2^{n-1}.$$

Once this is proved, combining this and (8), we obtain the conclusion. Let $v = v_1v_2\cdots v_n$ be an arbitrary vertex of $S^+(n, k)$ except the special vertex s. It is sufficient to show that there exists a vertex u such that

$$d_{S^+(n,k)}(u, v) > 2^{n-1},$$

where $d_{S^+(n,k)}(u, v)$ denotes the number of edges of a shortest path joining u with v in $S^+(n, k)$. Without loss of generality, we may assume that $v_1 = 1$. We choose the vertices u depends on v as follows.

$$\begin{cases} u = 23^{n-1} & (v_2 \in \{1, 3, 4, \cdots, k\}), \\ u = 32^{n-1} & (v_2 = 2). \end{cases}$$

Therefore, we can check that lengths of all paths joining u with v are strictly greater than 2^{n-1}. We end the proof of (4).

4 Proof of Theorem 3

In this section, we prove Theorem 3. Firstly, we investigate (5). For all vertices u and v of $S^{++}(n, k)$, by the same reason for $S^+(n, k)$, clearly we have

$$d_{S^{++}(n,k)}(u, v) \leq 2^n - 1.$$

Choosing two vertices $(u, v) = (12^{n-1}, k^n)$ with the direct path, the diameter is attained. This proves (5).

Secondly, we consider (6). We shall prove that for all vertices v of $S^{++}(n, k)$

$$e_{S^{++}(n,k)}(v) \geq 2^{n-1} + 2^{n-2} + \cdots + 2^{n-k}. \tag{9}$$

Let $v = v_1 v_2 \cdots v_n$ be an arbitrary vertex of $S^{++}(n,k)$. It is sufficient to show that there exists a vertex u such that

$$d_{S^{++}(n,k)}(u,v) \geq 2^{n-1} + 2^{n-2} + \cdots + 2^{n-k}. \tag{10}$$

To do this, we denote by $\#(A)$ the number of elements of the set A. We choose the vertices u depending on v with several paths need to be mentioned as follows. Consequently, we can show that lengths of all paths linking u and v are greater than or equal to the right hand side of (10).

The case for $v_1 \neq 0$. Without loss of generality, we may assume that $v_1 = 1$. We distinguish several cases.

Case 1: $\#(\{1,2,\cdots,k\} \setminus \{1, v_2, \cdots, v_k\}) \geq 2$. We can take $\alpha \neq \beta \in \{1,2,\cdots,k\} \setminus \{1, v_2, \cdots, v_k\}$ and set $u = \alpha\beta^{n-1}$. For the paths via $1j^{n-1}$, $j \in \{1,2,\cdots,k\} \setminus \{\alpha, \beta\}$, it is easy to see that their lengths are greater than or equal to 2^n and for the paths through $1j^{n-1}$, $j \in \{\alpha, \beta\}$, clearly we see that their lengths are greater than or equal to the right hand side of (10).

Case 2: $\#(\{1,2,\cdots,k\} \setminus \{1, v_2, \cdots, v_k\}) = 1$. We can take $\alpha \in \{1,2,\cdots,k\} \setminus \{1, v_2, \cdots, v_k\}$. When $1 \notin \{v_2, \cdots, v_k\}$, set $u = 0\alpha^{n-1}$. For the paths containing $1j^{n-1}$, $j \in \{2,\cdots,k\} \setminus \{\alpha\}$, it is easily checked that their lengths are strictly greater than 2^n. When $1 \in \{v_2, \cdots, v_k\}$, there only exists $\ell \in \{2,\cdots,k\}$ such that $v_\ell = 1$. Set $u = \alpha^{\ell-1}\beta\alpha^{n-\ell}$, where $\beta \neq 1$, $\beta \neq \alpha$. Note that for the path through 1^n, $d_{S(n-1,k)}(v_2 \cdots v_n, 1^{n-1})$ contains $2^{n-2} + 2^{n-3} + \cdots + 2^{n-k}$ except $2^{n-\ell}$ and $d_{S(n,k)}(\alpha^n, u)$ coincides $2^{n-\ell}$. Moreover, for the paths via $1j^{n-1}$, $j \in \{2,\cdots,k\} \setminus \{\alpha\}$, $d_{S(n-1,k)}(v_2 \cdots v_n, j^{n-1})$ contains $2^{n-2}+2^{n-3}+\cdots+2^{n-k}$ except $2^{n-\ell'}$, where $\ell' \in \{2,\cdots,k\} \setminus \{\ell\}$ is the unique number satisfies $v_{\ell'} = j$ and $d_{S(n,k)}(\alpha j^{n-1}, u)$ contains $2^{n-2} + 2^{n-3} + \cdots + 2^{n-k}$ except $2^{n-\ell}$, where $\ell \in \{2,\cdots,k\}$.

Case 3: $\#(\{1,2,\cdots,k\} \setminus \{1, v_2, \cdots, v_k\}) = 0$. Set $u = 0u_2 \cdots u_n$ satisfying $u_i \neq 1$ for all $i \in \{2,\cdots,n\}$ and $u_j \neq v_j$ for all $j \in \{2,\cdots,k\}$. It should be remarked that for the paths containing $1j^{n-1}$, $j \in \{2,\cdots,k\}$, there only exists $\ell \in \{2,\cdots,k\}$ such that $v_\ell = j$ and $u_\ell \neq j$. Furthermore, $d_{S(n-1,k)}(v_2 \cdots v_n, j^{n-1})$ contains $2^{n-2} + \cdots + 2^{n-k}$ except $2^{n-\ell}$ and $d_{S(n-1,k)}(j^{n-1}, u_2 \cdots u_n)$ contains $2^{n-\ell}$.

The case for $v_1 = 0$. We distinguish two cases.

Case 1: $\#(\{1,2,\cdots,k\} \setminus \{v_2, v_3, \cdots, v_k\}) \geq 2$. The proof of this case is similar to that of Case 1 for $v_1 \neq 0$ and we omit it.

Case 2: $\#(\{1,2,\cdots,k\} \setminus \{v_2, v_3, \cdots, v_k\}) = 1$. We can take $\alpha \in \{1,2,\cdots,k\} \setminus \{v_2, v_3, \cdots, v_k\}$. Set $u = \alpha u_2 \cdots u_n$ satisfying $u_i \neq \alpha$ for all $i \in \{2,\cdots,n\}$ and $u_j \neq v_j$ for all $j \in \{2,\cdots,k\}$. When $1 \notin \{v_2, \cdots, v_k\}$, this means $\alpha = 1$, the rest of the proof of this case is similar to that of Case 3 for $v_1 \neq 0$. When $1 \in \{v_2, \cdots, v_k\}$, this means $\alpha \neq 1$, the rest of the proof of this case is similar to that of Case 2 for $v_1 \neq 0$.

Finally, choosing the vertex $v = 012\cdots(k-1)k^{n-k}$, it follows that $e_{S^{++}(n,k)}(v)$ coincides the right hand side of (9), where the farthest vertex from v is $k(k-1)^{n-1}$ with the paths via $0k^{n-1}$ and $0(k-1)^{n-1}$. This completes the proof of (6).

Acknowledgments The second author thanks Takuto Imai for useful discussion.

References

1. Klavzar, S., Milutinovic, U.: Graphs $S(n,k)$ and a variant of the tower of Hanoi problem. Czechoslovak Math. J. **47**(122), 95–104 (1997)
2. Parisse, D.: On some metric properties of the Sierpiński graphs $S(n,k)$. Ars Combin. **90**, 145–160 (2009)
3. Klavzar, S., Mohar, B.: Crossing number of Sierpiński-like graphs. J. Graph Theory **50**, 186–198 (2005)
4. Lin, C-H., Liu, J-J., Wang, Y-L., Yen, W-C-K.: The hub number of Sierpiński-like graphs. Theory Comput. Syst. **49**, 588–600 (2011)

Some Inequalities for Parseval Frames

Takeshi Mandai, Ryuichi Ashino, and Akira Morimoto

Abstract Let $F = \{f_k\}_{k\in K}$ be a Parseval frame in a Hilbert space H, that is, $\|x\|^2 = \sum_{k\in K} |\langle x, f_k\rangle|^2$ holds for all $x \in H$. It is well known that if the norm $\|f_{k_0}\| = 1$, then $f_{k_0} \perp f_k$ for all $k \neq k_0$. In general, we might expect that if $\|f_{k_0}\|$ is close to 1, then the angles between other f_k's are close to $\pi/2$. We want to make it clear by some inequalities. In fact, we can prove several inequalities. The most typical one is

$$\frac{|\langle f_k, f_l\rangle|}{\|f_k\| \cdot \|f_l\|} \leq \frac{\sqrt{1-\|f_k\|^2}}{\|f_k\|} \cdot \frac{\sqrt{1-\|f_l\|^2}}{\|f_l\|}$$

for $k \neq l$. The meaning of the inequalities and some related topics will be given.

1 Introduction

Frame theory for Hilbert space, initiated by Duffin and Schaffer [4], has been widely used in various areas like signal analysis, especially wavelet theory [3, 5]. Parseval frames, also called normalized tight frames, generalize orthonormal bases. They have the same reconstruction formula as orthonormal bases [1, 2, 6].

We are interested in a configuration of vectors in a Parseval frame. We give several inequalities estimating the angle between two vectors in a Parseval frame by means of their norms, without proofs. The proofs will be given elsewhere.

T. Mandai (✉)
Osaka-Electro-Communication University, Osaka, Japan
e-mail: mandai@osakac.ac.jp

R. Ashino · A. Morimoto
Osaka Kyoiku University, Kashiwara, Japan
e-mail: ashino@cc.osaka-kyoiku.ac.jp; morimoto@cc.osaka-kyoiku.ac.jp

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_54

2 Parseval Frames

Let H be a Hilbert space over $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$. Let K be an index set like $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{N} \times \mathbb{Z}$ and so on, and $F = (f_k)_{k \in K}$ be a sequence of $f_k \in H$. The cardinality of K is denoted by $|K|$. A typical Hilbert space is

$$\ell^2(L) := \left\{ (c_l)_{l \in L} \,\middle|\, c_l \in \mathbb{K},\ \sum\nolimits_{l \in L} |c_l|^2 < \infty \right\}$$

for an index set L with inner product $\langle (c_l), (d_l) \rangle = \sum_{l \in L} c_l \overline{d_l}$ and norm $\|(c_l)\| = \sqrt{\sum_{l \in L} |c_l|^2}$.

Definition 1

(i) The sequence F is called a *Bessel sequence* for H, if there exists a constant $B > 0$ such that

$$\sum_{k \in K} |\langle f, f_k \rangle|^2 \le B \|f\|^2 \quad \text{for every } f \in H.$$

(ii) The sequence F is called a *frame* for H, if there exist two constants $A, B > 0$ such that

$$A \|f\|^2 \le \sum_{k \in K} |\langle f, f_k \rangle|^2 \le B \|f\|^2 \quad \text{for every } f \in H.$$

(iii) The sequence F is called a *Parseval frame*, or *normalized tight frame*, for H, if

$$\|f\|^2 = \sum_{k \in K} |\langle f, f_k \rangle|^2 \quad \text{for every } f \in H. \tag{1}$$

An orthonormal basis is a Parseval frame, but there are many Parseval frames which are not orthogonal systems.

Let F be a Bessel sequence for H. The *analysis operator* of F is a bounded operator $R_0 = R_0[F] : H \to \ell^2(K)$, defined by

$$R_0 f := (\langle f, f_k \rangle)_{k \in K} \in \ell^2(K) \quad \text{for } f \in H.$$

The *synthesis operator* is $R_0^* = R_0[F]^* : \ell^2(K) \to H$. We have

$$R_0^*(c_k)_k = \sum_{k \in K} c_k f_k \in H \quad \text{for } (c_k)_{k \in K} \in \ell^2(K).$$

The *frame operator* is $S = S[F] = R_0[F]^* R_0[F] : H \to H$. We have

$$Sf = \sum_{k \in K} \langle f, f_k \rangle f_k \in H \quad \text{for } f \in H.$$

Since

$$\langle Sf, f \rangle = \sum_{k \in K} |\langle f, f_k \rangle|^2, \tag{2}$$

S is a non-negative definite self-adjoint operator. If F is a frame, S is positive-definite.

2.1 Important Properties

We summarize some important properties of Parseval frames [1, 2, 6]. Let $F = (f_k)_{k \in H}$ be a Parseval frame for H.

1. The analysis operator $R_0[F]$ is an isometry, that is

$$\|R_0 f\| = \|f\| \quad \text{for every } f \in H.$$

$$\sum_{k \in K} |\langle f, f_k \rangle|^2 = \|f\| \quad \text{for every } f \in H.$$

   An isometry preserves inner products as well.

$$\langle R_0 f, R_0 g \rangle = \langle f, g \rangle \quad \text{for every } f, g \in H.$$

$$\sum_{k \in K} \langle f, f_k \rangle \overline{\langle g, f_k \rangle} = \langle f, g \rangle \quad \text{for every } f, g \in H.$$

2. We have a good reconstruction formula:

$$f = \sum_{k \in K} \langle f, f_k \rangle f_k \quad \text{for all } f \in H.$$

   In other words, $S[F] = Id_H$.
3. Let $H_1 (\subset H)$ be a closed subspace of H, and P_{H_1} be the orthogonal projection onto H_1. Then, $P_{H_1}(F) := (P_{H_1}(f_k))_{k \in H}$ is a Parseval frame for H_1.
4. Every Parseval frame is an orthogonal projection of an orthonormal basis. That is, there exist a Hilbert space $\widetilde{H} \supset H$ and its orthonormal basis $E = (e_k)_{k \in K}$ such that $f_k = P_H(e_k)$.

5. Let $H = \ell^2(L)$ and $F = (f_k)_{k\in K}$, $f_k = (f_{k,l})_{l\in L} \in \ell^2(L)$. Then, $\big((f_{k,l})_{k\in K}\big)_{l\in L}$ is an orthonormal system of $\ell^2(K)$. That is, $\sum_{k\in K} f_{k,l}\,\overline{f_{k,l'}} = \delta_{l,l'}$ for $l, l' \in L$, where $\delta_{l,l'}$ is Kronecker's delta.

   This is also a sufficient condition for F to be a Parseval frame.
6. The sum of the squares of norms depends only on the dimension of H: $\sum_{k\in K} \|f_k\|^2 = \dim H$.
7. If $\dim \mathrm{Span}\{f_1, \ldots, f_s\} = 1$, then we can replace $f_1, \ldots, f_s$ by $f_* := \sqrt{\sum_{j=1}^s \|f_j\|^2}\, e$, preserving that F is a Parseval frame, where $e \in \mathrm{Span}\{f_1, \ldots, f_s\}$, $\|e\| = 1$. Here, $\mathrm{Span} V$ denotes the subspace spanned by the vectors in V.

From a frame we can make a Parseval frame in the following manner. If F is a frame for H, then its frame operator $S = S[F]$ is positive-definite, and $S^{-1/2}F := (S^{-1/2}f_k)_{k\in K}$ is a Parseval frame. In fact, by (2) we have

$$\begin{aligned}\sum_k |\langle f, S^{-1/2}f_k\rangle|^2 &= \sum_k |\langle S^{-1/2}f, f_k\rangle|^2 = \langle SS^{-1/2}f, S^{-1/2}f\rangle \\ &= \langle f, f\rangle = \|f\|^2 .\end{aligned} \tag{3}$$

In other words, if we replace the inner product $\langle\ ,\ \rangle$ of H by new inner product $\langle\ ,\ \rangle_S$ defined by

$$\langle f, g\rangle_S := \langle S^{-1/2}f, S^{-1/2}g\rangle = \langle S^{-1}f, g\rangle,$$

then F is a Parseval frame for the new Hilbert space H equipped with this inner product $\langle\ ,\ \rangle_S$ and the norm $\|f\|_S = \sqrt{\langle f, f\rangle_S} = \left\|S^{-1/2}f\right\|$. In fact, by (3)

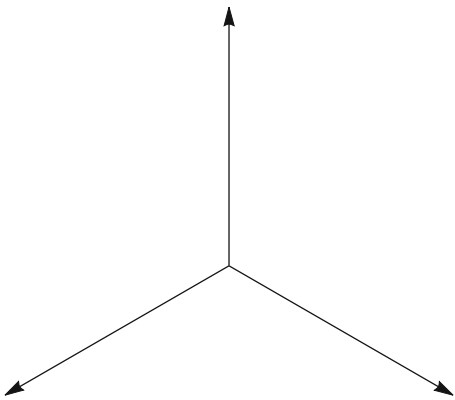
$$\sum_{k\in K} |\langle f, f_k\rangle_S|^2 = \sum_{k\in K} |\langle S^{-1/2}f, S^{-1/2}f_k\rangle|^2 = \left\|S^{-1/2}f\right\|^2 = \|f\|_S^2 .$$

2.2 *Interesting Examples*

In this section, we give several interesting examples of Parseval frames.

1. Let $F_j = (e_k^{(j)})_{k\in K_j}$ be orthonormal bases (or Parseval frames) for H ($j \in J$) and $\sum_{j\in J} |a_j|^2 = 1$. Then $\bigcup_{j\in J}(a_j e_k^{(j)})_{k\in K_j}$ is a Parseval frame. Especially, if $|J| < \infty$, then we can take $a_j = 1/\sqrt{|J|}$.
2. Let $(e_k)_{k\in K}$ be an orthonormal basis (or a Parseval frame) for $\widetilde{H} = L^2(\Omega_0)$, where $\Omega_0 \subset \mathbb{R}^n$. If $\Omega \subset \Omega_0$, then $(e_k|_\Omega)_{k\in K}$ is a Parseval frame for $H = L^2(\Omega)$.

Fig. 1 Mercedes Benz frame



3. In $\mathbb{R}^2$, let

$$f_1 = \begin{pmatrix} 0 \\ \frac{\sqrt{2}}{\sqrt{3}} \end{pmatrix}, \quad f_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{6}} \end{pmatrix}, \quad f_3 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{6}} \end{pmatrix}.$$

   Then, (f_1, f_2, f_3) is a Parseval frame for $\mathbb{R}^2$ (Fig. 1). This is called Mercedes Benz frame.
4. As for wavelet frames, we have the following theorem.
   Let $\alpha \in \mathbb{R},\ \alpha > 1$. For $f \in L^2(\mathbb{R}^n),\ j \in \mathbb{Z},\ u \in \mathbb{R}^n$, set

$$f_{j,u}(x) := (D_{\alpha^j} T_u f)(x) = \alpha^{jn/2} f(\alpha^j x - u).$$

   Let L be a finite index set, and $p_\ell > 0\ (\ell \in L)$.

Theorem 1 *Let Q_ℓ be a cube in $\mathbb{R}^n$ with the sides of length $\frac{2\pi}{p_\ell}$ $(\ell \in L)$. For $\psi^\ell \in L^2(\mathbb{R}^n)$ $(\ell \in L)$, if*

$$\operatorname{supp} \widehat{\psi^\ell} \subset Q_\ell, \quad \ell \in L,$$

$$\sum_{\ell \in L} \frac{1}{p_\ell^n} \sum_{j \in \mathbb{Z}} |\widehat{\psi^\ell}(\alpha^{-j}\xi)|^2 = 1 \qquad \text{for } 1 \le \xi \le \alpha,$$

where $\widehat{\psi}(\xi)$ is a Fourier transform of ψ, then $\left(\psi^\ell_{j,kp_\ell}\right)_{\ell \in L; j \in \mathbb{Z}, k \in \mathbb{Z}^n}$ is a Parseval frame for $L^2(\mathbb{R}^n)$.

3 What Do We Want to Know?

Let $F = (f_k)_{k\in K}$ be a Parseval frame. It is well known that if $\|f_{k_0}\| = 1$, then $f_k \perp f_{k_0}$ for all $k \neq k_0$. Hence, we might expect:

If $\|f_{k_0}\|$ is "close" to 1, then the angle θ_{k,k_0} is "close" to $\dfrac{\pi}{2}$,
where $\theta_{k,k_0} \in [0, \pi]$ is determined by $\cos\theta_{k,k_0} = \dfrac{\Re\langle f_k, f_{k_0}\rangle}{\|f_k\|\,\|f_{k_0}\|}$.

Here, $\Re z$ denotes the real part of $z \in \mathbb{C}$.

However, we can show that in case of $\|f_k\|^2 + \|f_{k_0}\|^2 \leq 1$, the angle θ_{k,k_0} can be arbitrary.

Theorem 2 *For every $\theta \in [0, \pi]$, and for every $(t, s) \in (0, 1)^2$ satisfying $t^2 + s^2 \leq 1$, there exist a Parseval frame F for $\mathbb{R}^2$, and $f_1, f_2 \in F$ such that*

$$\frac{\langle f_1, f_2\rangle}{\|f_1\|\,\|f_2\|} = \cos\theta, \quad \|f_1\| = t, \quad \|f_2\| = s.$$

How about the case $\|f_k\|^2 + \|f_{k_0}\|^2 > 1$? From the definition of Parseval frame, we can easily show the following inequality.

Proposition 1 *Let $|K| \geq 2$ and $f_k \neq 0$ for all $k \in K$. If $k \neq l$, then*

$$\frac{|\langle f_k, f_l\rangle|}{\|f_k\|\,\|f_l\|} \leq \frac{\sqrt{1 - \|f_k\|^2}}{\|f_l\|}. \tag{4}$$

In other words, if $\|f_k\|^2 + \|f_l\|^2 > 1$, then

$$\frac{|\langle f_k, f_l\rangle|}{\|f_k\|\,\|f_l\|} \leq \min\left\{\frac{\sqrt{1 - \|f_k\|^2}}{\|f_l\|}, \frac{\sqrt{1 - \|f_l\|^2}}{\|f_k\|}\right\} \tag{5}$$

$$= \frac{\sqrt{1 - \max\{\|f_k\|, \|f_l\|\}^2}}{\min\{\|f_k\|, \|f_l\|\}}. \tag{6}$$

The equality in (5) holds, if
(i) $f_k \perp f_j$ *for all* $j \neq k, l$, *or* (ii) $f_l \perp f_j$ *for all* $j \neq k, l$.

Note that $\dfrac{\sqrt{1 - \|f_k\|^2}}{\|f_l\|} < 1$ if and only if $\|f_k\|^2 + \|f_l\|^2 > 1$. If the right hand side of (4) is greater than 1, the inequality is meaningless since it is weaker than the

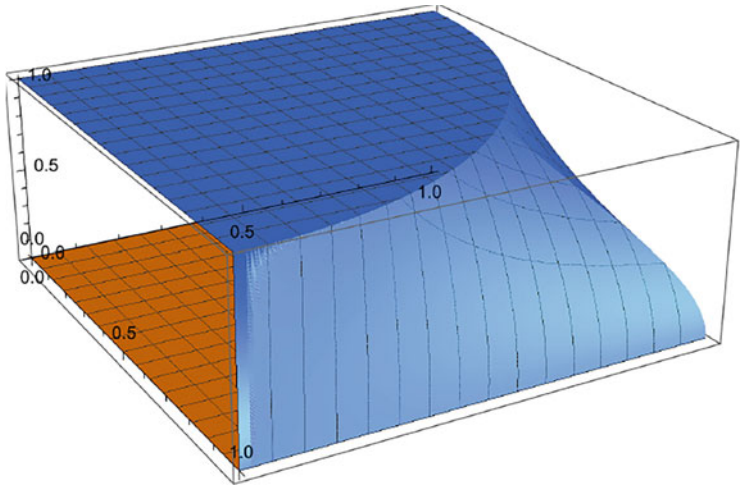


Fig. 2 Graph of $\min\left\{1, \dfrac{\sqrt{1-\max\{x,y\}^2}}{\min\{x,y\}}\right\}$

Schwarz inequality. Similarly, $\dfrac{\sqrt{1-\max\{\|f_k\|, \|f_l\|\}^2}}{\min\{\|f_k\|, \|f_l\|\}} < 1$ if and only if $\|f_k\|^2 + \|f_l\|^2 > 1$. (See Fig. 2.)

The equality in (6) follows from the following lemma.

Lemma 1 *If* $x, y \in (0, 1]$ *and* $x^2 + y^2 > 1$, *then*

$$\min\left\{\frac{\sqrt{1-x^2}}{y}, \frac{\sqrt{1-y^2}}{x}\right\} = \frac{\sqrt{1-\max\{x,y\}^2}}{\min\{x,y\}}. \tag{7}$$

Proof Since

$$\frac{1-x^2}{y^2} - \frac{1-y^2}{x^2} = \frac{(y-x)(x+y)(x^2+y^2-1)}{x^2y^2},$$

if $x > 0,\ y > 0,\ x^2 + y^2 > 1$, then

$$\frac{1-x^2}{y^2} \le \frac{1-y^2}{x^2} \iff y \le x.$$

Hence, if $y \le x$, then the both sides of (7) are $\dfrac{\sqrt{1-x^2}}{y}$, and if $x \le y$, then the both sides are $\dfrac{\sqrt{1-y^2}}{x}$. □

The inequality (4) is obtained by dropping many terms of the equality (1), and the conditions (i) and (ii) for the equality seem too strong. For example, Mercedes Benz frame does not satisfy the equality, since $\frac{1}{2} < \frac{1}{\sqrt{2}}$. We want stronger inequalities.

4 Main Result

In this section, we give a stronger inequality, which is best possible in the sense explained in the next section.

Theorem 3 *Let $F = (f_k)_{k\in K}$ be a Parseval frame. If $k \neq l$, then*

$$|\langle f_k, f_l\rangle| \leq \sqrt{1 - \|f_k\|^2}\sqrt{1 - \|f_l\|^2}, \tag{8}$$

$$\frac{|\langle f_k, f_l\rangle|}{\|f_k\|\,\|f_l\|} \leq \frac{\sqrt{1 - \|f_k\|^2}\sqrt{1 - \|f_l\|^2}}{\|f_k\|\,\|f_l\|}. \tag{9}$$

Mercedes Benz frame satisfies the equality $\left(\frac{1}{2} = \frac{1}{2}\right)$.

Note that the right-hand side $\dfrac{\sqrt{1 - \|f_k\|^2}\sqrt{1 - \|f_l\|^2}}{\|f_k\|\,\|f_l\|} < 1$ if and only if $\|f_k\|^2 + \|f_l\|^2 > 1$ (Fig. 3).

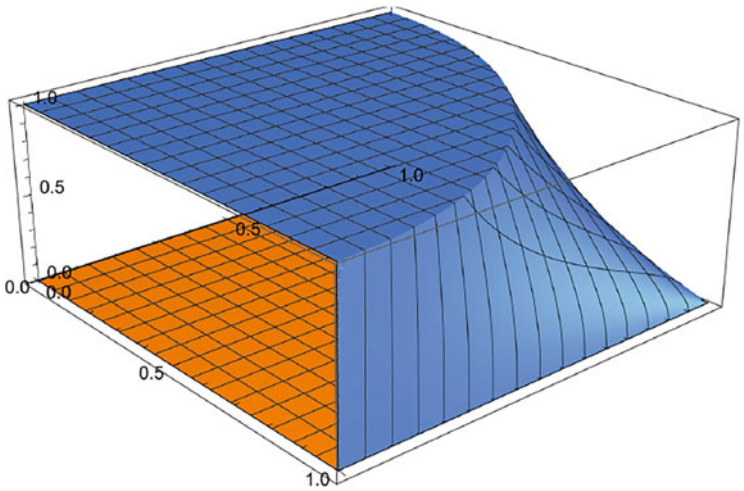


Fig. 3 Graph of $\min\left\{1, \dfrac{\sqrt{1-x^2}\sqrt{1-y^2}}{xy}\right\}$

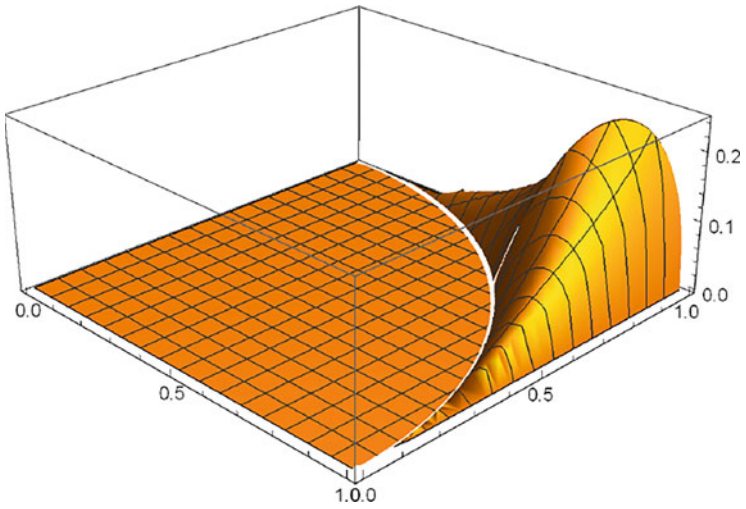


Fig. 4 Graph of $\min\left\{1,\ \dfrac{\sqrt{1-\max\{x,\,y\}^2}}{\min\{x,\,y\}}\right\} - \min\left\{1,\ \dfrac{\sqrt{1-x^2}\sqrt{1-y^2}}{xy}\right\}$

The inequality (9) is stronger than (5) if $\|f_k\|^2 + \|f_l\|^2 > 1$, since

$$\frac{\sqrt{1-\max\{x,\,y\}^2}}{\min\{x,\,y\}} > \frac{\sqrt{1-x^2}\sqrt{1-y^2}}{xy} \iff x^2 + y^2 > 1.$$

See Fig. 4. The maximum is $\dfrac{1}{4}$ at $(x,\,y) = (\dfrac{2}{\sqrt{5}}, \dfrac{2}{\sqrt{5}})$.

The inequality (9) is quantitatively explaining the following. If $\|f_k\|$ is close to 1, then $\cos\theta_{k,l}$ is close to 0, that is, $\theta_{k,l}$ is close to $\dfrac{\pi}{2}$.

Next, we consider when equality holds. The condition for the equality in (9) is

$$(E)_{(k,l)} \quad |\langle f_k,\, f_l\rangle|^2 = (1 - \|f_k\|^2)(1 - \|f_l\|^2).$$

Theorem 4

(1) $(E)_{(k,l)}$ *for all* $k \neq l$ *holds if and only if* $\operatorname{codim}\operatorname{Range} R_0[F] \leq 1$, *where* $\operatorname{Range} P$ *is the range of a linear operator* P, *and* $\operatorname{codim} V$ *is the codimension of a subspace* V.
Especially, if $\dim H < \infty$ *and* $|K| = \dim H + 1$, *then the equality holds for all* $k \neq l$. (*The Mercedes Benz frame is an example.*)

(2) *let* $k \neq l$ *be fixed.*

(i) *When* $\dim Span\{f_k,\, f_l\} = 1$ (*that is,* $\theta_{k,l} = 0$ *or* π),
$(E)_{(k,l)}$ *holds if and only if* $f_j \perp f_k,\, f_l$ *for all* $j \neq k, l$.
(*This implies* $\|f_k\|^2 + \|f_l\|^2 = 1$.)

(ii) *When* $\dim Span\{f_k, f_l\} = 2$,
$(E)_{(k,l)}$ *holds if and only if* $\dim Span\{P_{Span\{f_k,f_l\}} f_j \mid j \neq k, l\} \leq 1$.

5 Best Possibility

Inequality (8) is the best inequality in the following sense.

Theorem 5 *Let* $\dim H \geq 2$. *For every* $f, g \in H$ *satisfying*

$$|\langle f, g\rangle| \leq \sqrt{1 - \|f\|^2}\sqrt{1 - \|g\|^2},$$

there exists a Parseval frame $F = (f_k)_{k\in K}$ *for* H *such that* $f = f_k$, $g = f_l$ *for some* $k \neq l \in K$.

6 Another Type of Inequality

We have estimated the configuration by the angles or cosines of them. We can also consider the distance of two vectors.

Let $F = (f_k)_{k\in K}$ be a Parseval frame for H. If $a \in \mathbb{K}$ and $|a| = 1$, then we can replace f_l by af_l preserving that F is a Parseval frame. We should consider af_l with $|a| = 1$ is an "equivalent" vector to f_l. Note that

$$\min_{a\in\mathbb{K},|a|=1} \|f_k - af_l\|^2 = \|f_k\|^2 + \|f_l\|^2 - 2|\langle f_k, f_l\rangle|,$$

in general, since

$$\begin{aligned}\|f_k - af_l\|^2 &= \|f_k\|^2 + |a|^2\|f_l\|^2 - a\overline{\langle f_k, f_l\rangle} - \overline{a}\langle f_k, f_l\rangle \\ &= \|f_k\|^2 + \|f_l\|^2 - 2\Re a\overline{\langle f_k, f_l\rangle} \\ &\geq \|f_k\|^2 + \|f_l\|^2 - 2|\langle f_k, f_l\rangle|, \qquad (10)\end{aligned}$$

and the equality in (10) is attained if $a\overline{\langle f_k, f_l\rangle} > 0$.

Theorem 6 *Let* $k \neq l$. *For* $a \in \mathbb{K}$ *with* $|a| = 1$, *we have*

$$\|f_k - af_l\|^2 \geq \|f_k\|^2 + \|f_l\|^2 - 2\sqrt{1 - \|f_k\|^2}\sqrt{1 - \|f_l\|^2}.$$

If $\|f_k\|$ is close to 1, then the distance between f_k and af_l is close to $\|f_k\|^2 + \|f_l\|^2$, which is the case when $f_k \perp f_l$, which means f_k and f_l cannot be so "close" in a Parseval frame.

Similarly,

$$\min_{a\in\mathbb{K}} \|f_k - af_l\|^2 = \frac{\|f_k\|^2\,\|f_l\|^2 - |\langle f_k, f_l\rangle|^2}{\|f_l\|^2}$$

represents the distance between f_k and the line $L_{f_l} := \{af_l \mid a \in \mathbb{K}\}$.

Theorem 7 *Let $k \neq l$. For $a \in \mathbb{K}$, we have*

$$\|f_k - af_l\|^2 \geq \frac{\|f_k\|^2 + \|f_l\|^2 - 1}{\|f_l\|^2}.$$

If $\|f_k\|$ is close to 1, then the distance between f_k and L_{f_l} is close to 1, which means f_k and f_l cannot be so "close" in a Parseval frame.

References

1. Christensen, O.: Frames and bases: an introductory cource. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston (2008)
2. Christensen, O.: An Introduction to Frames and Riesz Bases. Applied and Numerical Harmonic Analysis, 2nd edn. Birkhäuser/Springer, Basel/Berlin (2016)
3. Daubechies, I.: Ten lectures on wavelets. CBS-NSF Regional Conferences in Applied Mathematics, vol. 61. SIAM, Philadelphia (1992)
4. Duffin, R.J., Schaeffer, A.C.: A class of nonharmonic Fourier series, Trans. Am. Math. Soc. **72**, 341–366 (1952)
5. Mallat, S.: A Wavelet Tour of Signal Processing, The Sparse Way, 3rd edn. Academic Press, Amsterdam (2009)
6. Saitoh, S., Sawano, Y.: Theory of Reproducing Kernels and Applications. Springer, Singapore (2016)

p-Adic Time-Frequency Analysis and Its Properties

Toshio Suzuki

Abstract $\mathbf{Q}_p$ is the field of p-adic numbers defined by the completion of the field of rational numbers with respect to the p-adic norm. The p-adic number field $\mathbf{Q}_p$ was introduced by Kurt Hensel in 1897. The p-adic analysis, which is the mathematical analysis of functions defined on $\mathbf{Q}_p$, has attracted attention in a variety of fields such as image processing and data compression. In this paper, we study the time-frequency analysis for complex valued functions on $\mathbf{Q}_p$. Especially we will construct the p-adic Stockwell transform and see its properties.

1 p-Adic Field $\mathbf{Q}_p$

The p-adic number field $\mathbf{Q}_p$ was introduced by Kurt Hensel in 1897. The applications of p-adic numbers have attracting attention not only in mathematics [1, 2] but also in various other fields. The topology of $\mathbf{Q}_p$ is quite different from the one of the real numbers field $\mathbf{R}$. In this section, we see the properties of the p-adic number field.

1.1 Definition of the p-Adic Field

For a prime number p, the rational number $x(\neq 0)$ can be represented as

$$x = p^{\gamma}\frac{m}{n},$$

T. Suzuki (✉)
Tokyo University of Science, Shinjuku-ku, Japan
e-mail: tosuzuki@rs.tus.ac.jp

U. Kähler et al. (eds.), *Analysis, Applications, and Computations*,
Research Perspectives, https://doi.org/10.1007/978-3-031-36375-7_55

where $\gamma = \gamma(x) \in \mathbf{Z}$ and $m, n \in \mathbf{Z}$ are not divisible by p. Then the p-adic norm is defined as

$$|x|_p = \begin{cases} 0 & (x = 0), \\ p^{-\gamma} & (x \neq 0). \end{cases}$$

Remark that the p-adic norm may take only countable set of values. The field of $\mathbf{Q}_p$ is defined by the completion of the field of rational numbers $\mathbf{Q}$ with respect to the p-adic norm $|\cdot|_p$ [9]. Ostrowski theorem gives us that every non-trivial norm on the set of rational numbers $\mathbf{Q}$ is equivalent to either the usual real absolute value or a p-adic norm. Therefore, it is natural idea to think of a p-adic number field $\mathbf{Q}_p$.

The p-adic norm has the following properties: For $x, y \in \mathbf{Q}_p$,

1. $|x|_p \geq 0, \quad |x|_p = 0 \Leftrightarrow x = 0,$
2. $|xy|_p = |x|_p |y|_p,$
3. $|x + y|_p \leq \max\{|x|_p, |y|_p\}$ (Strong triangle inequality).

Moreover, if $|x|_p \neq |y|_p$, the p-adic norm satisfies $|x + y|_p = \max\{|x|_p, |y|_p\}$. Since the p-adic norm satisfies the inequality 3, it is called non-Archimedean.

1.2 The p-Adic Canonical Form

Any p-adic number $x(\neq 0) \in \mathbf{Q}_p$ such that $|x|_p = p^{-\gamma} (\gamma \in \mathbf{Z})$ can be represented as the canonical form

$$x = p^{\gamma} \sum_{j=0}^{\infty} x_j p^j = p^{\gamma}(x_0 + x_1 p + x_2 p^2 + \cdots)$$

where $0 \leq x_j \leq p - 1 \quad (0 \leq j < \infty)$ and $x_0 \neq 0$. This series converges in the sense of p-adic norm. For example, the following equation holds:

$$-1 = (p-1) + (p-1)p + (p-1)p^2 + \cdots .$$

By adding 1 to both sides, we can verify that this equation is valid.

Using the canonical form, we can define the fractional part of the p-adic number. Let $x \in \mathbf{Q}_p$ have the canonical form $x = p^{\gamma}(x_0 + x_1 p + x_2 p^2 + \cdots)$. Then, the fractional part of x is defined as follows:

$$\{x\}_p = \begin{cases} 0 & (\gamma \geq 0 \text{ or } x = 0), \\ p^{\gamma}(x_0 + x_1 p + x_2 p^2 + \cdots + x_{-\gamma-1} p^{-\gamma-1}) & (\gamma < 0). \end{cases}$$

1.3 The Topology of $\mathbf{Q}_p$

Since the p-adic norm and the absolute value are different, the topology of $\mathbf{Q}_p$ is also different from the topology of $\mathbf{R}$. For $a \in \mathbf{Q}_p, \gamma \in \mathbf{Z}$, we put the p-adic ball and the sphere as

$$B_\gamma(a) = \{x \in \mathbf{Q}_p \mid |x-a|_p \le p^\gamma\}, \quad S_\gamma(a) = \{x \in \mathbf{Q}_p \mid |x-a|_p = p^\gamma\}.$$

Especially, when the center is at the origin, we write $B_\gamma = B_\gamma(0)$, $S_\gamma = S_\gamma(0)$. The following properties are valid:

1. $B_\gamma(a) = B_{\gamma+1}(a) \setminus S_{\gamma+1}(a)$.
2. $B_\gamma(a)$ and $S_\gamma(a)$ are both open and closed set in $\mathbf{Q}_p$.
3. Any two balls in $\mathbf{Q}_p$ either disjoint or one is contained in another.
4. $\mathbf{Q}_p = \bigcup_{\gamma\in\mathbf{Z}} B_\gamma(a) = \bigcup_{\gamma\in\mathbf{Z}} S_\gamma(a)$.

$\mathbf{Q}_p$ can be represented as the union of the p-adic spheres. These properties give us that $\mathbf{Q}_p$ is a totally disconnected space.

2 p-Adic Time-Frequency Analysis

Since the topology of $\mathbf{Q}_p$ is different from the topology of $\mathbf{R}$, the p-adic time-frequency analysis is also different from the case on $\mathbf{R}$.

2.1 p-Adic Calculus

There exists a Haar measure dx on $\mathbf{Q}_p$, which is positive, shift invariant $d(x+a) = dx$ and normalized by $\int_{B_0} dx = 1$. For f, which is mapping on $\mathbf{Q}_p$ we define the L^q norm $(1 \le q < \infty)$ as

$$\|f\|_{L^q(\mathbf{Q}_p)} = \left(\int_{\mathbf{Q}_p} |f(x)|^q dx\right)^{1/q}$$

and the L^q space is defined by

$$L^q(\mathbf{Q}_p) = \{f : \mathbf{Q}_p \to \mathbf{C} \mid \int_{\mathbf{Q}_p} |f(x)|^q dx < \infty\}.$$

If $q = 2$, the L^2 space is a Hilbert space with the inner product

$$(f, g)_{L^2(\mathbf{Q}_p)} = \int_{\mathbf{Q}_p} f(x)\overline{g(x)}dx,$$

for $f, g \in L^2(\mathbf{Q}_p)$ where $\overline{g(x)}$ is the complex conjugate of $g(x)$.

2.2 p-Adic Fourier Transform

We define the additive character of the field $\mathbf{Q}_p$ as $\chi_p(x) = \exp(2\pi i\{x\}_p)$. Then the p-adic Fourier transform of $f \in L^2(\mathbf{Q}_p)$ is defined by

$$\mathcal{F}f(\xi) = \hat{f}(\xi) = \int_{\mathbf{Q}_p} f(x)\chi_p(\xi x)dx \quad (\xi \in \mathbf{Q}_p)$$

and its inverse by

$$\mathcal{F}^{-1}f(\xi) = \check{f}(\xi) = \int_{\mathbf{Q}_p} f(x)\chi_p(-\xi x)dx \quad (\xi \in \mathbf{Q}_p).$$

We can check that the p-adic Fourier transformation $f \to \hat{f}$ is a linear isomorphism from $L^2(\mathbf{Q}_p)$ onto $L^2(\mathbf{Q}_p)$ and for any $f, g \in L^2(\mathbf{Q}_p)$,

$$(f, g)_{L^2(\mathbf{Q}_p)} = (\hat{f}, \hat{g})_{L^2(\mathbf{Q}_p)}, \qquad \|f\|_{L^2(\mathbf{Q}_p)} = \|\hat{f}\|_{L^2(\mathbf{Q}_p)}$$

hold. On the other hand, the integration of the additive character on $\mathbf{Q}_p$ has the following property. See [6] for the proof.

Proposition 1 *For $\gamma \in \mathbf{Z}$,*

$$\lambda(\xi, \gamma) = \int_{S_\gamma} \chi_p(\xi x)dx = \begin{cases} p^\gamma\left(1 - \frac{1}{p}\right) & (|\xi|_p \le p^{-\gamma}), \\ -p^{\gamma-1} & (|\xi|_p = p^{-\gamma+1}), \\ 0, & (|\xi|_p \ge p^{-\gamma+2}) \end{cases}$$

and more generally,

$$\lambda(\xi, \gamma; k_0, \cdots, k_l) = \int_{S_\gamma, x_0 = k_0, \cdots, x_l = k_l} \chi_p\left(|\xi|_p^{-1}x\right)dx$$

$$= \begin{cases} \chi_p\left(|\xi|_p^{-1}p^{-\gamma}\left(k_0 + \cdots + k_l p^l\right)\right)p^{\gamma-l-1}, & \text{if } |\xi|_p \le p^{\gamma-l-1}, \\ 0, & \text{if } |\xi|_p \ge p^{\gamma-l}. \end{cases}$$

Moreover let $f \in L^2(\mathbf{Q}_p)$ be a function which depends only on the value of $|x|_p$. Then, we have

$$\int_{\mathbf{Q}_p} f(x)\chi_p(\xi x)dx = \sum_{\gamma\in\mathbf{Z}} f\left(p^{\gamma}\right)\int_{S_\gamma}\chi_p(|\xi|_p^{-1}x)dx.$$

This proposition gives us that a Fourier transform of the function which depends only on the value of $|x|_p$ is also a function which depends only on the value of $|\xi|_p$.

2.3 *p-Adic Time Frequency Analysis*

First, we see the definition of the p-adic windowed Fourier transform. See [6] for the properties of the p-adic windowed Fourier transform and the proof of the following theorem.

Definition 1 Let $g \in L^1(\mathbf{Q}_p) \cap L^2(\mathbf{Q}_p)$. For $f \in L^2(\mathbf{Q}_p)$, the definition of the windowed (short-time) Fourier transform with the window function $g \in L^1(\mathbf{Q}_p) \cap L^2(\mathbf{Q}_p)$ is as follows:

$$(G_g f)(b,\xi) = \frac{1}{\|g\|_2}\int_{\mathbf{Q}_p} f(x)\overline{g(x-b)}\chi_p(\xi x)dx, \quad b,\xi\in\mathbf{Q}_p. \tag{1}$$

Under some assumptions, the p-aidc windowed Fourier transform of a function can be represented as the sum of the values of the function.

Theorem 1 *Assume that $f, g \in L^2(\mathbf{Q}_p)$ are functions which depend only on the value of $|x|_p$, then*

$$\begin{aligned}\left(G_g f\right)(b,\xi) = {} & \frac{1}{\|g\|_2}\sum_{k=1}^{p-1}\sum_{\gamma>\gamma_b} f\left(p^{-\gamma}\right)\bar{g}\left(p^{\gamma}\right)\lambda(\xi,\gamma;k)\\ & + \frac{\bar{g}\left(|b|_p\right)}{\|g\|_2}\sum_{k=1}^{p-1}\sum_{\gamma<\gamma_b} f\left(p^{-\gamma}\right)\lambda(\xi,\gamma;k)\\ & + \frac{f\left(p^{-\gamma_b}\right)}{\|g\|_2}\sum_{k=0}^{\infty}\bar{g}\left(p^{\gamma_q-k}\right)[\lambda\left(\xi,\gamma_b;b_0,\cdots,b_{k-1}\right)\\ & -\lambda\left(\xi,\gamma_b;b_0,\cdots,q_k\right)]\end{aligned}$$

where the $\gamma_b = |b|_p$.

Next, we see the definition of the p-adic wavelet transform. See [5] for the properties of the p-adic wavelet transform and the proofs of the following Proposition and Theorem. The (continuous) wavelet transform is an integral transform which provides a representation of a signal by the varing the translation and scale parameters of a wavelet [3].

Definition 2 If $\psi \in L^1(\mathbf{Q}_p) \cap L^2(\mathbf{Q}_p)$ satisfies the admissible condition

$$c_\psi = \int_{\mathbf{Q}_p} \frac{|\hat{\psi}(a)|^2}{|a|_p} da < \infty,$$

then we call ψ a wavelet. Let $\alpha \in \mathbf{R}$, $\psi \in L^1(\mathbf{Q}_p) \cap L^2(\mathbf{Q}_p)$ be a wavelet.

For $f \in L^2(\mathbf{Q}_p)$, we define the p-adic (continuous) wavelet transform by

$$(\Omega_\psi f)(b, a) = \frac{1}{\sqrt{c_\psi}|a|_p^\alpha} \int_{\mathbf{Q}_p} f(x) \overline{\psi\left(\frac{x-b}{a}\right)} dx.$$

Proposition 2 *Let $\psi \in L^2(\mathbf{Q}_p)$ satisfy the admissible condition. Then, for $f \in L^2(\mathbf{Q}_p)$ and $\alpha, \beta \in \mathbf{R}$ such that $2\alpha + \beta = 3$,*

$$f(x) = \frac{1}{\sqrt{c_\psi}} \int_{\mathbf{Q}_p} \frac{da}{|a|_p^{\alpha+\beta}} \int_{\mathbf{Q}_p} \left(\Omega_\psi f\right)(b, a) \psi\left(\frac{x-b}{a}\right) db.$$

Under some assumptions, the p-adic wavelet transform of a function can be also represented as the sum of the values of the function.

Theorem 2 *Assume that $f \in L^2(\mathbf{Q}_p)$ depends only on the value of $|x|_p$, and ψ is a wavelet. Then,*

$$\begin{aligned}(\Omega_\psi f)(b, a) = \frac{1}{\sqrt{c_\psi}|a|_p^\alpha} \Bigg\{ & \left(1 - \frac{1}{p}\right) \sum_{\gamma > \gamma_b} f\left(p^{-\gamma}\right) \overline{\psi\left(|a|_p p^{-\gamma}\right)} p^\gamma \\ & + \left(1 - \frac{1}{p}\right) \overline{\psi\left(\frac{|a|_p}{|b|_p}\right)} \sum_{\gamma < \gamma_b} f\left(p^{-\gamma}\right) p^\gamma \\ & + \left(1 - \frac{1}{p}\right) |b|_p f\left(|b|_p^{-1}\right) \sum_{k=1}^{\infty} \overline{\psi\left(\left|\frac{a}{b}\right|_p p^k\right)} p^{-k} \\ & + \left(1 - \frac{2}{p}\right) |b|_p f\left(|b|_p^{-1}\right) \overline{\psi\left(\left|\frac{a}{b}\right|_p\right)} \Bigg\}\end{aligned}$$

where $\gamma_b = |b|_p$

See [5] for the proofs.

For $f \in L^2(\mathbf{Q}_p)$, $a, b \in \mathbf{Q}_p$, we define the translate, dilation, and modulation operators as

$$(T_b f)(x) = f(x+b)$$
$$(D_{1/a} f)(x) = \frac{1}{|a|^\alpha} f\left(\frac{x}{a}\right)$$
$$(M_\xi f)(x) = \chi_p(\xi x) f(x)$$

Then, the p-adic windowed Fourier transform and the p-adic wavelet transform can be represented as follows:

$$(G_g f) = (f, M_{-\xi} T_{-b} g)_{L^2(\mathbf{Q}_p)}, \quad (\Omega_\varphi f) = (f, T_{-b} D_{1/a} \frac{\varphi}{\sqrt{c_\psi}})_{L^2(\mathbf{Q}_p)}.$$

2.4 *p-Adic Stockwell Transform*

The Stockwell transform (S transform) was introduced by R. G. Stockwell (see [8]) for analyzing geophysics data. This transform is said that it is the hybrid transform of the windowed Fourier transform and the wavelet transform. There are a lot of studies of the Stockwell transform [4, 10].

First, we see the definition of the p-adic Stockwell transform. The following results are our previous results. See [7] for the proofs.

Definition 3 Let $g \in L^2(\mathbf{Q}_p)$ be a function with compact support. For $f \in L^2(\mathbf{Q}_p)$, we define the Stockwell transform S_g by

$$(S_g f)(b, \xi) = |\xi|_p \int_{\mathbf{Q}_p} f(x) \overline{g(\xi(x-b))} \chi_p(x\xi) dx.$$

We can find that the p-adic Stockwell transform contains the translation, dilation, and modulation factors. Especially, we can represent the p-adic Stockwell transform with the p-adic wavelet transform.

Proposition 3 *Let $\psi \in L^2(\mathbf{Q}_p)$ be a wavelet and $\psi(x) = g(x)\chi_p(-x)$. Then, for the wavelet transform Ω and Stockwell transform S, the following relation has hold:*

$$(S_g f)(b, \xi) = \sqrt{c_\psi} |\xi|_p^{-\alpha+1} \chi_p(b\xi)(\Omega_\psi f)(b, 1/\xi).$$

The following theorem is the Parseval-Steklov type identity for the p-adic Stockwell transform.

Theorem 3 *Assume that* $g \in L^1(\mathbf{Q}_p) \cap L^2(\mathbf{Q}_p)$ *satisfies* $\|g\|_{L^2(\mathbf{Q}_p)} = 1$ *and*

$$c_g = \int_{\mathbf{Q}_p} \frac{|\hat{g}(\xi - 1)|^2}{|\xi|_p} d\xi < \infty.$$

Then, for any $f, h \in L^2(\mathbf{Q}_p)$,

$$(f, h)_{L^2(\mathbf{Q}_p)} = \frac{1}{c_g} \int_{\mathbf{Q}_p} \int_{\mathbf{Q}_p} S_g f(b, \xi) \overline{S_g h(b, \xi)} \frac{db d\xi}{|\xi|_p}.$$

Especially, if $f = h$, we get

$$\|f\|_{L^2(\mathbf{Q}_p)} = \frac{1}{\sqrt{c_g}} \|S_g f\|_{L^2(\mathbf{Q}_p^2/|\xi|_p)}.$$

Similar to the case of the Stockwell transform for functions on real numbers, we can obtain the inversion formula of the p-adic Stockwell transform.

Theorem 4 *Let* $g \in L^2(\mathbf{Q}_p)$ *satisfy* $\|g\|_{L^2(\mathbf{Q}_p)} = 1$ *and*

$$c_g = \int_{\mathbf{Q}_p} \frac{|\hat{g}(\xi - 1)|^2}{|\xi|_p} d\xi < \infty.$$

Then, for any $f \in L^2(\mathbf{Q}_p)$,

$$f(x) = \frac{1}{c_g} \int_{\mathbf{Q}_p} \int_{\mathbf{Q}_p} S_g f(b, \xi) g(\xi(x - b)) \chi_p(-bx) \frac{d\xi db}{|\xi|_p}.$$

The Stockwell transform of a function which depends only on the value of $|x|_p$ can be represented as the sum of the function like the windowed Fourier transform and the wavelet transform.

Theorem 5 *Let* $f \in L^2(\mathbf{Q}_p)$, $g \in L^1(\mathbf{Q}_p) \cap L^2(\mathbf{Q}_p)$ *be functions depend only on the value of* $|x|_p$. *Let* $b, \xi \in \mathbf{Q}_p$ *and* $b = p^{-\gamma_b}(b_0 + b_1 p + b_2 p^2 + \cdots)$. *Then, we have*

$$\begin{aligned}
\left(S_g f\right)(b, \xi) = {} & |\xi|_p \sum_{\gamma > \gamma_b} f(p^\gamma) \overline{g\left(|\xi|_p p^\gamma\right)} \lambda(\xi, \gamma) \\
& + |\xi|_p \overline{g(|\xi| p^{\gamma_b})} \sum_{\gamma < \gamma_b} f(p^\gamma) \lambda(\xi, \gamma) \\
& + |\xi|_p f(p^{\gamma_b}) \sum_{k=0}^{\infty} \overline{g(|\xi|_p p^{\gamma_b - k})} \\
& \quad \times \left(\lambda(\xi, \gamma_b; b_0, \ldots, b_{k-1}) - \lambda(\xi, \gamma_b; b_0, \ldots, b_k)\right).
\end{aligned}$$

References

1. Albeverio, S., Khrennikov, A.Y., Shelkovich, V.M.: Theory of p-Adic Distributions: Linear and Nonlinear Models. Cambridge University Press, Cambridge (2010)
2. Chuong, N.M., Egorov, Y.V., Khrennikov, A., Meyer, Y., Mumford, D. (eds.): Harmonic, Wavelet and p-Adic Analysis. World Scientific, Singapore (2007)
3. Daubechies, I.: Ten lectures on wavelets. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 61. SIAM, Philadelphia (1992)
4. Du, J., Wong, M.W., Zhu, H.: Continuous and discrete inversion formulas for the Stockwell transform. Integr. Trans. Spec. Funct. **18**, 537–543 (2007)
5. C. Minggen, G. Gao, P.U. Chung, On the wavelet transform in the field $\mathbf{Q}_p$ of p-adic number. Appl. Comput. Harmon. Anal. **13**, 162–168 (2002)
6. S.Y. Park, P.U. Chung, An application of p-adic analysis to windowed fourier transform. Kangweon-Kyungki Math. J. **12**(2), 193–200 (2004)
7. Suzuki, T.: The construction of p-adic stockwell transform and its spectra, p-adic numbers. Ultrametric Anal. Appl. **13**(2), 166–173 (2021)
8. Stockwell, R.G., Mansinha L., Lowe, R.P.: Localization of the complex spectrum, the S transform. IEEE Trans. Signal Process. **44**, 998–1001 (1996)
9. Vladimirov, V.S., Volovich, I.V., Zelenov, E.I.: p-Adic Analysis and Mathematical Physics. World Scientific, Singapore (1994)
10. Wong, M.W., Zhu, H.: A characterization of the Stockwell spectrum. In: Toft, J., Wong, M.W., Zhu, H. (eds.) Modern Trends in Pseudo-Differential Operators. Operator Theory: Advances and Applications, pp. 251–257. Birkhäuser, Basel (2007)