Springer Optimization and Its Applications 134

Themistocles M. Rassias *Editor*

Applications of Nonlinear Analysis

Springer

Springer Optimization and Its Applications

Volume 134

Aims and Scope

Optimization has been expanding in all directions at an astonishing rate during the last few decades. New algorithmic and theoretical techniques have been developed, the diffusion into other disciplines has proceeded at a rapid pace, and our knowledge of all aspects of the field has grown even more profound. At the same time, one of the most striking trends in optimization is the constantly increasing emphasis on the interdisciplinary nature of the field. Optimization has been a basic tool in all areas of applied mathematics, engineering, medicine, economics and other sciences.

The series *Springer Optimization and Its Applications* aims to publish state-of-the-art expository works (monographs, contributed volumes, textbooks) that focus on algorithms for solving optimization problems and also study applications involving such problems. Some of the topics covered include nonlinear optimization (convex and nonconvex), network flow problems, stochastic optimization, optimal control, discrete optimization, multi-objective programming, description of software packages, approximation techniques and heuristic approaches.

More information about this series at http://www.springer.com/series/7393

Themistocles M. Rassias
Editor

Applications of Nonlinear Analysis

Springer

Editor
Themistocles M. Rassias
Department of Mathematics
National Technical University of Athens
Athens, Greece

ISSN 1931-6828 ISSN 1931-6836 (electronic)
Springer Optimization and Its Applications
ISBN 978-3-319-89814-8 ISBN 978-3-319-89815-5 (eBook)
https://doi.org/10.1007/978-3-319-89815-5

Library of Congress Control Number: 2018946611

Mathematics Subject Classification: 26-XX, 28-XX, 30-XX, 32-XX, 34-XX, 35-XX, 37-XX, 39-XX, 41-XX, 43-XX, 45-XX, 46-XX, 47-XX, 49-XX, 52-XX, 53-XX, 54-XX, 57-XX, 58-XX, 65-XX

Printed on acid-free paper

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

Preface

The *Applications of Nonlinear Analysis* presents some classical and new results in important subjects of nonlinear analysis and its applications.

The contributing papers have been written by experts from the international mathematical community. These papers deepen our understanding of some of the most essential research problems and theories of nonlinear nature.

Effort has been made for the presentation of the concepts, theories, and methods to reach wide readership.

I would like to express my thanks to all the scientists who contributed to the preparation of this volume. I would also like to acknowledge the superb assistance of the staff of Springer for the publication of this book.

Athens, Greece

Themistocles M. Rassias

Contents

New Applications of γ-Quasiconvexity

Shoshana Abramovich

1 Introduction

This survey deals with inequalities satisfied by γ-quasiconvex functions which are one of the many variants of convex functions. Convex functions and their variants are dealt with extensively (see for instance the classical books [11, 14, 16] and their references) The γ-quasiconvex functions have already been dealt with by S. Abramovich, L.-E. Persson and N. Samko. The basic facts on γ-quasiconvexity on which this paper is built, can be found in [2, 8], and [9]. This survey is an extension and continuation of [2], which include previous results related to γ-quasiconvexity.

We state here additional results to those in Survey [2]. In particular results concerning Hölder, Minkowski, Jensen-Steffensen and Slater-Pečarić type inequalities for γ-quasiconvex functions.

We start with a definition of and lemmas about γ-quasiconvexity, see [1, 2, 8], and [9]:

Definition 1 Let γ be a real number. A real-valued function f defined on an interval $[0, b)$ with $0 < b \leq \infty$ is called γ-quasiconvex (γ-quasiconcave) if it can be represented as the product of a convex (concave) function and the power function x^{γ}.

A convex function φ on $[0, b)$, $0 < b \leq \infty$ is characterized by the inequality

$$\varphi(y) - \varphi(x) \geq C_{\varphi}(x)(y - x), \quad \forall x, y \in [0, b), \quad C_{\varphi} \in \mathbb{R}, \tag{1}$$

from which the following lemmas is easily established:

S. Abramovich (✉)
Department of Mathematics, University of Haifa, Haifa, Israel
e-mail: abramos@math.haifa.ac.il

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization
and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_1

Lemma 1 ([8, Lemma 1]) *Let $\psi_\gamma(x) = x^\gamma \varphi(x)$, $\gamma \in \mathbb{R}$, where φ is convex on $[0,b)$, that is, ψ_γ is a γ-quasiconvex function. Then*

$$\psi_\gamma(y) - \psi_\gamma(x) \geq \varphi(x)\left(y^\gamma - x^\gamma\right) + C_\varphi(x)\, y^\gamma (y-x), \tag{2}$$

holds for all $x \in [0,b)$, $y \in [0,b)$, where $C_\varphi(x)$ is defined by (1).

It is obvious that $\psi(x) = x^{p+\gamma}$, $x > 0$, $p \geq 1$ is γ-quasiconvex and when $0 < p < 1$ is γ-quasiconcave.

The following lemma is derived by some computation on the right handside of (2), (see also [9, Lemma 2]):

Lemma 2 ([9]) *Let φ be convex differentiable function on $[a,b)$ and let $\psi_k(x) = x^k\varphi(x)$, $k = 1,2,\ldots,N$. Then the N-quasiconvex function $\psi_N(x) = x^N\varphi(x)$ satisfies for $a \leq x < y < b$, $a \geq 0$*

$$\begin{aligned} &\psi_N(y) - \psi_N(x) \qquad\qquad (3)\\ &\geq (\psi_N(x))'(y-x) + (y-x)^2 \sum_{k=1}^{N} y^{k-1}(\psi_{N-k}(x))' \\ &= (\psi_N(x))'(y-x) + (y-x)^2 \frac{\partial}{\partial x}\left(\frac{x^N - y^N}{x-y}\varphi(x)\right). \end{aligned}$$

In Sect. 2 we state results about Jensen's type and Slater-Pečarić type inequalities when the coefficients $\alpha_i \geq 0$, $i = 1,\ldots,n$. Also, we quote inequalities for which the coefficients are not always non-negative. We call these coefficients **Steffensen's coefficients**.

In Sect. 3 Hardy type inequalities are presented.

By using the results stated about Jensen type inequalities for γ-quasiconvex functions we get in Sect. 4 Hölder's type inequalities which are of the type

$$\int fg\,d\nu \lesseqgtr \left(\int g^q d\nu\right)^{1/q}\left(\int f^p d\nu\right)^{1/p} H(f,g)$$

that lately are widely discussed (see for instance [12, 13, 15, 19] and their references).

In Sect. 5 we state Minkowski type inequalities which are derived by using again the Jensen type inequalities for the γ-quasiconvex functions $f(x) = x^{p+1}$, $x \geq 0$, $p \geq 1$.

Finally, in Sect. 6 we get by the γ-quasiconvexity technique an estimation of Jensen Gap, in particular, for functions that have Taylor power series representation.

2 Jensen and Slater-Pečarić Type Inequalities for N-quasiconvex Functions

2.1 *Jensen and Slater-Pečarić Type Inequalities for N-quasiconvex Functions with Non-negative Coefficients*

We quote here some of the basic results which appear in [9], which are used to prove the theorems stated in the sequel.

Theorem 1 *Let $\varphi : [a, b) \to \mathbb{R}$, $a \geq 0$ be convex differentiable function, and let $\psi_k(x)$ be*

$$\psi_k(x) = x^k \varphi(x), \qquad k = 0, 1, \dots, N, \qquad \psi_0 = \varphi.$$

Let

$$\alpha_i \geq 0, \qquad x_i \in [a, b), \ i = 1, \dots, n, \qquad \sum_{i=1}^{n} \alpha_i = 1.$$

Denote

$$\overline{x} = \sum_{i=1}^{n} \alpha_i x_i,$$

then:

1) A Jensen's type inequality holds:

$$\begin{aligned} & \sum_{i=1}^{n} \alpha_i \psi_N(x_i) - \psi_N(\overline{x}) \\ \geq & \sum_{i=1}^{n} \alpha_i \varphi(\overline{x}) \left(x_i^N - \overline{x}^N \right) + \sum_{i=1}^{n} \alpha_i \varphi'(\overline{x}) x_i^N (x_i - \overline{x}) \\ = & \sum_{i=1}^{n} \sum_{k=1}^{N} \alpha_i (x_i - \overline{x})^2 x_i^{k-1} (\psi_{N-k}(\overline{x}))' \\ = & \sum_{i=1}^{n} \alpha_i (x_i - \overline{x})^2 \frac{\partial}{\partial \overline{x}} \left(\frac{\overline{x}^N - x_i^N}{\overline{x} - x_i} \varphi(\overline{x}) \right). \end{aligned} \tag{4}$$

If φ is also non-negative and increasing then for $N = 2, \dots,$ the above inequality refines Jensen's inequality. For $N = 1$ we get for $\psi_1(x) = x\varphi(x)$

$$\begin{aligned}&\sum_{i=1}^{n}\alpha_i\psi_1\left(x_i\right)-\psi_1\left(\overline{x}\right) \qquad (5)\\ &\geq\sum_{i=1}^{n}\alpha_i\varphi^{'}\left(\overline{x}\right)x_i\left(x_i-\overline{x}\right)=\sum_{i=1}^{n}\alpha_i\varphi^{'}\left(\overline{x}\right)\left(x_i-\overline{x}\right)^2.\end{aligned}$$

If φ is increasing and convex (and not necessarily non-negative) then again (5) is a refinement of Jensen's inequality.

2) *For a fixed $C\in[a,b)$ we get when*

$$\alpha_i\geq 0,\qquad i=1,\ldots,n,\qquad \sum_{i=1}^{n}\alpha_i=1$$

that

$$\begin{aligned}&C^N\varphi\left(C\right)-\sum_{i=1}^{n}\alpha_i x_i^N\varphi\left(x_i\right)=\psi_N\left(C\right)-\sum_{i=1}^{n}\alpha_i\psi_N\left(x_i\right)\\ &\geq\sum_{i=1}^{n}\alpha_i\left(x_i^N\varphi\left(x_i\right)\right)'\left(C-x_i\right)+\sum_{i=1}^{n}\alpha_i\left(C-x_i\right)^2\sum_{k=1}^{N}C^{k-1}\left(\psi_{N-k}\left(x_i\right)\right)'\\ &=\sum_{i=1}^{n}\alpha_i\left(x_i^N\varphi\left(x_i\right)\right)'\left(C-x_i\right)+\sum_{i=1}^{n}\alpha_i\left(C-x_i\right)^2\frac{\partial}{\partial x_i}\left(\frac{x_i^N-C^N}{x_i-C}\varphi\left(x_i\right)\right).\end{aligned}$$

3) *Especially if*

$$\sum_{i=1}^{n}\alpha_i\psi_N'\left(x_i\right)>0,$$

and if

$$C=M_{\psi_N}=\frac{\sum_{i=1}^{n}\alpha_i x_i\psi_N'\left(x_i\right)}{\sum_{i=1}^{n}\alpha_i\psi_N'\left(x_i\right)}\in[a,b),$$

then we get a Slater-Pečarić type inequality

$$\begin{aligned}&\psi_N\left(M_{\psi_N}\right)-\sum_{i=1}^{n}\alpha_i\psi_N\left(x_i\right)\\ &\geq\sum_{i=1}^{n}\sum_{k=1}^{N}\alpha_i\left(M_{\psi_N}-x_i\right)^2M_{\psi_N}^{k-1}\left(\psi_{N-k}\left(x_i\right)\right)'\\ &=\sum_{i=1}^{n}\alpha_i\left(M_{\psi_N}-x_i\right)^2\frac{\partial}{\partial x_i}\left(\frac{M_{\psi_N}^N-x_i^N}{M_{\psi_N}-x_i}\varphi\left(x_i\right)\right).\end{aligned}$$

If φ is also non-negative and increasing then for $N = 1, \ldots$ the above inequality is a refinement of Slater Pečarić inequality.

We get in [9, Theorem 1] the integral form of Jensen's type inequality for γ-quasiconvex functions and the special case when $\gamma = 1$ is:

Lemma 3 ([2, 9]) *Let f be a non-negative function. Let f and $\varphi \circ f$ be μ-integrable functions on the probability measure space (Ω, μ) and*

$$\int_{\Omega} f(s)\, d\mu(s) > 0.$$

Let also $\psi(x) = x\varphi(x)$. If φ is a differentiable convex on $[0, b)$, $0 < b \leq \infty$

$$\int_{\Omega} \psi(f(s))\, d\mu(s) - \psi\left(\int_{\Omega} f(s)\, d\mu(s)\right)$$
$$\geq \int_{\Omega} \varphi'\left(\int_{\Omega} f(\sigma)\, d\mu(\sigma)\right)\left(f(s) - \int_{\Omega} f(\sigma)\, d\mu(\sigma)\right)^2 d\mu(s).$$

hold. If φ is also increasing we get a refinement of Jensen's inequality.

Example 1 Let

$$\varphi(x) = e^{x^3}, \qquad \psi(x) = xe^{x^3}$$

then from the convexity of ψ we get that

$$\int_0^1 \psi(x)\, dx \geq \frac{e^{\frac{1}{8}}}{2}$$

and from the 1-quasiconvexity we get the better result

$$\int_0^1 \psi(x)\, dx \geq \frac{5e^{\frac{1}{8}}}{8}.$$

2.2 Jensen and Slater-Pečarić Type Inequalities for Steffensen's Coefficients

We state now a Jensen-Steffensen type inequality and a Slater-Pečarić type inequality for N-quasiconvex functions, when N is an integer, and the coefficients are not necessarily non-negative. These coefficients $\rho_1, \ldots, \rho_n$ are called **Steffensen coefficients**, and satisfy

$$0 \leq P_k = \sum_{i=1}^{k} \rho_i \leq P_n, \qquad \overline{P}_k = \sum_{i=k}^{n} \rho_i \geq 0,\, P_n > 0, \qquad k = 1, \ldots, n.$$

For 1-quasiconvex function ψ we present a Jensen's type inequality obtained in [9, Theorem 3]:

Theorem 2 *Let $\rho_1, \dots, \rho_n$ be Steffensen coefficients, and let*

$$\mathbf{x} = (x_1, \dots, x_n) > 0$$

satisfy

$$0 < x_1 \leq \dots \leq x_n.$$

Let φ be non-negative, increasing differentiable convex function defined on $x \geq 0$, and let $\psi(x) = x\varphi(x)$. Let

$$\overline{x} = \sum_{i=1}^{n} \frac{\rho_i x_i}{P_n}.$$

Let s be the integer that satisfies $0 < x_s \leq \overline{x} \leq x_{s+1} \leq x_n$. Then we get

$$\begin{aligned}
&\sum_{i=1}^{n} \rho_i \psi(x_i) - P_n \psi(\overline{x}) \\
&\geq \varphi'(x_1) \left(\sum_{j=1}^{s} P_j + \sum_{j=s+1} \overline{P}_j \right) \left(\frac{\sum_{i=1}^{n} \rho_i |x_i - \overline{x}|}{\sum_{j=1}^{s} P_i + \sum_{j=s+1}^{n} \overline{P}_j} \right)^2 \\
&\geq \varphi'(x_1) P_n \max\{s, n-s\} \left(\frac{\sum_{i=1}^{n} \rho_i |x_i - \overline{x}|}{P_n \max\{s, n-s\}} \right)^2 \\
&\geq \varphi'(x_1)(n-1) P_n \left(\frac{\sum_{i=1}^{n} \rho_i |x_i - \overline{x}|}{(n-1) P_n} \right)^2 \geq 0.
\end{aligned}$$

We state now a Jensen-Steffensen type inequality and Slater Pečarić type inequality for N-quasiconvex functions, when N is an integer. The proof of this theorem appears in [1], uses (3) and some of the techniques used in [6] and [7].

Theorem 3 *Let $\rho_1, \dots, \rho_n$ be Jensen-Steffensen coefficients, and let $\mathbf{x} = (x_1, \dots, x_n)$ satisfy $0 < x_1 \leq \dots \leq x_n$. Let φ be non-negative, increasing differentiable convex function defined on $x \geq 0$, and let*

$$\psi_N(x) = x^N \varphi(x)$$

where N is an integer. Let

$$\overline{x} = \sum_{i=1}^{n} \frac{\rho_i x_i}{P_n}.$$

Let s be the integer that satisfies

$$0 < x_s \leq \overline{x} \leq x_{s+1} \leq x_n.$$

Then

$$\begin{aligned}
&\sum_{i=1}^{n} \rho_i \psi_N(x_i) - P_n \psi_N(\overline{x}) \qquad (6)\\
&\geq \sum_{k=1}^{N} x_1^{k-1} \psi'_{N-k}(x_1) \left(\sum_{j=1}^{s} P_j + \sum_{j=s+1} \overline{P}_j\right) \left(\frac{\sum_{j=1}^{n} \rho_j \left|x_j - \overline{x}\right|}{\sum_{j=1}^{s} P_j + \sum_{j=s+1}^{n} \overline{P}_j}\right)^2\\
&= \left(\sum_{j=1}^{s} P_j + \sum_{j=s+1} \overline{P}_j\right) \left(\frac{\sum_{j=1}^{n} \rho_j \left|x_j - \overline{x}\right|}{\sum_{j=1}^{s} P_j + \sum_{j=s+1}^{n} \overline{P}_j}\right)^2 \frac{\partial}{\partial x}\left(\frac{x^N - x_1^N}{x - x_1}\varphi(x)\right)/_{x=x_1}\\
&\geq (P_n \max\{s, n-s\})^{-1} \left(\sum_{i=1}^{n} \rho_i |x_i - \overline{x}|\right)^2 \frac{\partial}{\partial x}\left(\frac{x^N - x_1^N}{x - x_1}\varphi(x)\right)/_{x=x_1}\\
&\geq ((n-1) P_n)^{-1} \left(\sum_{i=1}^{n} \rho_i |x_i - \overline{x}|\right)^2 \frac{\partial}{\partial x}\left(\frac{x^N - x_1^N}{x - x_1}\varphi(x)\right)/_{x=x_1} \geq 0
\end{aligned}$$

holds, unless one of the following two cases occurs:

1) *either $\overline{x} = x_1$ or $\overline{x} = x_n$,*
2) *there exists $k \in \{3, \ldots, n-2\}$ such that $\overline{x} = x_k$ and*

$$P_j\left(x_j - x_{j+1}\right) = 0, \qquad j = 1, \ldots, k-1, \quad \overline{P}_j\left(x_j - x_{j-1}\right) = 0, \\ j = k+1, \ldots, n.$$

In these two cases

$$\sum_{i=1}^{n} \rho_i \psi(x_i) - P_n \psi(\overline{x}) = 0.$$

A refinement of Slater-Pečarić inequality in case that ψ_N is N-quasiconvex functions is proved in [1] and is as follows:

Theorem 4 *Under the same conditions as in Theorem 3 on $(\rho_1, \ldots, \rho_n)$, $(x_1, \ldots, x_n)$ and on*

$$\psi_k(x) = x^k \varphi(x), \qquad k = 0, 1, \ldots, N,$$

if

$$\sum_{i=1}^{n}\rho_i\psi_N'(x_i)\neq 0,$$

we define

$$M_{\psi_N}=\frac{\sum_{i=1}^{n}\rho_i x_i\psi_N'(x_i)}{\sum_{i=1}^{n}\rho_i\psi_N'(x_i)}.$$

Let

$$\overline{x}=\sum_{i=1}^{n}\frac{\rho_i x_i}{P_n}.$$

Case A: for s satisfying

$$x_s\leq M_{\psi_N}\leq x_{s+1},\quad s+1\leq n,$$

then,

$$\begin{aligned}
&\sum_{i=1}^{n}\rho_i\psi_N(x_i)-P_n\psi_N\left(M_{\psi_N}\right) \qquad (7)\\
&\leq-\sum_{k=1}^{N}x_1^{k-1}\psi_{N-k}'(x_1)\left(\sum_{j=1}^{s}P_j+\sum_{j=s+1}\overline{P}_j\right)\left(\frac{\sum_{j=1}^{n}\rho_j\left|x_j-\overline{x}\right|}{\sum_{j=1}^{s}P_j+\sum_{j=s+1}^{n}\overline{P}_j}\right)^2\\
&=-\left(\sum_{j=1}^{s}P_j+\sum_{j=s+1}\overline{P}_j\right)^{-1}\left(\sum_{j=1}^{n}\rho_j\left|x_j-\overline{x}\right|\right)^2\frac{\partial}{\partial x}\left(\frac{x^N-x_1^N}{x-x_1}\varphi(x)\right)/_{x=x_1}\\
&\leq-\left(P_n\max\{s,n-s\}\right)^{-1}\left(\sum_{i=1}^{n}\rho_i\left|x_i-\overline{x}\right|\right)^2\frac{\partial}{\partial x}\left(\frac{x^N-x_1^N}{x-x_1}\varphi(x)\right)/_{x=x_1}\\
&\leq-\left((n-1)P_n\right)^{-1}\left(\sum_{i=1}^{n}\rho_i\left|x_i-\overline{x}\right|\right)^2\frac{\partial}{\partial x}\left(\frac{x^N-x_1^N}{x-x_1}\varphi(x)\right)/_{x=x_1}\leq 0
\end{aligned}$$

holds, unless one of the following two cases occurs:

1) *either $\overline{x}=x_1$ or $\overline{x}=x_n$,*
2) *there exists $k\in\{3,\ldots,n-2\}$ such that $\overline{x}=x_k$ and*

$$P_j\left(x_j-x_{j+1}\right)=0,\quad j=1,\ldots,s-1,\quad \overline{P}_j\left(x_j-x_{j-1}\right)=0,\\ j=s+1,\ldots,n$$

In these two cases

$$\sum_{i=1}^{n} \rho_i \psi\left(x_i\right) - P_n \psi\left(M_{\psi_N}\right) = 0.$$

Case B: Then, for

$$M_{\psi_N} > x_n,$$

we obtain

$$\begin{aligned} & \sum_{i=1}^{n} \rho_i \psi_N\left(x_i\right) - P_n \psi_N\left(M_{\psi_N}\right) \\ & \leq -\left(n P_n\right)^{-1} \left(\sum_{i=1}^{n} \rho_i \left|x_i - M_{\psi_N}\right|\right)^2 \frac{\partial}{\partial x}\left(\frac{x^N - x_1^N}{x - x_1} \varphi\left(x\right)\right) / x = x_1. \end{aligned}$$

Theorem 4, is a refinement of Slater-Pečarić inequality.

3 Hardy Type Inequalities for γ-Quasiconvex Functions

The original Hardy's inequality has a "turning point" (the point where the inequality is reversed) at $p = 1$. One of its versions is:

$$\begin{aligned} & \int_0^b \left(\frac{1}{x}\int_0^x f\left(y\right) dy\right)^p x^{\alpha} dx \\ & \leq \left(\frac{p}{p-\alpha-1}\right)^p \int_0^b f^p\left(x\right) x^{\alpha} \left(1 - \left(\frac{x}{b}\right)^{\frac{p-\alpha-1}{p}}\right) dx \end{aligned} \tag{8}$$

for

$$p \geq 1, \qquad \alpha < p-1, \qquad 0 \leq b \leq \infty$$

or

$$p < 0, \qquad \alpha > p-1, \qquad 0 \leq b \leq \infty.$$

This inequality can be proved directly by the properties of convex functions (The proof can be found in [17] and its references). But by using the γ-quasiconvexity we get a refined variant of the original Hardy's inequality where the turning point is any $p > 1$ (see [8, Theorem 2] and [2, Theorem 12]).

Theorem 5 *Let*

$$p \geq 2, \qquad k > 1, \qquad 0 < b \leq \infty,$$

and $\gamma \in \mathbb{R}_+$, *and let the function* f *be nonnegative and locally integrable on* $(0, b)$. *Then*

$$\begin{aligned}
&\left(\frac{p+\gamma}{k-1}\right)^{p+\gamma} \int_0^b \left[\left(1-\left(\frac{x}{b}\right)^{\frac{k-1}{p+\gamma}}\right) x^{p+\gamma} f^{p+\gamma}(x) - \left(\int_0^x f(t)\,dt\right)^{p+\gamma}\right] \frac{dx}{x^k} \\
\geq &\left(\frac{k-1}{p+\gamma}\right) \int_0^b \int_t^b \left(\left(f(t)\frac{p+\gamma}{k-1}\left(\frac{t}{x}\right)^{1-\frac{k-1}{p+\gamma}}\right)^{\gamma} - \left(\frac{1}{x}\int_0^x f(\sigma)\,d\sigma\right)^{\gamma}\right) \\
&\times \left(\frac{1}{x}\int_0^x f(\sigma)\,d\sigma\right)^{p} x^{\left(1-\frac{k-1}{p+\gamma}\right)(p+\gamma-1)} t^{\frac{k-1}{p+\gamma}-1} \frac{dx}{x^2} dt \\
&+p\left(\frac{k-1}{p+\gamma}\right)^{1-\gamma} \int_0^b \int_t^b \left(f(t)\, t^{1-\frac{k-1}{p+\gamma}}\right)^{\gamma} \left(f(t)\frac{p+\gamma}{k-1}\left(\frac{t}{x}\right)^{1-\frac{k-1}{p+\gamma}}\right. \\
&\left.-\frac{1}{x}\int_0^x f(\sigma)\,d\sigma\right)\left(\frac{1}{x}\int_0^x f(\sigma)\,d\sigma\right)^{p-1} x^{\left(1-\frac{k-1}{p+\gamma}\right)(p+1)} t^{\frac{k-1}{p+\gamma}-1} \frac{dx}{x^2} dt \geq 0
\end{aligned} \tag{9}$$

holds, and when $\gamma = 0$, *inequality (9) coincide with (8).*

4 γ-Quasiconvexity and New Hölder Type Inequalities

In [1] Jensen's type inequalities are used to prove new Hölder type inequalities and reversed Hölder type inequalities, in particular Lemma 3 and the following Lemmas 4 and 5 are used there to get refinements for $p \geq 2$ of Hölder inequality, lower bounds for $1 < p \leq 2$ and upper bounds when $0 < p < 1$:

Lemma 4 ([9, Corollary 1]) *Let* $0 < p \leq 1$, *and let* f *be a* μ*-measurable and positive function on the probability measure space* (μ, Ω) *and*

$$x = \int_\Omega f(s)\,d\mu(s) > 0.$$

Then

$$-I_1 + \left(\int_\Omega f(s)\,d\mu(s)\right)^p \leq \int_\Omega (f(s))^p\,d\mu(s) \leq \left(\int_\Omega f(s)\,d\mu(s)\right)^p,$$

where

$$I_1 = p\left(\int_{\Omega} f(s)\, d\mu(s)\right)^p \left(1 - \int_{\Omega} f(s)\, d\mu(s) \int_{\Omega} (f(s))^{-1}\, d\mu(s)\right) > 0.$$

Lemma 5 ([9, Corollary 2]) *Let* $0 < p \le 1$,*let* f *be a non-negative* μ*-measurable function on the probability measure space* (Ω, μ)*and*

$$x = \int_{\Omega} f(s)\, d\mu(s) > 0.$$

Then

$$-I_2 + \left(\int_{\Omega} f(s)\, d\mu(s)\right)^p \le \int_{\Omega} (f(s))^p\, d\mu(s) \le \left(\int_{\Omega} f(s)\, d\mu(s)\right)^p, \qquad (10)$$

where

$$I_2 = p\left(\int_{\Omega} f(s)\, d\mu(s)\right)^{p-1} \int_{\Omega} \frac{(f(s) - x)^2}{f(s)} d\mu(s). \qquad (11)$$

From Lemma 3 we get that for the 1-quasiconvex functions

$$\varphi(x) = x^p, \qquad x \ge 0, \qquad p \ge 2$$

the inequality

$$\int_{\Omega} (f(s))^p\, d\mu(s) - \left(\int_{\Omega} f(s)\, d\mu(s)\right)^p \qquad (12)$$
$$\ge (p-1)\left(\int_{\Omega} f(s)\, d\mu(s)\right)^{p-2} \int_{\Omega} \left(f(s) - \int_{\Omega} f(s)\, d\mu(s)\right)^2 d\mu(s)$$

holds.

By using (12) we get:

Theorem 6 *Let* $p \ge 2$ *and define* q *by* $\frac{1}{p} + \frac{1}{q} = 1$. *Then for any two nonnegative* ν*-measurable functions* f *and* g

$$\int_{\Omega} fg d\nu \qquad (13)$$
$$\le \left(\int_{\Omega} f^p d\nu - (p-1)\left(\frac{\int_{\Omega} fg d\nu}{\int_{\Omega} g^q d\nu}\right)^{p-2} \int_{\Omega} \left(fg^{(1-q)} - \frac{\int_{\Omega} fg d\nu}{\int_{\Omega} g^q d\nu}\right)^2 g^q d\nu\right)^{\frac{1}{p}}$$
$$\times \left(\int_{\Omega} g^q d\nu\right)^{\frac{1}{q}}.$$

If $1 < p \leq 2$ *we get when*

$$\int_\Omega f^p d\nu \geq (p-1)\left(\frac{\int_\Omega fgd\nu}{\int_\Omega g^q d\nu}\right)^{p-2} \int_\Omega \left(fg^{(1-q)} - \frac{\int_\Omega fgd\nu}{\int_\Omega g^q d\nu}\right)^2 g^q d\nu,$$

that

$$\begin{aligned} &\left(\int_\Omega f^p d\nu\right)^{\frac{1}{p}} \left(\int_\Omega g^q d\nu\right)^{\frac{1}{q}} \\ &\geq \int_\Omega fgd\nu \\ &\geq \left(\int_\Omega f^p d\nu - (p-1)\left(\frac{\int_\Omega fgd\nu}{\int_\Omega g^q d\nu}\right)^{p-2} \int_\Omega \left(fg^{(1-q)} - \frac{\int_\Omega fgd\nu}{\int_\Omega g^q d\nu}\right)^2 g^q d\nu\right)^{\frac{1}{p}} \\ &\quad \times \left(\int_\Omega g^q d\nu\right)^{\frac{1}{q}}. \end{aligned} \tag{14}$$

The last inequalities emphasize that through the 1*-quasiconvexity and* 1*-quasiconcavity notions we get refined Hölder inequality for* $p \geq 2$ *in (13) and a lower bound in (14) for* $1 < p \leq 2$.

Similar results appear in [5] and [18], there by using another variant of convex functions called superquadratic functions.

From Lemma 5 we get a two sided Hölder type inequality:

Theorem 7 *Let* $0 < p \leq 1$*, f and g be non-negative μ-measurable functions on the probability measure space* (Ω, ν) *then*

$$\begin{aligned} &\left(\int_\Omega f^p d\nu\right)^{\frac{1}{p}} \left(\int_\Omega g^q d\nu\right)^{\frac{1}{q}} \leq \int_\Omega fgd\nu \\ &\leq \left(\int_\Omega f^p d\nu + p\left(\frac{\int_\Omega fgd\nu}{\int_\Omega g^q d\nu}\right)^{p-1} \int_\Omega \left(fg^{(1-q)} - \frac{\int_\Omega fgd\nu}{\int_\Omega g^q d\nu}\right)^2 \frac{g^{2q-1}}{f} d\nu\right)^{\frac{1}{p}} \\ &\quad \times \left(\int_\Omega g^q d\nu\right)^{\frac{1}{q}}. \end{aligned} \tag{15}$$

Similarly we get from Lemma 4 that

Theorem 8 *Let* $0 < p \leq 1$*, f and g be non-negative μ-measurable functions on the probability measure space* (Ω, ν)*, then*

$$\int_\Omega fgd\nu$$

$$\leq \left(\int_\Omega f^p d\nu + p\left(\frac{\int_\Omega fgd\nu}{\int_\Omega g^q d\nu}\right)^p \left(\int_\Omega g^q d\nu - \frac{\int_\Omega fgd\nu}{\int_\Omega g^q d\nu}\int_\Omega \frac{g^{2q-1}}{f} d\nu\right)\right)^{\frac{1}{p}}$$
$$\times \left(\int_\Omega g^q d\nu\right)^{\frac{1}{q}}.$$

Hölder type inequality for $0 < p \leq \frac{1}{2}$ and for $\frac{1}{2} \leq p < 1$ which we state now, are derived again from the theorems related to 1-quasiconvex functions but are obtained by different substitutions than those employed in the proof of Theorems 6–8.

Theorem 9 *Let* $0 < p \leq \frac{1}{2}$ *and define* $\frac{1}{p} + \frac{1}{q} = 1$. *Then for any positive* ν-*measurable function* f *and* g

$$\int_\Omega fgd\nu \geq \left(\int_\Omega f^p d\nu\right)^{\frac{1}{p}} \left(\int_\Omega g^q d\nu\right)^{\frac{1}{q}} \tag{16}$$
$$\times \left[1 + \left(\frac{1}{p} - 1\right)\int_\Omega \left(\frac{f^p \int_\Omega g^q d\nu - g^q \int_\Omega f^p d\nu}{\int_\Omega f^p d\nu}\right)^2 \frac{g^{-q}}{\int_\Omega g^q d\nu} d\nu\right]$$

is derived, which is a refinement of Hölder inequality.

For $\frac{1}{2} \leq p < 1$, *we get the reverse of inequality (16) and together with Hölder inequality for* $0 < p < 1$

$$\left(\int_\Omega g^q d\nu\right)^{\frac{1}{q}} \left(\int_\Omega f^p d\nu\right)^{\frac{1}{p}} \leq \int_\Omega fgd\nu \leq \left(\int_\Omega g^q d\nu\right)^{\frac{1}{q}} \left(\int_\Omega f^p d\nu\right)^{\frac{1}{p}}$$
$$\times \left[1 + \left(\frac{1}{p} - 1\right)\int_\Omega \left(\frac{f^p \int_\Omega g^q d\nu - g^q \int_\Omega f^p d\nu}{\int_\Omega f^p d\nu}\right)^2 \frac{g^{-q}}{\int_\Omega g^q d\nu} d\nu\right] \tag{17}$$

is derived.

5 Minkowski Type Inequalities Using 1-Quasiconvexity

By using Theorem 6 in [1] we get Minkowski type inequalities:

Theorem 10 *Let* $p \geq 2$ *and let* $\frac{1}{q} = 1 - \frac{1}{p}$. *Then for any two non-negative* ν-*measurable functions* f *and* g

$$\left(\int (f+g)^p d\nu\right)^{\frac{1}{p}}$$
$$\leq \left(\int f^p d\nu - D\left(\int f(f+g)^{p-1} d\nu\right)^{p-2}\right)^{\frac{1}{p}}$$

$$+\left(\int g^{p}d\nu - D\left(\int g\,(f+g)^{p-1}\,d\nu\right)^{p-2}\right)^{\frac{1}{p}}$$

where

$$D=(p-1)\left(\int\left(\frac{\left(g\int f\,(f+g)^{p-1}\,d\nu - f\int g\,(f+g)^{p-1}\,d\nu\right)^{2}(f+g)^{p-2}}{\left(\int (f+g)^{p}\,d\nu\right)^{p}}\right)d\nu\right).$$

The following Theorem 11 follows from inequality (14) and is proved in [1].

Theorem 11 *Let $1 < p \leq 2$ and let $\frac{1}{q} = 1 - \frac{1}{p}$. Then for any two non-negative ν-measurable functions f and g*

$$\left(\int f^{p}d\nu\right)^{\frac{1}{p}}+\left(\int g^{p}d\nu\right)^{\frac{1}{p}} \geq \left(\int (f+g)^{p}\,d\nu\right)^{\frac{1}{p}}$$

$$\geq \left(\int f^{p}d\nu - D\left(\int f\,(f+g)^{p-1}\,d\nu\right)^{p-2}\right)^{\frac{1}{p}}$$

$$+\left(\int g^{p}d\nu - D\left(\int g\,(f+g)^{p-1}\,d\nu\right)^{p-2}\right)^{\frac{1}{p}}$$

where

$$D=(p-1)\int\left(\frac{\left(g\int f\,(f+g)^{p-1}\,d\nu - f\int g\,(f+g)^{p-1}\,d\nu\right)^{2}(f+g)^{p-2}}{\left(\int (f+g)^{p}\,d\nu\right)^{p}}\right)d\nu.$$

and

$$\int f^{p}d\nu \geq D\left(\int f\,(f+g)^{p-1}\,d\nu\right)^{p-2},\ \int g^{p}d\nu \geq D\left(\int g\,(f+g)^{p-1}\,d\nu\right)^{p-2}.$$

We now quote from [1] Minkowski's type inequalities when $0 < p \leq \frac{1}{2}$ and when $\frac{1}{2} \leq p < 1$.

Theorem 12 *Let $0 < p \leq \frac{1}{2}$ and define $\frac{1}{p} + \frac{1}{q} = 1$. Then for any two non-negative ν-measurable functions f and g*

$$\left(\int (f+g)^{p}\,d\nu\right)^{\frac{1}{p}}$$

$$\geq \left(\int f^p d\nu\right)^{\frac{1}{p}}$$
$$\times\left[1+\left(\frac{1}{p}-1\right)\int\left(\frac{(f+g)^p\int f^p d\nu - f^p\int (f+g)^p\,d\nu}{\int f^p d\nu}\right)^2\frac{(f+g)^{-p}}{\int (f+g)^p\,d\nu}d\nu\right]$$
$$+\left(\int g^p d\nu\right)^{\frac{1}{p}}$$
$$\times\left[1+\left(\frac{1}{p}-1\right)\int\left(\frac{(f+g)^p\int g^p d\nu - g^p\int (f+g)^p\,d\nu}{\int g^p d\nu}\right)^2\frac{(f+g)^{-p}}{\int (f+g)^p\,d\nu}d\nu\right].$$

When $\frac{1}{2}\leq p<1$ *we get*

$$\left(\int f^p d\nu\right)^{\frac{1}{p}}+\left(\int g^p d\nu\right)^{\frac{1}{p}}\leq\left(\int (f+g)^p\,d\nu\right)^{\frac{1}{p}}$$

$$\leq \left(\int f^p d\nu\right)^{\frac{1}{p}}$$
$$\times\left[1+\left(\frac{1}{p}-1\right)\int\left(\frac{(f+g)^p\int f^p d\nu - f^p\int (f+g)^p\,d\nu}{\int f^p d\nu}\right)^2\frac{(f+g)^{-p}}{\int (f+g)^p\,d\nu}d\nu\right]$$
$$+\left(\int g^p d\nu\right)^{\frac{1}{p}}$$
$$\times\left[1+\left(\frac{1}{p}-1\right)\int\left(\frac{(f+g)^p\int g^p d\nu - g^p\int (f+g)^p\,d\nu}{\int g^p d\nu}\right)^2\frac{(f+g)^{-p}}{\int (f+g)^p\,d\nu}d\nu\right].$$

6 Bounds of “Jensen’s Gap” for N-quasiconvex Functions

6.1 Bounds for Difference Between Two “Jensen’s Gaps” for N-quasiconvex Functions

We state here one of many results that can be derived from the previous theorems. First we quote a result from [10] about the difference between two “Jensen’s gaps”

$$\sum_{i=1}^{n} p_i\psi\left(x_i\right)-\psi\left(\overline{x}_p\right)$$

and

$$\sum_{i=1}^{n} q_i \psi\left(x_i\right) - \psi\left(\overline{x}_q\right).$$

Then we present a new theorem proved in [1] with results when ψ_N is a N-quasiconvex function. In particular for a 1-quasiconvex function ψ_1 the result is interesting.

These results are refinements of the following theorem by Dragomir in [10] (see also [3]):

Theorem 13 *Let*

$$x_i \in I, \quad i = 1, \ldots, n, \quad \overline{x}_p = \sum_{i=1}^{n} p_i x_i, \quad p_i \geq 0, \quad i = 1, \ldots, n, \quad \sum_{i=1}^{n} p_i = 1$$

and

$$\overline{x}_q = \sum_{i=1}^{n} q_i x_i, \quad q_i > 0, \quad i = 1, \ldots, n, \quad \sum_{i=1}^{n} q_i = 1,$$

$$m = \min_{1 \leq i \leq n} \left(\frac{p_i}{q_i}\right) \quad \text{and} \quad M = \max_{1 \leq i \leq n} \left(\frac{p_i}{q_i}\right).$$

If ψ is convex then

$$M\left(\sum_{i=1}^{n} q_i \psi\left(x_i\right) - \psi\left(\overline{x}_q\right)\right) \geq \sum_{i=1}^{n} p_i \psi\left(x_i\right) - \psi\left(\overline{x}_p\right) \tag{18}$$
$$\geq m\left(\sum_{i=1}^{n} q_i \psi\left(x_i\right) - \psi\left(\overline{x}_q\right)\right).$$

Now we state a refinement of Theorem 13 for N-quasiconvex function ψ_N.

Theorem 14 ([1, Theorem 18]) *Suppose that $\psi_N : [a, b) \to \mathbb{R}$, $0 \leq a < b \leq \infty$ is N-quasiconvex function, that is $\psi_N = x^N \varphi(x)$, $N = 1, 2, \ldots$ where φ is convex on $[a, b)$.*

Let

$$x_i \in I, \quad i = 1, \ldots, n, \quad \overline{x}_p = \sum_{i=1}^{n} p_i x_i, \quad p_i \geq 0, \quad i = 1, \ldots, n, \quad \sum_{i=1}^{n} p_i = 1$$

and

$$\overline{x}_q = \sum_{i=1}^{n} q_i x_i, \quad q_i > 0, \quad i = 1, \ldots, n, \quad \sum_{i=1}^{n} q_i = 1.$$

Then, for $m=\min\limits_{1\le i\le n}\left(\frac{p_i}{q_i}\right)$ *and* $M=\max\limits_{1\le i\le n}\left(\frac{p_i}{q_i}\right)$

$$\left(\sum_{i=1}^{n} p_i\psi_N\left(x_i\right)-\psi_N\left(\overline{x}_p\right)\right)-m\left(\sum_{i=1}^{n} q_i\psi_N\left(x_i\right)-\psi_N\left(\overline{x}_q\right)\right) \tag{19}$$
$$\ge\sum_{i=1}^{n}\left(p_i-mq_i\right)\left(x_i-\overline{x}_p\right)^2\frac{\partial}{\partial\overline{x}_p}\left(\frac{x_i^N-\overline{x}_p^N}{x_i-\overline{x}_p}\varphi\left(\overline{x}_p\right)\right)$$
$$+m\left(\overline{x}_q-\overline{x}_p\right)^2\left(\frac{\overline{x}_q^N-\overline{x}_p^N}{\overline{x}_q-\overline{x}_p}\varphi\left(\overline{x}_p\right)\right),$$

and

$$\left(\sum_{i=1}^{n} p_i\psi_N\left(x_i\right)-\psi_N\left(\overline{x}_p\right)\right)-M\left(\sum_{i=1}^{n} q_i\psi_N\left(x_i\right)-\psi_N\left(\overline{x}_q\right)\right) \tag{20}$$
$$\le\sum_{i=1}^{n}\left(p_i-Mq_i\right)\left(x_i-\overline{x}_q\right)^2\frac{\partial}{\partial\overline{x}_q}\left(\frac{x_i^N-\overline{x}_q^N}{x_i-\overline{x}_q}\varphi\left(\overline{x}_q\right)\right)$$
$$-M\left(\overline{x}_q-\overline{x}_p\right)^2\frac{\partial}{\partial\overline{x}_q}\left(\frac{\overline{x}_q^N-\overline{x}_p^N}{\overline{x}_q-\overline{x}_p}\varphi\left(\overline{x}_q\right)\right).$$

For $N=1$ *we get that*

$$\left(\sum_{i=1}^{n} p_i\psi_1\left(x_i\right)-\psi_1\left(\overline{x}_p\right)\right)-m\left(\sum_{i=1}^{n} q_i\psi_1\left(x_i\right)-\psi_1\left(\overline{x}_q\right)\right) \tag{21}$$
$$\ge\varphi^{'}\left(\overline{x}_p\right)\left(\left(\sum_{i=1}^{n} p_ix_i^2-\left(\overline{x}_p\right)^2\right)-m\left(\sum_{i=1}^{n} q_ix_i^2-\left(\overline{x}_q\right)^2\right)\right),$$

and

$$\left(\sum_{i=1}^{n} p_i\psi_1\left(x_i\right)-\psi_1\left(\overline{x}_p\right)\right)-M\left(\sum_{i=1}^{n} q_i\psi_1\left(x_i\right)-\psi_1\left(\overline{x}_q\right)\right) \tag{22}$$
$$\le\varphi^{'}\left(\overline{x}_q\right)\left(\left(\sum_{i=1}^{n} p_ix_i^2-\left(\overline{x}_p\right)^2\right)-M\left(\sum_{i=1}^{n} q_ix_i^2-\left(\overline{x}_q\right)^2\right)\right).$$

In particular if φ *is also non-negative increasing then (19)–(22) are refinements of (18).*

6.2 Jensen Gap and Taylor Power Series

In [4] we get by the γ-quasiconvexity technique an estimation of Jensen Gap, in particular, for functions that have Taylor power series representation, which we state in this section.

Theorem 15 ([4, Theorem 1]) *Let $\phi : [0, A) \to \mathbb{R}$ have the Taylor power series representation on $[0, A)$,*

$$0 < A \leq \infty : \phi(x) = \sum_{n=0}^{\infty} a_n x^n.$$

Let φ be a convex increasing function on $[0, A)$ that is related to ϕ by

$$\varphi(x) = \frac{\phi(x) - \phi(0)}{x} = \sum_{n=0}^{\infty} a_{n+1} x^n.$$

If $f \geq 0$ and f, f^2 and $\phi \circ f$ are integrable functions on Ω and

$$z = \int_{\Omega} f d\mu > 0,$$

where μ is a probability measure on Ω, then:

a)

$$\int_{\Omega} \phi(f) d\mu - \phi(z) \geq \left(\frac{\phi(z) - \phi(0)}{z} \right)' \left(\int_{\Omega} f^2 d\mu - z^2 \right) \geq 0.$$

In other words:

$$\begin{aligned} J(\phi, \mu, f) &= \int_{\Omega} \phi(f) d\mu - \phi(z) \\ &= \sum_{n=0}^{\infty} a_{n+1} \int_{\Omega} f^{n+1} d\mu - \sum_{n=0}^{\infty} a_{n+1} z^{n+1} \\ &\geq \sum_{n=0}^{\infty} (n+1) a_{n+2} z^n \left(\int_{\Omega} f^2 d\mu - z^2 \right) \geq 0. \end{aligned}$$

b) For

$$\overline{x} = \sum_{i=1}^{m} \alpha_i x_i, \quad \sum_{i=1}^{m} \alpha_i = 1, \quad 0 \leq \alpha_i \leq 1, \quad 0 \leq x_i < A, \quad i = 1, \ldots, m,$$

it yields that

$$\sum_{i=1}^{m}\alpha_i\phi\left(x_i\right)-\phi\left(\overline{x}\right)\geq\left(\frac{\phi\left(\overline{x}\right)-\phi(0)}{\overline{x}}\right)'\left(\sum_{i=1}^{m}\alpha_i x_i^2-\overline{x}^2\right)\geq 0.$$

In other words,

$$\sum_{i=1}^{m}\sum_{n=0}^{\infty}\alpha_i a_{n+1}x_i^{n+1}-\sum_{n=0}^{\infty}a_{n+1}\overline{x}^{n+1}\geq\sum_{n=0}^{\infty}(n+1)\,a_{n+2}\overline{x}^n\left(\sum_{i=1}^{m}\alpha_i x_i^2-\overline{x}^2\right)\geq 0.$$

Corollary 1 ([4, Corollary 3]) *Let* $0<A\leq\infty$ *and let* $\phi:[0,A)$ *have a Taylor expansion*

$$\phi\left(x\right)=\sum_{n=0}^{\infty}a_n x^n,$$

on $[0,A)$ *. If*

$$\overline{x}=\sum_{i=1}^{m}\alpha_i x_i,\quad \sum_{i=1}^{m}\alpha_i=1,\quad 0\leq\alpha_i\leq 1,\quad 0\leq x_i\leq A,\quad i=1.2,\ldots m,$$

then

$$J=\sum_{i=1}^{m}\alpha_i\phi\left(x_i\right)-\phi\left(\overline{x}\right)=\sum_{n=2}^{\infty}a_n\left(\sum_{i=1}^{m}\alpha_i x_i^2-\overline{x}^2\right)\sum_{k=1}^{n-1}\left(n-k\right)x^{k-1}\overline{x}^{n-k-1}.$$

Concluding Remark In this survey we show refinements and extensions of important type inequalities related to convexity. In future survey we will show other refinements of important inequalities for instance the Hermite-Hadamard and the Fejer inequalities.

References

1. S. Abramovich, Jensen, Hölder, Minkowski, Jensen-Steffensen and Slater-Pečarić inequalities derived through N-quasiconvexity. Math. Inequal. Appl. **19**(4), 1203–1226 (2016)
2. S. Abramovich, Applications of quasiconvexity, in *Contributions in Mathematics and Engineering in Honor of Constantin Caratheodori* (Springer, Basel, 2016), pp. 1–23
3. S. Abramovich, S.S. Dragomir, Normalized Jensen functional, superquadracity and related inequalities, in *Inequalities and Applications*, ed. by C. Bandle, L. Losonczi, A. Gilányi, Z. Páles, M. Plum. International Series of Numerical Mathematics, vol. 157 (Birkhäuser Verlag, Basel, 2008), pp. 217–228

4. S. Abramovich, L.-E. Persson, Some new estimates of the “Jensen Gap”. J. Inequal. Appl. **2016**(39), 9 (2016)
5. S. Abramovich, G. Jameson, G. Sinnamon, Refining Jensen’s Inequality. Bull. Math. Soc. Sci. Math. Roumanie (N.S) **47**(95), 3–14 (2004)
6. S. Abramovich, M. Klaricić Bacula, M. Matić, J. Pečarić, A variant of Jensen-Steffensen’s inequality and quazi arithmetic means. J. Math. Anal. Appl. **307**, 370–386 (2005)
7. S. Abramovich, S. Banić, M. Matić, J. Pečarić, Jensen Steffensen’s and related inequalities, for superquadratic functions. Math. Inequal. Appl. **11**, 23–41 (2008)
8. S. Abramovich, L.-E. Persson, N. Samko, Some new scales of refined Jensen and Hardy type inequalities. Math. Inequal. Appl. **17**, 1105–1114 (2014)
9. S. Abramovich, L.-E. Persson, N. Samko, On γ-quasiconvexity, superquadracity and two-sided reversed Jensen type inequalities. Math. Inequal. Appl. **18**(2), 615–627 (2015)
10. S.S. Dragomir, Bounds for the normalised Jensen functional. Bull. Aust. Math. Soc. **74**, 471–478 (2006)
11. G.H. Hardy, J.E. Littlewood, G. Pólya, *Inequalities* (Cambridge University Press, Cambridge, 1964)
12. E.G. Kwon, J.E. Bae, On a refined Hölder’s inequality. J. Math. Inequal. **10**(1), 261–268 (2016)
13. J. Matkowski, A converse of Hölder inequality theorem. Math. Inequal. Appl. **12**(1), 21–32 (2009)
14. C. Niculescu, L.-E. Persson, *Convex Functions and Their Applications, a Contemporary Approach*. CMS Books in Mathematics, vol. 23 (Springer, New York, 2006)
15. L. Nikolova, S. Varošanec, Refinement of Hölder’s inequality derived from functions $\psi_{p,q,\lambda}$ and $\phi_{p,q,\lambda}$. Ann. Funct. Anal **2**(1), 72–83 (2011)
16. J. Pečarić, F. Proschan, Y.L. Tong, *Convex Functions, Partial Orderings, and Statistical Applications* (Academic, New York, 1992)
17. L.-E. Persson, N. Samko, What should have happened if Hardy had discovered this? J. Inequal. Appl. **29**(11) (2012)
18. G. Sinnamon, Refining the Hölder and Minkowski inequalities. J. Inequal. Appl. **6**, 633–640 (2001)
19. J.-F. Tian, Property of Hölder-type inequality and its application. Math. Inequal. Appl. **16**(3), 831–841 (2013)

Criteria for Convergence of Iterates in a Compression-Expansion Fixed Point Theorem of Functional Type

Richard I. Avery, Douglas R. Anderson, and Johnny Henderson

1 Introduction

Fixed point theorems have widely been used to verify the existence of solutions to boundary value problems [1, 4, 6–9, 11, 13]. There are results, for example Petryshyn [12], that culminate in a solution to a boundary value problem; however, the difficulty in applying these theorems lies in the invariance assumptions. Intervals of functional type can be used to narrow the underlying set used with a fixed point theorem in such a way that once a unique fixed point is established in an interval of functional type, one can employ k-contraction principles to iterate to the solution in the interval of functional type. Intervals of functional type which were introduced in the extension of the compression-expansion fixed point theorem of functional type [3] provide a means to narrow the search for a fixed point using the properties of the operator. We conclude with an application that demonstrates some conditions that can be included in a functional type interval that have not been used in existence of solutions arguments in the past.

R. I. Avery
College of Arts and Sciences, Dakota State University, Madison, SD, USA
e-mail: rich.avery@dsu.edu

D. R. Anderson (✉)
Concordia College, Department of Mathematics, Moorhead, MN, USA
e-mail: andersod@cord.edu

J. Henderson
Department of Mathematics, Baylor University, Waco, TX, USA
e-mail: Johnny_Henderson@baylor.edu

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization
and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_2

2 Preliminaries

For completeness we provide the following definitions and theorems which are nearly identical to the presentation in other compression-expansion fixed point papers, in particular [2].

Definition 1 Let E be a real Banach space. A nonempty closed convex set $P \subset E$ is called a *cone* if for all $x \in P$ and $\lambda \geq 0$, $\lambda x \in P$ and if $x, -x \in P$ then $x = 0$.

Every cone $P \subset E$ induces an ordering in E given by $x \leq y$ if and only if $y - x \in P$.

Definition 2 An operator is called completely continuous if it is continuous and maps bounded sets into precompact sets.

Definition 3 A map α is said to be a nonnegative continuous concave functional on a cone P of a real Banach space E if $\alpha : P \to [0, \infty)$ is continuous and

$$\alpha(tx + (1-t)y) \geq t\alpha(x) + (1-t)\alpha(y)$$

for all $x, y \in P$ and $t \in [0, 1]$. Similarly we say the map β is a nonnegative continuous convex functional on a cone P of a real Banach space E if $\beta : P \to [0, \infty)$ is continuous and

$$\beta(tx + (1-t)y) \leq t\beta(x) + (1-t)\beta(y)$$

for all $x, y \in P$ and $t \in [0, 1]$.

Definition 4 Let A be an open subset of a cone P, a and b be nonnegative numbers, α be a concave functional on P, and β be a convex functional on P. Then the set

$$A(\beta, b, \alpha, a) = \{x \in A \ : \ a < \alpha(x) \ \text{ and } \ \beta(x) < b\}$$

is an interval of functional type.

Definition 5 Let D be a subset of a real Banach space E. If $r : E \to D$ is continuous with $r(x) = x$ for all $x \in D$, then D is a *retract* of E, and the map r is a *retraction*.

Dugundji's Theorem, which is stated below, is applied to the cone in our main result so the fixed point index can be applied; a proof can be found in [5, p. 44]. The *convex hull* of a subset D of a real Banach space X is given by

$$conv(D) = \left\{ \sum_{i=1}^{n} \lambda_i x_i : x_i \in D, \ \lambda_i \in [0, 1], \ \sum_{i=1}^{n} \lambda_i = 1, \ \text{and } n \in \mathbb{N} \right\}.$$

Theorem 1 *For Banach spaces X and Y, let $D \subset X$ be closed and let $F : D \to Y$ be continuous. Then F has a continuous extension $\tilde{F} : X \to Y$ such that $\tilde{F}(X) \subset \overline{conv(F(D))}$.*

Corollary 1 *Every closed convex set of a Banach space is a retract of the Banach space.*

The proof of our fixed point theorem relies on properties of the fixed point index, which are stated below; a proof can be found in [5, p. 238].

Theorem 2 *Let X be a retract of a real Banach space E. Then, for every bounded relatively open subset U of X and every completely continuous operator $A : \overline{U} \to X$ which has no fixed points on ∂U (relative to X), there exists an integer $i(A, U, X)$ satisfying the following conditions:*

(i) Normality: $i(A, U, X) = 1$ if $Ax \equiv y_0 \in U$ for any $x \in \overline{U}$;
(ii) Additivity: $i(A, U, X) = i(A, U_1, X) + i(A, U_2, X)$ whenever U_1 and U_2 are disjoint open subsets of U such that A has no fixed points on $\overline{U} - (U_1 \cup U_2)$;
(iii) Homotopy Invariance: $i(H(t, \cdot), U, X)$ is independent of $t \in [0, 1]$ whenever $H : [0, 1] \times \overline{U} \to X$ is completely continuous and $H(t, x) \neq x$ for any $(t, x) \in [0, 1] \times \partial U$;
(iv) Solution: If $i(A, U, X) \neq 0$, then A has at least one fixed point in U.

Moreover, $i(A, U, X)$ is uniquely defined.

In the following theorem we extend the compression-expansion fixed point theorem [3] that utilizes intervals of functional type with a condition for iterates of an operator T to converge to a unique fixed point. In the spirit of the Leggett-Williams fixed point theorem [10], which many of the compression-expansion fixed point theorems of functional type have generalized, we do not know that T is invariant on $A(\beta, b, \alpha, a)$. However, if suitable k-contractive conditions are met we can still use the same arguments as presented in the Banach fixed point theorem to prove that iterates converge and that the fixed point of T in $A(\beta, b, \alpha, a)$ is unique. See [14, p. 17] for a presentation of these concepts; one can also see these techniques in the work of Petryshyn [12]. The key is to first prove that there is a unique fixed point in $A(\beta, b, \alpha, a)$, and then to show under additional conditions that the iterates will converge to this fixed point.

For completeness and clarification, the entire proof of the extension is presented below, as there was a hidden assumption in the original paper [3], namely A being convex; moreover, x_0 also needed to lie in the set $A(\theta, c, \psi, d)$, and the assumptions that were made on the set $\partial A(\beta, b, \alpha, a)$ only needed to be made on the set $\overline{A} \cap \partial P(\beta, b, \alpha, a)$.

Theorem 3 *Suppose P is a cone in a real Banach space E, A is a relatively open subset of P, α and ψ are nonnegative continuous concave functionals on P, β and θ are nonnegative continuous convex functionals on P, and $T : P \to P$ is a completely continuous operator. If there exist nonnegative numbers $a, b, c,$ and d and $x_0 \in A(\beta, b, \alpha, a) \cap A(\theta, c, \psi, d)$ such that:*

$(A0)$ *$A(\beta, b, \alpha, a)$ is bounded;*
$(A1)$ *if $x \in \partial A \cap \overline{P(\beta, b, \alpha, a)}$ and $\tau \in [0, 1]$, then $(1-\tau)Tx + \tau x_0 \neq x$;*
$(A2)$ *if $x \in \overline{A} \cap \partial P(\beta, b, \alpha, a)$ with $\alpha(x) = a$ and $\theta(x) \leq c$, then $\alpha(Tx) > a$;*
$(A3)$ *if $x \in \overline{A} \cap \partial P(\beta, b, \alpha, a)$ with $\alpha(x) = a$ and $\theta(Tx) > c$, then $\alpha(Tx) > a$;*
$(A4)$ *if $x \in \overline{A} \cap \partial P(\beta, b, \alpha, a)$ with $\beta(x) = b$ and $\psi(Tx) < d$, then $\beta(Tx) < b$;*
$(A5)$ *if $x \in \overline{A} \cap \partial P(\beta, b, \alpha, a)$ with $\beta(x) = b$ and $\psi(x) \geq d$, then $\beta(Tx) < b$;*

then T has a fixed point $x_* \in A(\beta, b, \alpha, a)$. Moreover, if for all $x \in A(\beta, b, \alpha, a)$ there exists a $k \in [0, 1)$ such that

$$\|Tx - x_*\| \leq k\|x - x_*\|,$$

then x_* is the unique fixed point of T in $A(\beta, b, \alpha, a)$. Furthermore, if $y_0 \in A(\beta, b, \alpha, a)$ and

$$\|T^n y_0 - x_*\| \leq k\|T^{n-1} y_0 - x_*\|$$

for all positive integers n, then

$$T^n y_0 \to x_*.$$

Proof By Corollary 1, P is a retract of the Banach space E since it is convex and closed.
Since the functional interval satisfies

$$A(\beta, b, \alpha, a) = A \cap P(\beta, b, \alpha, a),$$

we have that

$$\begin{aligned}
&\partial A(\beta, b, \alpha, a)\\
&= \partial(A \cap P(\beta, b, \alpha, a))\\
&= \overline{(A \cap P(\beta, b, \alpha, a))} \cap \overline{(P - (A \cap P(\beta, b, \alpha, a)))}\\
&= \overline{(A \cap P(\beta, b, \alpha, a))} \cap \overline{(P - A) \cup (P - P(\beta, b, \alpha, a))}\\
&= \overline{(A \cap P(\beta, b, \alpha, a))} \cap (\overline{(P - A)} \cup \overline{(P - P(\beta, b, \alpha, a))})\\
&\subseteq (\overline{A} \cap \overline{P(\beta, b, \alpha, a)}) \cap (\overline{(P - A)} \cup \overline{(P - P(\beta, b, \alpha, a))})\\
&= (\overline{A} \cap \overline{P(\beta, b, \alpha, a)} \cap \overline{(P - A)}) \cup (\overline{A} \cap \overline{P(\beta, b, \alpha, a)} \cap \overline{(P - P(\beta, b, \alpha, a))})\\
&= (\partial A \cap \overline{P(\beta, b, \alpha, a)}) \cup (\overline{A} \cap \partial P(\beta, b, \alpha, a)).
\end{aligned}$$

Define $H : [0, 1] \times \overline{A(\beta, b, \alpha, a)} \to P$ by

$$H(\tau, x) = (1-\tau)Tx + \tau x_0.$$

Clearly, H is continuous and $H\left([0,1]\times\overline{A(\beta,b,\alpha,a)}\right)$ is precompact.

Claim $H(\tau,x)\neq x$ for all $(\tau,x)\in[0,1]\times\partial A(\beta,b,\alpha,a)$.

Suppose not; that is, suppose there exists $(t_0,y_0)\in[0,1]\times\partial A(\beta,b,\alpha,a)$ such that

$$H(t_0,y_0)=y_0.$$

Since $y_0\in\partial A(\beta,b,\alpha,a)$, we have that

$$y_0\in\overline{A}\cap\partial P(\beta,b,\alpha,a),$$

so either $\beta(y_0)=b$ or $\alpha(y_0)=a$, as

$$\partial A(\beta,b,\alpha,a)\subseteq\left(\partial A\cap\overline{P(\beta,b,\alpha,a)}\right)\cup\left(\overline{A}\cap\partial P(\beta,b,\alpha,a)\right)$$

and $H(\tau,x)=(1-\tau)Tx+\tau x_0\neq x$ for all $(\tau,x)\in[0,1]\times\partial A\cap\overline{P(\beta,b,\alpha,a)}$ by condition $(A1)$.

Case 2.1 $\beta(y_0)=b$.
Either $\psi(Ty_0)<d$ or $\psi(Ty_0)\geq d$.

Subcase 2.1.1 $\psi(Ty_0)<d$.
By condition (A4) we have $\beta(Ty_0)<b$, thus it follows that

$$b=\beta(y_0)=\beta\left((1-t_0)Ty_0+t_0x_0\right)\leq(1-t_0)\beta(Ty_0)+t_0\beta(x_0)<b,$$

which is a contradiction.

Subcase 2.1.2 $\psi(Ty_0)\geq d$.
Since $x_0\in A(\theta,c,\psi,d)$, $\psi(x_0)>d$, hence we have that $\psi(y_0)\geq d$ because

$$\psi(y_0)=\psi((1-t_0)Ty_0+t_0x_0)\geq(1-t_0)\psi(Ty_0)+t_0\psi(x_0)\geq d,$$

and thus by condition (A5) we have $\beta(Ty_0)<b$, which is the same contradiction we arrived at in the previous subcase.

Case 2.2 $\alpha(y_0)=a$.
Either $\theta(Ty_0)\leq c$ or $\theta(Ty_0)>c$.

Subcase 2.2.1 $\theta(Ty_0)>c$.
By condition (A3) we have $\alpha(Ty_0)>a$, thus we have

$$a=\alpha(y_0)=\alpha((1-t_0)Ty_0+t_0x_0)\geq(1-t_0)\alpha(Ty_0)+t_0\alpha(x_0)>a,$$

which is a contradiction.

Subcase 2.2.2 $\theta(Tx_0)\leq c$.

Since $x_0 \in A(\theta, c, \psi, d)$, $\theta(x_0) < c$, hence we have that $\theta(y_0) \leq c$ because

$$\theta(y_0) = \theta((1 - t_0)Ty_0 + t_0 x_0) \leq (1 - t_0)\theta(Ty_0) + t_0\theta(x_0) \leq c,$$

and thus by condition (A2) we have $\alpha(Ty_0) > a$, which is the same contradiction we arrived at in the previous case.

Therefore, we have shown that $H(\tau, x) \neq x$ for all $(\tau, x) \in [0, 1] \times \partial A(\beta, b, \alpha, a)$. Note, this also verifies that T does not have any fixed points on $\partial A(\beta, b, \alpha, a)$ (let $\tau = 0$). Thus by the homotopy invariance property of the fixed point index

$$i(T, A(\beta, b, \alpha, a), P) = i(x_0, A(\beta, b, \alpha, a), P),$$

and by the normality property of the fixed point index

$$i(T, A(\beta, b, \alpha, a), P) = i(x_0, A(\beta, b, \alpha, a), P) = 1.$$

Consequently by the solution property of the fixed point index, T has a fixed point $x_* \in A(\beta, b, \alpha, a)$.

Moreover, if for all $x \in A(\beta, b, \alpha, a)$ there exists a $k \in [0, 1)$ such that

$$\|Tx - x_*\| \leq k\|x - x_*\|,$$

then for any fixed point $z^* \in A(\beta, b, \alpha, a)$ we have that

$$\|z^* - x_*\| = \|Tz^* - x_*\| \leq k\|z^* - x_*\|.$$

Therefore $\|z^* - x_*\| = 0$ as $k < 1$, and we have verified that under this condition T has a unique fixed point in $A(\beta, b, \alpha, a)$. Furthermore, if $y_0 \in A(\beta, b, \alpha, a)$ and

$$\|T^n y_0 - x_*\| \leq k\|T^{n-1} y_0 - x_*\|$$

for all positive integers n, then by induction

$$\|T^n y_0 - x_*\| \leq k^n \|y_0 - x_*\|;$$

hence the iterates converge to the fixed point x_*, that is,

$$T^n y_0 \to x_*.$$

□

The following corollary condenses the Leggett-Williams type conditions of Theorem 3 into invariance-like conditions.

Corollary 2 *Suppose P is a cone in a real Banach space E, A is a relatively open subset of P, α is a nonnegative continuous concave functional on P, β is a nonnegative continuous convex functional on P, and $T : P \to P$ is a completely continuous operator. If there exist nonnegative numbers a and b and $x_0 \in A(\beta, b, \alpha, a)$ such that*

$(A0)$ *$A(\beta, b, \alpha, a)$ is bounded;*
$(A1)$ *if $x \in \partial A \cap \overline{P(\beta, b, \alpha, a)}$ and $\tau \in [0, 1]$, then $(1 - \tau)Tx + \tau x_0 \neq x$;*
$(H2)$ *if $x \in \overline{A} \cap \partial P(\beta, b, \alpha, a)$ with $\alpha(x) = a$, then $\alpha(Tx) > a$; and,*
$(H3)$ *if $x \in \overline{A} \cap \partial P(\beta, b, \alpha, a)$ with $\beta(x) = b$, then $\beta(Tx) < b$;*

then T has a fixed point $x_* \in A(\beta, b, \alpha, a)$. Moreover, if for all $x \in A(\beta, b, \alpha, a)$ there exists a $k \in [0, 1)$ such that

$$\|Tx - x_*\| \leq k\|x - x_*\|,$$

then x_* is the unique fixed point of T in $A(\beta, b, \alpha, a)$. Furthermore, if $y_0 \in A(\beta, b, \alpha, a)$ and

$$\|T^n y_0 - x_*\| \leq k\|T^{n-1} y_0 - x_*\|$$

for all positive integers n, then

$$T^n y_0 \to x_*.$$

In our application in the next section we show how one can arrive at invariance conditions to invoke Corollary 2 through a clever choice of the set A, and we show that the iterates will converge to the unique fixed point in our interval of functional type.

3 Application

In this section, using an interval of functional type, we will illustrate the key techniques for verifying the existence and uniqueness of a positive solution for a conjugate boundary value problem in an interval of functional type. Note that the resulting conditions for a fixed point to exist in our functional-type interval will force the boundaries to be mapped in and out respectively in the set $A(\beta, \alpha, b, a)$, however the conditions do not force the boundaries to be mapped in and out respectively in the set $P(\beta, \alpha, b, a)$, which is an important contribution of Theorem 3 and Corollary 2 to the literature. We consider the classical conjugate boundary value problem

$$x''(t) + f(x(t)) = 0, \quad t \in (0, 1), \tag{1}$$

$$x(0) = 0 = x(1), \tag{2}$$

where $f : \mathbb{R} \to [0, \infty)$ is continuous. It is well known that if x is a fixed point of the operator T defined by

$$Tx(t) := \int_0^1 G(t,s) f(x(s)) ds,$$

where

$$G(t,s) = \begin{cases} t(1-s) \text{ if } 0 \le t \le s \le 1 \\ \\ s(1-t) \text{ if } 0 \le s \le t \le 1, \end{cases}$$

then x is a solution of the boundary value problem (1), (2).
Define the cone $P \subset E = C[0,1]$, which is a Banach space with the norm

$$\|x\| = \sup_{t \in [0,1]} |x(t)|,$$

by

$$P = \left\{ y \in E \;\middle|\; \begin{array}{l} y \text{ is concave, symmetric, and} \\ \text{nonnegative valued on [0,1] with} \\ y(0) = 0 = y(1) \end{array} \right\}.$$

One can show that for all $x \in P$, applying the symmetry of x, that

$$(Tx)\left(\frac{1}{4}\right) = \int_0^{\frac{1}{4}} sf(x(s))\, ds + \int_{\frac{1}{4}}^{\frac{1}{2}} \frac{f(x(s))}{4}\, ds$$

and

$$(Tx)\left(\frac{1}{2}\right) = \int_0^{\frac{1}{2}} s\; f(x(s))\, ds.$$

For $x \in P$ define the convex functional β on P by

$$\beta(x) := \max_{t \in \left[0, \frac{1}{2}\right]} x(t) = x\left(\frac{1}{2}\right)$$

and the concave functional α on P by

$$\alpha(x) := \min_{t \in [\frac{1}{4}, \frac{3}{4}]} x(t) = x\left(\frac{1}{4}\right).$$

We are now ready to prove the existence of a unique positive solution to (1), (2) in our functional-type interval if the conditions in the following theorem are satisfied, and show that a sequence of iterates will converge to this unique positive solution in our interval of functional type.

Theorem 4 *If* $0 < \frac{9a}{2} < b$, $0 < M < 16$ *and* $f : [0, \infty) \to [0, \infty)$ *is a continuously differentiable function such that*

(a) $6x < f(x) < 62b$ *for* $x \in [0, a]$,
(b) $\frac{96x}{11} + \frac{36a}{11} < f(x) < 8b$ *for* $x \in \left[a, \frac{35a}{24}\right]$,
(c) $16a < f(x) < 8b$ *for* $x \in \left[\frac{35a}{24}, \frac{3b}{4}\right]$,
(d) $16a < f(x) < 6b$ *for* $x \in \left[\frac{3b}{4}, b\right]$,
(e) $16a < f(x)$ *for* $x \in [b, 2b]$,
(f) $|f'(x)| < M$ *for* $x \in [0, a]$, *and*
(g) $|f'(x)| < \frac{16}{3}$ *for* $x \in [a, 2b]$,

then the conjugate boundary value problem (1), (2) *has a positive solution* $x_* \in A(\beta, b, \alpha, a)$. *Moreover, for every* $y \in A(\beta, b, \alpha, a)$ *we have that*

$$T^n y_0 \to x_*.$$

Proof Let

$$A = \left\{x \in P : x\left(\frac{1}{2}\right) - x\left(\frac{1}{4}\right) < \frac{b}{4}, \frac{35a}{24} < x\left(\frac{1}{2}\right), \right.$$
$$\left. \text{and } (f \circ x)(t) > 16at + 8a \text{ for } t \in \left[\frac{1}{4}, \frac{1}{2}\right]\right\},$$

$c \in \mathbb{R}$ such that $\frac{9a}{2} < c < b$, and

$$x_0(t) = \frac{4ct(1-t)}{3}.$$

Then x_0 satisfies

$$x_0\left(\frac{1}{4}\right) = \frac{c}{4} > a \quad \text{and} \quad x_0\left(\frac{1}{2}\right) = \frac{c}{3} > \frac{9a}{6} > \frac{35a}{24},$$

so that

$$x_0\left(\frac{1}{2}\right) - x_0\left(\frac{1}{4}\right) = \frac{c}{3} - \frac{c}{4} = \frac{c}{12} < \frac{b}{12} < \frac{b}{4}.$$

Let $t \in \left[\frac{1}{4}, \frac{1}{2}\right]$. If $x_0(t) \in \left[a, \frac{35a}{24}\right]$, then by the concavity of x_0 and assumption (b) we have

$$\begin{aligned}(f \circ x_0)(t) &> \frac{96}{11}\, x_0(t) + \frac{36a}{11}\\ &> \frac{96}{11}\left(\frac{11a}{6}t + \frac{13a}{24}\right) + \frac{36a}{11}\\ &= 16at + 8a;\end{aligned}$$

if $x_0(t) > \frac{35a}{24}$, then by assumption (c) we have

$$(f \circ x_0)(t) > 16a \geq 16at + 8a.$$

Thus, in either case we have that $(f \circ x_0)(t) > 16at + 8a$ for $t \in \left[\frac{1}{4}, \frac{1}{2}\right]$. Therefore $x_0 \in A(\beta, b, \alpha, a)$.

For any $x \in \overline{A(\beta, b, \alpha, a)}$ we see that $b \geq x(t) \geq a$ for $t \in \left[\frac{1}{4}, \frac{1}{2}\right]$, thus

$$\begin{aligned}(Tx)\left(\frac{1}{2}\right) - (Tx)\left(\frac{1}{4}\right) &= \int_{\frac{1}{4}}^{\frac{1}{2}} \left(s - \frac{1}{4}\right) f(x(s))\, ds\\ &< \int_{\frac{1}{4}}^{\frac{1}{2}} 8b\left(s - \frac{1}{4}\right) ds = \frac{b}{4}.\end{aligned}$$

Also, for any $x \in \overline{A(\beta, b, \alpha, a)}$ we have that $x\left(\frac{1}{4}\right) \geq a$; thus, by concavity we have $x(t) \geq 4at$ for $t \in \left[0, \frac{1}{4}\right]$, so that $f(x(t)) > 24at$ for $t \in \left[0, \frac{1}{4}\right]$. It follows that

$$\begin{aligned}(Tx)\left(\frac{1}{2}\right) &= \int_0^{\frac{1}{2}} s\, f(x(s))\, ds\\ &= \int_0^{\frac{1}{4}} s\, f(x(s))\, ds + \int_{\frac{1}{4}}^{\frac{1}{2}} s\, f(x(s))\, ds\\ &> \int_0^{\frac{1}{4}} s\, 6(4as)\, ds + \int_{\frac{1}{4}}^{\frac{1}{2}} s\, (16as + 8a)\, ds = \frac{35a}{24}\end{aligned}$$

and

$$(Tx)\left(\frac{1}{4}\right) = \int_0^{\frac{1}{4}} s\, f(x(s))\, ds + \int_{\frac{1}{4}}^{\frac{1}{2}} \frac{f(x(s))}{4}\, ds$$

$$
\begin{aligned}
&> \int_0^{\frac{1}{4}} s\ 6(4as)\ ds + \int_{\frac{1}{4}}^{\frac{1}{2}} \frac{16as+8a}{4}\ ds \\
&\geq \int_0^{\frac{1}{4}} s\ 6(4as)\ ds + \int_{\frac{1}{4}}^{\frac{1}{2}} 4as+2a\ ds = a.
\end{aligned}
$$

By the concavity of Tx and x_0 we have for all $\tau \in [0, 1]$ that

$$(1-\tau)Tx + \tau x_0 \quad \text{is concave.}$$

Also, since $x_0(\frac{1}{4}) > a$, $(Tx)(\frac{1}{4}) > a$, $x_0(\frac{1}{2}) > \frac{35a}{24}$ and $(Tx)(\frac{1}{2}) > \frac{35a}{24}$, we have that

$$a < (1-\tau)Tx\left(\frac{1}{4}\right) + \tau x_0\left(\frac{1}{4}\right) \quad \text{and} \quad \frac{35a}{24} < (1-\tau)Tx\left(\frac{1}{2}\right) + \tau x_0\left(\frac{1}{2}\right).$$

By the concavity of $(1-\tau)Tx + \tau x_0$ we have that

$$((1-\tau)Tx + \tau x_0)(t) \geq \frac{11a}{6}t + \frac{13a}{24} \quad \text{for} \quad t \in \left[\frac{1}{4}, \frac{1}{2}\right].$$

Consequently for $t \in \left[\frac{1}{4}, \frac{1}{2}\right]$, if $((1-\tau)Tx + \tau x_0)(t) \in \left[a, \frac{35a}{24}\right]$, then

$$
\begin{aligned}
(f \circ ((1-\tau)Tx + \tau x_0))(t) &> \frac{96}{11}\ ((1-\tau)Tx + \tau x_0)(t) + \frac{36a}{11} \\
&\geq \frac{96}{11}\ \left(\frac{11a}{6}t + \frac{13a}{24}\right) + \frac{36a}{11} \\
&= 16at + 8a,
\end{aligned}
$$

and if $((1-\tau)Tx + \tau x_0)(t) > \frac{35a}{24}$ then

$$(f \circ ((1-\tau)Tx + \tau x_0))(t) > 16a \geq 16at + 8a.$$

Thus, in either case we have that $(f \circ ((1-\tau)Tx + \tau x_0))(t) > 16at + 8a$. Therefore we have that

$$(1-\tau)Tx + \tau x_0 \neq x$$

for all $x \in \partial A \cap \overline{P(\beta, b, \alpha, a)}$ and $\tau \in [0, 1]$.

Clearly $A(\beta, b, \alpha, a)$ is a nonempty, bounded, open subset of P, and thus condition $(A1)$ of Corollary 2 is satisfied. We have also shown that

$$\alpha(x) > a$$

for all $x \in \overline{A(\beta, b, \alpha, a)}$, so that condition $(H2)$ of Corollary 2 is also satisfied. If $x \in \partial A(\beta, b, \alpha, a)$ with $\beta(x) = b$, then

$$x\left(\frac{1}{2}\right) - x\left(\frac{1}{4}\right) < \frac{b}{4}$$

hence

$$b - \frac{b}{4} = \frac{3b}{4} < x\left(\frac{1}{4}\right),$$

thus by the concavity of x, for $t \in \left[\frac{1}{12}, \frac{1}{4}\right]$, we have that

$$a < \frac{b}{4} \leq x(t).$$

It follows that

$$\begin{aligned}
\beta(Tx) = Tx\left(\frac{1}{2}\right) &= \int_0^{\frac{1}{2}} s\ f(x(s))\ ds \\
&= \int_0^{\frac{1}{12}} s\ f(x(s))\ ds + \int_{\frac{1}{12}}^{\frac{1}{4}} s\ f(x(s))\ ds + \int_{\frac{1}{4}}^{\frac{1}{2}} s\ f(x(s))\ ds \\
&< \int_0^{\frac{1}{12}} 62bs\ ds + \int_{\frac{1}{12}}^{\frac{1}{4}} 8bs\ ds + \int_{\frac{1}{4}}^{\frac{1}{2}} 6bs\ ds \\
&= b.
\end{aligned}$$

As a result, condition $(H3)$ of Corollary 2 is satisfied. Therefore by Corollary 2, T has at least one fixed point $x_* \in A(\beta, b, \alpha, a)$ which is a solution of the boundary value problem (1), (2).
For any $y \in A(\beta, b, \alpha, a)$ we have

$$\begin{aligned}
\|Ty - Tx_*\| &= \max_{t\in[0,1]} \left| \int_0^1 G(t,s)f(y(s))\ ds - \int_0^1 G(t,s)f(x_*(s))\ ds \right| \\
&\leq \max_{t\in[0,1]} \int_0^1 G(t,s)\left|f(y(s)) - f(x_*(s))\right|\ ds \\
&= \int_0^{\frac{1}{2}} s\left|f(y(s)) - f(x_*(s))\right|\ ds \\
&= \int_0^{\frac{1}{4}} s\left|f(y(s)) - f(x_*(s))\right|\ ds
\end{aligned}$$

$$
\begin{aligned}
&+\int_{\frac{1}{4}}^{\frac{1}{2}} s\,|f(y(s)) - f(x_*(s))|\ ds \\
&\leq \int_0^{\frac{1}{4}} s\left(\max_{w\in[0,b]} \left|f'(w)\right|\right)\|y - x_*\|\ ds \\
&\quad+\int_{\frac{1}{4}}^{\frac{1}{2}} s\left(\max_{w\in[a,b]} \left|f'(w)\right|\right)\|y - x_*\|\ ds \\
&< \left(\frac{M}{32}\right)\|y - x_*\| + \left(\frac{1}{2}\right)\|y - x_*\| \\
&= k\|y - x_*\|,
\end{aligned}
$$

where

$$k = \frac{M}{32} + \frac{1}{2} < 1$$

since $M < 16$. Therefore x_* is the unique solution for the boundary value problem (1), (2) in $A(\beta, b, \alpha, a)$.

Claim For all $n \in \mathbb{N}$, $\|T^n y\| \leq 2b$, $(T^n y)(\frac{1}{4}) > a$, and $\|T^n y - x_*\| \leq k\|T^{n-1}$ $y - x_*\|$.

Clearly this is true for $n = 1$ since we have already shown that

$$\|Ty - x_*\| = \|Ty - Tx_*\| < k\|y - x_*\| \leq b,$$

which also verifies that

$$\|Ty\| < 2b.$$

We also have that

$$
\begin{aligned}
(Ty)\left(\frac{1}{4}\right) &= \int_0^{\frac{1}{4}} s\ f(y(s))\ ds + \int_{\frac{1}{4}}^{\frac{1}{2}} \frac{f(y(s))}{4}\ ds \\
&> \int_{\frac{1}{4}}^{\frac{1}{2}} \frac{f(y(s))}{4}\ ds \\
&\geq \int_{\frac{1}{4}}^{\frac{1}{2}} \frac{16a}{4}\ ds = a.
\end{aligned}
$$

Let $m \geq 1$ and suppose that $\|T^j y\| \leq 2b$, $(T^j y)(\frac{1}{4}) > a$, and $\|T^j y - x_*\| \leq k\|T^{j-1} y - x_*\|$ for all $j \leq m$. Thus,

$$\begin{aligned}
(T^{m+1}y)\left(\frac{1}{4}\right) &= \int_0^{\frac{1}{4}} s\ f((T^m y)(s))\, ds + \int_{\frac{1}{4}}^{\frac{1}{2}} \frac{f((T^m y)(s))}{4}\, ds \\
&> \int_{\frac{1}{4}}^{\frac{1}{2}} \frac{f((T^m y)(s))}{4}\, ds \\
&\geq \int_{\frac{1}{4}}^{\frac{1}{2}} \frac{16a}{4}\, ds = a,
\end{aligned}$$

and

$$\begin{aligned}
\|T^{m+1}y - x_*\| &= \|T^{m+1}y - Tx_*\| \\
&= \max_{t\in[0,1]} \left| \int_0^1 G(t,s) f(T^m y(s))\, ds - \int_0^1 G(t,s) f(x_*(s))\, ds \right| \\
&\leq \max_{t\in[0,1]} \int_0^1 G(t,s) \left| f((T^m y)(s)) - f(x_*(s)) \right|\, ds \\
&= \int_0^{\frac{1}{2}} s \left| f((T^m y)(s)) - f(x_*(s)) \right|\, ds \\
&= \int_0^{\frac{1}{4}} s \left| ((T^m y)(s)) - f(x_*(s)) \right|\, ds \\
&\quad + \int_{\frac{1}{4}}^{\frac{1}{2}} s \left| f((T^m y)(s)) - f(x_*(s)) \right|\, ds \\
&\leq \int_0^{\frac{1}{4}} s \left(\max_{w\in[0,2b]} \left| f'(w) \right| \right) \|T^m y - x_*\|\, ds \\
&\quad + \int_{\frac{1}{4}}^{\frac{1}{2}} s \left(\max_{w\in[a,2b]} \left| f'(w) \right| \right) \|T^m y - x_*\|\, ds \\
&< \left(\frac{M}{32}\right) \|T^m y - x_*\| + \left(\frac{1}{2}\right) \|T^m y - x_*\| \\
&= k\|T^m y - x_*\|.
\end{aligned}$$

Consequently we have

$$\|T^{m+1}y - x_*\| \leq k^{m+1}\|y - x_*\|,$$

since $\|T^j y - x_*\| \leq k\|T^{j-1}y - x_*\|$ for all $j \leq m$. Note, this also verifies that $\|T^{m+1}y\| < 2b$, and the claim is proven by the principle of mathematical induction. Therefore, for every $y \in A(\beta, b, \alpha, a)$ we have that

$$T^n y \to x_*.$$

This completes the proof. □

References

1. D.R. Anderson, R.I. Avery, Fixed point theorem of cone expansion and compression of functional type. J. Differ. Equ. Appl. **8**, 1073–1083 (2002)
2. D.R. Anderson, R.I. Avery, J. Henderson, Functional expansion - compression fixed point theorem of Leggett-Williams type. Electron. J. Differ. Equ. **2010**(63), 1–9 (2010)
3. D.R. Anderson, R.I. Avery, J. Henderson, An extension of the compression-expansion fixed point theorem of functional type. Electron. J. Differ. Equ. **2016**(253), 1–9 (2016)
4. R.I. Avery, A generalization of the Leggett-Williams fixed point theorem. MSR Hot-Line **3**(7), 9–14 (1999)
5. K. Deimling, *Nonlinear Functional Analysis* (Springer, New York, 1985)
6. D. Guo, A new fixed point theorem. Acta Math. Sin. **24**, 444–450 (1981)
7. D. Guo, Some fixed point theorems on cone maps. Kexeu Tongbao **29**, 575–578 (1984)
8. D. Guo, V. Lakshmikantham, *Nonlinear Problems in Abstract Cones* (Academic, San Diego, 1988)
9. M.A. Krasnosel'skii, *Positive Solutions of Operator Equations* (P. Noordhoff, Groningen, 1964)
10. R.W. Leggett, L.R. Williams, Multiple positive fixed points of nonlinear operators on ordered Banach spaces. Indiana Univ. Math. J. **28**, 673–688 (1979)
11. W.V. Petryshyn, Multiple positive solutions of multivalued condensing mappings with some applications. J. Math. Anal. Appl. **124**, 237–253 (1987)
12. W.V. Petryshyn, Existence of fixed points of positive k-set-contractive maps as consequences of suitable boundary conditions. J. Lond. Math. Soc. **38**, 503–512 (1988)
13. J. Sun, G. Zhang, A generalization of the cone expansion and compression fixed point theorem and applications. Nonlinear Anal. **67**, 579–586 (2007)
14. E. Zeidler, *Nonlinear Functional Analysis and Its Applications I, Fixed Point Theorems* (Springer, New York, 1986)

On Lagrangian Duality in Infinite Dimension and Its Applications

Antonio Causa, Giandomenico Mastroeni, and Fabio Raciti

1 Introduction

The last decade has witnessed a renewed interest in the topic of Lagrangian duality in infinite dimensional spaces, mainly motivated by the need to deal with time-dependent or stochastic equilibrium problems in a Lebesgue space setting, or with unilateral problems described by elliptic partial differential equations. Indeed, in these applications, the classical theory is not applicable, because the ordering cones used have empty topological interior and the usual constraint qualifications (see e.g. [15, 20]) are not useful. In particular, in the Lebesgue spaces L^p, with $p > 1$, it is well known that the cone of the almost everywhere nonnegative functions, often used to describe inequality constraints in equilibrium problems (see e.g. [12, 17, 18]), has empty topological interior, and the same happens in some Sobolev spaces widely used in unilateral problems [8]. In order to overcome this problem, the concept of quasi relative interior has been proposed and used by some scholars (see e.g. [1, 2, 22]). A conical regularization method to cope with empty interior cones has been proposed, from both the theoretical and numerical aspects in [21], while a theory applicable to nonconvex problems can be found in [13]. It seems, however, that the paper which mostly influenced the research on this topic has been [7], because after its publication several scholars devoted their efforts to apply or develop the results therein (see e.g. [6, 9, 16, 23, 24]; C. Zălinescu, Private communications, September 4th, 5th, 7th, October 10th, 2007). Furthermore, the new theory has been further improved in [3–5].

A. Causa · F. Raciti (✉)
Dipartimento di Matematica e Informatica dell'Università di Catania, Catania, Italy
e-mail: causa@dmi.unict.it; fraciti@dmi.unict.it

G. Mastroeni
Dipartimento di Informatica dell'Università di Pisa, Pisa, Italy
e-mail: giandomenico.mastroeni@unipi.it

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_3

We now specify the aim and character of the present contribution which is meant for a broad audience, not necessarily skilled in duality theory. As a consequence we make no attempt to provide a long list of references and of complementary results. On the contrary, we aim for self-consistency and would like to convey to the reader some important ideas and tools of duality theory. In the following section we present a somewhat classical duality theory, along the same lines as in [19], which utilizes a methodology similar to the conical extension widely used in [14]. In Sect. 3,we present the main result of the new duality theory in the revised version of [4]. In Sect. 4, we apply this new approach to investigate a theoretical property of Nash-Rosen equilibria in infinite dimension [11]. At last, in the concluding section we summarize our analysis and offer some research perspectives.

2 Lagrangian Duality in a Classical Framework

We establish some notation and useful definitions. For all the notions of general convex analysis we refer to the book [26]. All the linear spaces we consider are real. We denote by $\mathbb{R}^+$ the open interval $(0, +\infty)$ and by $\mathbb{R}^+_0$ the interval $[0, +\infty)$, and analogously, $\mathbb{R}^- = (-\infty, 0)$, $\mathbb{R}^-_0 = (-\infty, 0]$. The topological dual space of a topological linear space Y will be denoted by Y^*, while $\langle \cdot, \cdot \rangle$ denotes the usual duality pairing between Y and Y^*, i.e., if $y \in Y$ and $u \in Y^*$, $\langle u, y \rangle$ is the value of u at y. Moreover if C is a cone in Y, the dual cone of C is defined by

$$C^* := \{l \in Y^* | \langle l, y \rangle \geq 0, \ \forall y \in C\}.$$

If K is a convex subset of a linear space and Y is a linear space partially ordered by a convex cone C, a mapping $g : K \to Y$ is called C-convex iff $\forall x, y \in K$, $\forall t \in [0, 1]$ it holds

$$t\, g(x) + (1 - t)g(y) - g[t\, x + (1 - t)y] \in C.$$

The notion of convexity has been generalized in several ways. For our purposes the notion of a *convex-like* function will be useful.

Definition 1 A map $g : K \to Y$ is called *convex-like w.r.t. the convex cone* $C \subset Y$ if the set $g(K) + C$ is convex.

Let us consider the following setting that will be assumed throughout the paper.

Let $\hat{S}$ be a convex set of a real linear topological space.

Let $(Y, ||\cdot||_Y)$ and $(Z, ||\cdot||_Z)$ be a partially ordered normed space with ordering cone C and a real normed space, respectively.

Let $f : \hat{S} \to \mathbb{R}$ be a given objective functional.

Let $g : \hat{S} \to Y$ and $h : \hat{S} \to Z$ be given constraints mappings with h affine-linear.

Let the mapping $(f, g, h) : \hat{S} \to \mathbb{R} \times Y \times Z$ be convex-like w.r.t. the product cone $\mathbb{R}_0^+ \times C \times \{0_Z\}$.

Furthermore, assume that the set $S = \{x \in \hat{S} | g(x) \in -C, \ h(x) = 0_Z\}$ is nonempty and C is a closed convex cone whose interior will be assumed nonempty whenever, in Sect. 2, we will user Slater's assumption. We consider the following optimization problem which, in the sequel, will be called Primal Problem.

$$\min_{x \in S} f(x). \tag{1}$$

The following lemma states that the Primal Problem is equivalent to the optimization problem

$$\min_{x \in \hat{S}} \ \sup_{(u,v) \in C^* \times Z^*} \ L(x, u, v), \tag{2}$$

where the functional $L : \hat{S} \times C^* \times Z^* \to \mathbb{R}$ defined by

$$L(x, u, v) := f(x) + \langle u, g(x) \rangle + \langle v, h(x) \rangle$$

is called Lagrange functional associated with problem (1).

Lemma 1 *$\overline{x} \in S$ is an optimal solution of problem (1) if and only if $\overline{x}$ is an optimal solution of problem (2). In this case the extremal values of both problems are equal.*

Proof Assume that $\overline{x} \in S$ is a minimum point of f on S. Let us notice that for $x \in \hat{S}$ such that $g(x) \in -C$ one has $\langle u, g(x) \rangle \leq 0, \forall u \in C^*$, hence $\sup_{u \in C^*} \langle u, g(x) \rangle = 0$.

Thus, for all $x \in \hat{S}$ with $g(x) \in -C, h(x) = 0_Z$ we get $\langle u, g(x) \rangle + \langle v, h(x) \rangle \leq 0$ for all $(u, v) \in C^* \times Z^*$, hence

$$\sup_{(u,v) \in C^* \times Z^*} \langle u, g(x) \rangle + \langle v, h(x) \rangle = 0.$$

Consider now an arbitrary x of $\hat{S}$ such that $g(x) \notin -C$. Since C is convex and closed there is a $\overline{u} \in C^*$ such that $\langle \overline{u}, g(x) \rangle > 0$ and we can prove that

$$\sup_{(u,v) \in C^* \times Z^*} \langle \overline{u}, g(x) \rangle + \langle v, h(x) \rangle = +\infty.$$

Indeed, if $g(x) \notin -C$ the separation Theorem 11 (see the Appendix) ensures that $\exists \overline{u} \in Y^* \setminus \{0_{Y^*}\}$ such that $\langle \overline{u}, -g(x) \rangle < \inf_{y \in C} \langle \overline{u}, y \rangle$, which implies $\langle \overline{u}, g(x) \rangle > 0$. Moreover, it is not difficult to prove that $\overline{u}$ actually belongs to C^*. For each $\lambda > 0$ we also get $\lambda \overline{u} \in C^*$, whence:

$$\sup_{u \in C^*} \langle u, g(x) \rangle = +\infty.$$

If $h(x) \neq 0_Z$, we immediately obtain that

$$\sup_{v \in Z^*} \langle v, h(x) \rangle = +\infty.$$

For all $x \in \hat{S}$ we thus get

$$\begin{aligned} \sup_{(u,v) \in C^* \times Z^*} \{f(\overline{x}) + \langle u, g(\overline{x}) \rangle + \langle v, h(\overline{x}) \rangle\} &= f(\overline{x}) + \sup_{(u,v) \in C^* \times Z^*} \{\langle u, g(\overline{x}) \rangle + \langle v, h(\overline{x}) \rangle\} \\ &= f(\overline{x}) \\ &\leq f(\overline{x}) + \sup_{(u,v) \in C^* \times Z^*} \langle u, g(x) \rangle + \langle v, h(x) \rangle \\ &\leq \sup_{(u,v) \in C^* \times Z^*} \{f(x) + \langle u, g(x) \rangle + \langle v, h(x) \rangle\}, \end{aligned}$$

which means that $\overline{x} \in S$ is also an optimal solution of the optimization problem (2).

Conversely, let $\overline{x} \in \hat{S}$ be a minimal point of the functional $\varphi : \hat{S} \longrightarrow \mathbb{R} \cup \{+\infty\}$ defined by

$$\varphi(x) = \sup_{(u,v) \in C^* \times Z^*} f(x) + \langle u, g(x) \rangle + \langle v, h(x) \rangle.$$

It can be easily seen that

$$\text{if } g(\overline{x}) \notin -C, \;\; \text{or } h(\overline{x}) \neq 0 \quad \text{then} \quad \sup_{(u,v) \in C^* \times Z^*} \langle u, g(x) \rangle + \langle v, h(x) \rangle = +\infty,$$

which is an obstruction to solvability of problem (2). It follows that $\overline{x} \in S$ and, with the same reasoning as in the first part of the proof,

$$\sup_{(u,v) \in C^* \times Z^*} \{\langle u, g(\overline{x}) \rangle + \langle v, h(\overline{x}) \rangle\} = 0.$$

Thus, for all $x \in S$ we get

$$\begin{aligned} f(\overline{x}) &= f(\overline{x}) + \sup_{(u,v) \in C^* \times Z^*} \{\langle u, g(\overline{x}) \rangle + \langle v, h(\overline{x}) \rangle\} \\ &\leq \sup_{(u,v) \in C^* \times Z^*} \{f(x) + \langle u, g(x) \rangle + \langle v, h(x) \rangle\} \\ &= f(x) + \sup_{(u,v) \in C^* \times Z^*} \{\langle u, g(x) \rangle + \langle v, h(x) \rangle\} = f(x), \end{aligned}$$

which means that $\overline{x}$ is a minimal point of f on S. □

We can associate with the Primal Problem (1) the following optimization problem

$$\max_{(u,v) \in C^* \times Z^*} \inf_{x \in \hat{S}} L(x, u, v), \tag{3}$$

which is called the Dual Problem associated with the Primal Problem (1) or equivalently (2). There are some relationships between the Primal Problem and the Dual Problem. The first of these is the weak duality theorem.

Theorem 1 *The maximal value of the Dual Problem is bounded from above by the minimal value of the Primal Problem i.e.*

$$\max_{(u,v)\in C^*\times Z^*}\ \inf_{x\in\hat{S}} L(x,u,v) \le \min_{x\in\hat{S}}\ \sup_{(u,v)\in C^*\times Z^*} L(x,u,v). \tag{4}$$

Proof It is a consequence of the standard inequality

$$\sup_{b\in B}\inf_{a\in A} F(a,b) \le \inf_{a\in A}\sup_{b\in B} F(a,b),$$

where $F : A\times B \to \mathbb{R}$. □

With the aid of additional sufficient conditions one can show that the Primal Problem and the Dual Problem are equivalent, i.e., inequality is replaced by equality in (4):

$$\max_{(u,v)\in C^*\times Z^*}\ \inf_{x\in\hat{S}} L(x,u,v) = \min_{x\in\hat{S}}\ \sup_{(u,v)\in C^*\times Z^*} L(x,u,v).$$

In such a case we say that strong duality holds for problem (1). Theorems that guarantee this equivalence are called strong duality theorems.

Theorem 2 *Assume that $h(\hat{S})$ contains a neighborhood of 0_Z and that there exists $\hat{x}\in\hat{S}$ such that $g(\hat{x})\in -\operatorname{int} C$, $h(\hat{x}) = 0_Z$. The last hypothesis is called Slater's condition.*

If the Primal Problem (1) admits an optimal solution then the Dual Problem admits an optimal solution and

$$\max_{(u,v)\in C^*\times Z^*}\ \inf_{x\in\hat{S}} L(x,u,v) = \min_{x\in\hat{S}}\ \sup_{(u,v)\in C^*\times Z^*} L(x,u,v).$$

Proof Let us consider the following subset of the space $\mathbb{R}\times Y\times Z$, endowed with the product topology,

$$\begin{aligned} M &= \{(f(x)+\alpha, g(x)+y, h(x)) \in \mathbb{R}\times Y\times Z | x\in\hat{S}, \alpha\ge 0, y\in C\}\\ &= (f,g,h)(\hat{S}) + \mathbb{R}_0^+ \times C\times\{0_Z\}. \end{aligned}$$

By our initial assumptions, the composite mapping $(f,g,h) : \hat{S}\to\mathbb{R}\times Y\times Z$ is convex-like, hence the set M is convex. Since $\operatorname{int} C \ne \varnothing$ and $h(\hat{S})$ contains a neighborhood of 0_Z, then $\operatorname{int} M\ne\varnothing$. By hypothesis there exists $\overline{x}\in S$ such that

$$f(\overline{x}) \le f(x) \quad \text{for all } x\in S.$$

Obviously $(f(\overline{x}), 0_Y, 0_Z)\in M$. We now prove that $(f(\overline{x}), 0_Y, 0_Z)\notin \operatorname{int} M$.

Indeed, let us consider $(a, b, c) \in M$ arbitrarily and notice that if $x \in S$ then $a \geq f(\overline{x})$, otherwise at least one of the two conditions: $b \notin -C$, $c \neq 0_Z$ is fulfilled. Let us now consider a neighborhood of $(f(\overline{x}), 0_Y, 0_Z))$:

$$]f(\overline{x}) - \varepsilon, f(\overline{x}) + \varepsilon[\times I(0_Y) \times I'(0_Z), \quad \text{for some } \varepsilon > 0$$

and observe that in such an arbitrary neighborhood there are points which do not belong to M. To this aim, choose $a \in]f(\overline{x}) - \varepsilon, f(\overline{x})[$, $c = 0$ and $b \in I(0_Y)$ such that $b \in -C$. As a consequence, $(f(\overline{x}), 0_Y, 0_Z)) \notin \operatorname{int} M$.

By a separation theorem there are $(\mu, u, v) \in \mathbb{R} \times Y^* \times Z^*$, $\gamma \in \mathbb{R}$ with $(\mu, u, v) \neq (0, 0_{Y^*}, 0_{Z^*})$ such that

$$\mu\beta + \langle u, z\rangle + \langle v, z'\rangle > \gamma \geq \mu f(\overline{x}) \quad \text{for all } (\beta, z, z') \in \operatorname{int} M. \tag{5}$$

Since for a convex set K with nonempty interior the equality $\operatorname{cl}(K) = \operatorname{cl}(\operatorname{int} K)$ holds true, we can conclude from inequality (5) that

$$\mu(f(x)+\alpha)+\langle u, g(x)+y\rangle+\langle v, h(x)\rangle \geq \gamma \geq \mu f(\overline{x}) \quad \text{for all } x \in \hat{S},\ \alpha \geq 0,\ y \in C. \tag{6}$$

For $x = \overline{x}$, $\alpha = 0$ we obtain from (6) that

$$\langle u, y\rangle \geq -\langle u, g(\overline{x})\rangle,\ \forall y \in C \tag{7}$$

and because C is a cone one can prove that $u \in C^*$. Since $g(\overline{x}) \in -C$, the preceding formula, for $y = 0$ implies

$$\langle u, g(\overline{x})\rangle = 0.$$

For $x = \overline{x}$, $y = 0_Y$ from (6) we obtain $\mu\alpha \geq 0$, $\forall \alpha \geq 0$, hence $\mu \geq 0$. In order to prove that $\mu > 0$, we assume $\mu = 0$ and will then get the false equality $(\mu, u, v) = (0, 0_{Y^*}, 0_{Z^*})$. Indeed, if $\mu = 0$, (6) gives

$$\langle u, g(x) + y\rangle + \langle v, h(x)\rangle \geq 0, \forall x \in \hat{S}, \forall y \in C, \tag{8}$$

Because of the Slater assumption, $\exists \hat{x} \in \hat{S} : g(\hat{x}) \in -\operatorname{int}(C), h(\hat{x}) = 0_Z$.

From (8) we obtain:

$$\langle u, g(\hat{x}) + y\rangle \geq 0\ \forall y \in C$$

which, for $y = 0$ reads as

$$\langle u, g(\hat{x})\rangle \geq 0.$$

Since $g(\hat{x}) \in -\operatorname{int}(C)$, $u \neq 0$ implies $\langle u, g(\hat{x})\rangle < 0$, which yields to $\mu = 0$.
Now, if $v \neq 0$ holds true, we get

$$\langle v, h(x)\rangle \geq 0, \forall x \in \hat{S},$$

but since $h(\hat{S})$ contains a neighborhood of 0_Z and h being linear-affine, it follows that $\langle v, h(\hat{S})\rangle$ assumes both negative and positive values, contradicting the above inequality, hence also $v = 0_{Z^*}$.

From inequality (6) with $\alpha = 0$ and $y = 0_Y$ we get

$$\mu f(x) + \langle u, g(x)\rangle + \langle v, h(x)\rangle \geq \mu f(\overline{x}) \quad \text{for all } x \in \hat{S}$$

which implies

$$f(x) + \frac{1}{\mu}\langle u, g(x)\rangle + \frac{1}{\mu}\langle v, h(x)\rangle \geq f(\overline{x}) \quad \text{for all } x \in \hat{S}.$$

Let us denote $\overline{u} = \dfrac{1}{\mu}u \in C^*$, $\overline{v} = \dfrac{1}{\mu}v \in Z^*$. We obtain that $\langle \overline{u}, g(\overline{x})\rangle = 0$ and, since $h(\overline{x}) = 0$, that $\langle \overline{v}, h(\overline{x})\rangle = 0$. It follows

$$\inf_{x \in \hat{S}} f(x) + \langle \overline{u}, g(x)\rangle + \langle \overline{v}, h(x)\rangle \geq f(\overline{x}) + \langle \overline{u}, g(\overline{x})\rangle + \langle \overline{v}, h(\overline{x})\rangle.$$

Hence we get

$$f(\overline{x}) = \inf_{x \in \hat{S}} f(x) + \langle \overline{u}, g(x)\rangle + \langle \overline{v}, h(x)\rangle,$$

and due to the weak duality theorem, $(\overline{u}, \overline{v})$ is an optimal solution of the dual problem. □

We can describe the relationships between the Primal Problem and the Dual Problem by the notion of saddle point for the Lagrange functional L.

Definition 2 A point $(\overline{x}, \overline{u}, \overline{v}) \in \hat{S} \times C^* \times Z^*$ is called a saddle point of the Lagrange functional L on $\hat{S} \times C^* \times Z^*$ if

$$L(\overline{x}, u, v) \leq L(\overline{x}, \overline{u}, \overline{v}) \leq L(x, \overline{u}, \overline{v}) \qquad \forall x \in \hat{S}, u \in C^*, v \in Z^*.$$

A saddle point of the Lagrange functional can be characterized by a minimax theorem as follows.

Theorem 3 *A point $(\overline{x}, \overline{u}, \overline{v}) \in \hat{S} \times C^* \times Z^*$ is a saddle point of the Lagrange functional L iff*

$$\min_{x \in \hat{S}} \sup_{(u,v) \in C^* \times Z^*} L(x, u, v) = \max_{(u,v) \in C^* \times Z^*} \inf_{x \in \hat{S}} L(x, u, v) \tag{9}$$

and $\overline{x}$ and $(\overline{u}, \overline{v})$ are the optimal solutions of the problems which appear in the left side and in the right side of (9) respectively.

Proof Let us assume that Eq. (9) is satisfied. For $\overline{x} \in \hat{S}$ and $(\overline{u}, \overline{v}) \in C^* \times Z^*$ we get

$$\sup_{(u,v)\in C^*\times Z^*} L(\overline{x}, u, v) = \inf_{x\in\hat{S}} L(x, \overline{u}, \overline{v})$$

so it follows

$$L(\overline{x}, \overline{u}, \overline{v}) \leq \sup_{(u,v)\in C^*\times Z^*} L(\overline{x}, u, v) = \inf_{x\in\hat{S}} L(x, \overline{u}, \overline{v}) \leq L(\overline{x}, \overline{u}, \overline{v})$$

which obviously gives

$$L(\overline{x}, \overline{u}, \overline{v}) = \sup_{(u,v)\in C^*\times Z^*} L(\overline{x}, u, v) = \inf_{x\in\hat{S}} L(x, \overline{u}, \overline{v}).$$

Hence $(\overline{x}, \overline{u}, \overline{v})$ is a saddle point of the Lagrange functional L.

Now let us assume that $(\overline{x}, \overline{u}, \overline{v}) \in \hat{S} \times C^* \times Z^*$ is a saddle point of L. By definition we get

$$\max_{(u,v)\in C^*\times Z^*} L(\overline{x}, u, v) = L(\overline{x}, \overline{u}, \overline{v}) = \min_{x\in\hat{S}} L(x, \overline{u}, \overline{v}) \tag{10}$$

Given $\hat{x}$ and $(\hat{u}, \hat{v}) \in C^* \times Z^*$ we get $\inf_{x\in\hat{S}} L(x, \hat{u}, \hat{v}) \leq L(\hat{x}, \hat{u}, \hat{v})$ and so

$$\sup_{(u,v)\in C^*\times Z^*} \inf_{x\in\hat{S}} L(x, u, v) \leq \sup_{(u,v)\in C^*\times Z^*} L(\hat{x}, u, v)$$

hence

$$\sup_{(u,v)\in C^*\times Z^*} \inf_{x\in\hat{S}} L(x, u, v) \leq \inf_{x\in\hat{S}} \sup_{(u,v)\in C^*\times Z^*} L(\hat{x}, u, v).$$

From this inequality and Eq. (10) we get

$$\begin{aligned} L(\overline{x}, \overline{u}, \overline{v}) &= \inf_{x\in\hat{S}} L(x, \overline{u}, \overline{v}) \leq \sup_{(u,v)\in C^*\times Z^*} \inf_{x\in\hat{S}} L(x, u, v) \\ &\leq \inf_{x\in\hat{S}} \sup_{(u,v)\in C^*\times Z^*} L(x, u, v) \leq \sup_{(u,v)\in C^*\times Z^*} L(\overline{x}, u, v) \\ &= L(\overline{x}, \overline{u}, \overline{v}) \end{aligned}$$

which easily gives the claim

$$L(\overline{x}, \overline{u}, \overline{v}) = \max_{(u,v)\in C^*\times Z^*} \inf_{x\in\hat{S}} L(x, u, v) = \min_{x\in\hat{S}} \sup_{(u,v)\in C^*\times Z^*} L(x, u, v).$$

□

The following theorem provides a relationship between a saddle point of the Lagrange functional and the optimal solutions of the Primal and the Dual problems.

Theorem 4 *A point* $(\overline{x}, \overline{u}, \overline{v}) \in \hat{S} \times C^* \times Z^*$ *is a saddle point of the Lagrange functional* L *iff* $\overline{x}$ *is an optimal solution of the Primal Problem (2),* $(\overline{u}, \overline{v})$ *is an optimal solution of the Dual Problem (3) and strong duality holds for* (1).

Proof Let us first assume that the point $(\overline{x}, \overline{u}, \overline{v}) \in \hat{S} \times C^* \times Z^*$ is a saddle point of the Lagrange functional $L : \hat{S} \times C^* \times Z^* \to \mathbb{R}$. By Theorem (3) we get

$$L(\overline{x}, \overline{u}, \overline{v}) = \min_{x\in\hat{S}} \sup_{(u,v)\in C^*\times Z^*} L(x, u, v) = \max_{(u,v)\in C^*\times Z^*} \inf_{x\in\hat{S}} L(x, u, v).$$

It follows that $\overline{x}$ is an optimal solution of problem (2) and, by Lemma 1, $\overline{x}$ is also an optimal solution of the primal problem (1). By the preceding equation, $(\overline{u}, \overline{v})$ is an optimal solution of the dual problem (3) and the extremal values of the two problems are equal.

Conversely, let us assume that $\overline{x}$ is an optimal solution of the primal problem (1) and that $(\overline{u}, \overline{v})$ is an optimal solution of the dual problem (3) and that the extremal values of these problems are equal.

Hence we get

$$\lambda := \inf_{x\in\hat{S}} L(x, \overline{u}, \overline{v}) = \max_{(u,v)\in C^*\times Z^*} \inf_{x\in\hat{S}} L(x, u, v)$$

and, by Lemma 1,

$$f(\overline{x}) = \sup_{(u,v)\in C^*\times Z^*} L(\overline{x}, u, v) = \min_{x\in\hat{S}} \sup_{(u,v)\in C^*\times Z^*} L(x, u, v).$$

It follows that $\lambda = f(\overline{x})$ and

$$\begin{aligned} \langle \overline{u}, g(\overline{x})\rangle = \langle \overline{u}, g(\overline{x})\rangle + \langle \overline{v}, h(\overline{x})\rangle &\geq -f(\overline{x}) + \inf_{x\in\hat{S}} f(x) + \langle \overline{u}, g(x)\rangle + \langle \overline{v}, h(x)\rangle \\ &= -f(\overline{x}) + \lambda = 0. \end{aligned}$$

Since $g(\overline{x}) \in -C$ and $\overline{u} \in C^*$ we get $\langle \overline{u}, g(\overline{x})\rangle \leq 0$, which implies $\langle \overline{u}, g(\overline{x})\rangle = 0$ and finally $f(\overline{x}) = L(\overline{x}, \overline{u}, \overline{v})$.

Thus we get

$$L(\overline{x}, \overline{u}, \overline{v}) = \min_{x\in\hat{S}} \sup_{(u,v)\in C^*\times Z^*} L(x, u, v) = \max_{(u,v)\in C^*\times Z^*} \inf_{x\in\hat{S}} L(x, u, v)$$

which, by the preceding Theorem 3, means that $(\overline{x}, \overline{u}, \overline{v})$ is a saddle point of the Lagrange functional. □

As a consequence of the Strong Duality Theorem 2 we get the following corollary which gives a sufficient condition for the existence of a saddle point of the Lagrange functional.

Corollary 1 *Assume that $h(\hat{S})$ contains a neighborhood of 0_Z and that there exists $\hat{x} \in \hat{S}$ such that $g(\hat{x}) \in -\operatorname{int} C$, $h(\hat{x}) = 0_Z$.*

If $\overline{x} \in S$ is an optimal solution of the Primal Problem (1), then there exists $(\overline{u}, \overline{v}) \in C^* \times Z^*$ such that $(\overline{x}, \overline{u}, \overline{v})$ is a saddle point of the Lagrange functional L.

3 A Characterization of Strong Duality in Infinite Dimension

As already explained in the introduction, in several concrete infinite dimensional problems the ordering cones C have empty interior. To overcome this issue one can use the notion of quasi relative interior for a convex set which allows for the use of a new kind of separation theorems. Based on this kind of analysis, suitable characterizations of strong duality have been obtained in the image space associated with the Primal Problem (1) (see, e.g. [7]). We first provide some definitions and propositions which will be useful in the sequel. The related proofs can be found for instance in [19] or [26].

Definition 3 Let $\hat{S}$ be a nonempty subset of a linear space. The set

$$\operatorname{Cone}(\hat{S}) := \{ts \,|\, t \geq 0, \;\; s \in \hat{S}\}$$

is called the cone generated by $\hat{S}$.

Definition 4 The set defined as

$$T_{\hat{S}}(\overline{x}) = \{y \in X \,|\; y = \lim_{n\to\infty} \lambda_n(x_n - \overline{x}),\; \lambda_n > 0,\; x_n \in \hat{S},\; \lim_{n\to\infty} x_n = \overline{x},\; \forall n \in \mathbb{N}\}$$

is called the contingent (or tangent) cone of the set $\hat{S}$ at the point $\overline{x}$.

Proposition 1 *If $\hat{S}$ is starshaped with respect to $\overline{x} \in \hat{S}$, then*

$$\operatorname{Cone}(\hat{S} - \{\overline{x}\}) \subset T_{\hat{S}}(\overline{x}).$$

Proposition 2 *Let $\hat{S} \neq \varnothing$. For each $\overline{x} \in \hat{S}$ the following inclusion holds true:*

$$T_{\hat{S}}(\overline{x}) \subset \operatorname{cl}\operatorname{Cone}(\hat{S} - \{\overline{x}\}).$$

Proposition 3 *For each $\overline{x} \in \hat{S}$, $T_{\hat{S}}(\overline{x})$ is closed. Hence, if $\hat{S}$ is starshaped with respect to $\overline{x} \in S$, then*

$$T_{\hat{S}}(\overline{x}) = \operatorname{cl}\operatorname{Cone}(\hat{S} - \{\overline{x}\}).$$

Furthermore, if $\hat{S}$ is convex, then also $T_{\hat{S}}(\overline{x})$ is convex, for every $\overline{x} \in \hat{S}$.

Let us now consider the Primal and Dual problems, (1) and (3). We recall now the new assumption introduced in [7].

Definition 5 We say that *Assumption S* is fulfilled at the point $x_0 \in S$ iff

$$T_{\widetilde{M}}(f(x_0), 0_Y, 0_Z) \cap (\mathbb{R}^- \times \{0_Y\} \times \{0_Z\}) = \varnothing$$

where

$$\widetilde{M} = \{(f(x), g(x), h(x)) | x \in \hat{S} \setminus S\} + (\mathbb{R}^+_0 \times C \times \{0_Z\}).$$

In the sequel we follow the development given in [4]. First let us introduce the set:

$$\mathscr{E} = \{(f(x_0), 0_Y, 0_Z) - (f, g, h)(\hat{S}) - (\mathbb{R}^+_0 \times C \times \{0_Z\})\}.$$

We observe that (f, g, h) convex-like implies that $\mathscr{E}$ is convex.

Moreover, the optimality of a feasible point x_0 can be expressed by means of the set $\mathscr{E}$ as shown in the following result.

Proposition 4 *$x_0 \in S$ is an optimal solution for the Primal Problem iff*

$$\mathscr{E} \cap (\mathbb{R}^+ \times C \times \{0_Z\}) = \varnothing. \qquad (11)$$

Proof Note that $x_0 \in S$ is an optimal solution for (1) iff the following system is impossible

$$\begin{cases} f(x_0) - f(x) > 0 \\ g(x) \in -C \\ h(x) = 0_Z \\ x \in \hat{S} \end{cases}$$

which can be equivalently written as

$$((f(x_0), 0_Y, 0_Z) - (f, g, h)(\hat{S})) \cap (\mathbb{R}^+ \times C \times \{0_Z\}) = \varnothing.$$

or,

$$(0, 0_Y, 0_Z) \notin [(f(x_0), 0_Y, 0_Z) - (f, g, h)(\hat{S}) - (\mathbb{R}^+ \times C \times \{0_Z\})] =$$

$$[(f(x_0), 0_Y, 0_Z) - (f, g, h)(\hat{S}) - (\mathbb{R}^+ \times C \times \{0_Z\}) - (\mathbb{R}^+_0 \times C \times \{0_Z\})].$$

Thus, $x_0 \in S$ is an optimal solution for (1) iff

$$(0, 0_Y, 0_Z) \notin [(f(x_0), 0_Y, 0_Z) - (f, g, h)(\hat{S}) - (\mathbb{R}^+_0 \times C \times \{0_Z\})] - (\mathbb{R}^+ \times C \times \{0_Z\})$$

which is equivalent to (11).

With a similar reasoning as in the proof of Theorem 2 it is possible to show that $(f(x_0), 0_Y, 0_Z) \in \mathrm{cl}(\widetilde{M})$.

Proposition 5 *Let* $x_0 \in S$ *be an optimal solution of* (1). *Then* Assumption S *is satisfied iff*

$$T_{\mathscr{E}}(0, 0_Y, 0_Z) \cap (\mathbb{R}^+ \times \{0_Y\} \times \{0_Z\}) = \varnothing.$$

Proof Suppose first that $T_{\widetilde{M}}(f(x_0), 0_Y, 0_Z) \cap (\mathbb{R}^- \times \{0_Y\} \times \{0_Z\}) = \varnothing$ and, by contradiction, assume that there exists $(t, 0_Y, 0_Z) \in T_{\mathscr{E}}(0, 0_Y, 0_Z)$ with $t > 0$. Hence there exist sequences $(x_n) \subset \hat{S}$, $(y_n) \subset C$, $(\beta_n) \subset \mathbb{R}^+$, $(\alpha_n) \subset \mathbb{R}_0^+$ such that

$$\lim_n (f(x_n) + \alpha_n) = f(x_0) \quad \lim_n (g(x_n) + y_n) = 0_Y \quad \lim_n h(x_n) = 0_Z \tag{12}$$

$$\lim_n \beta_n (f(x_0) - f(x_n) - \alpha_n) = t \quad \lim_n \beta_n (g(x_n) + y_n) = 0_Y \quad \lim_n \beta_n h(x_n) = 0_Z. \tag{13}$$

Since $t > 0$ and $\beta_n > 0$ it follows that there exists $\overline{n} \in \mathbb{N}$ such that $f(x_n) < f(x_0)$ for all $n > \overline{n}$. Since x_0 is an optimal solution of the Primal Problem it follows that $x_n \in \hat{S} \setminus S$ for all $n > \overline{n}$. This means $(-t, 0_Y, 0_Z) \in T_{\widetilde{M}}(f(x_0), 0_Y, 0_Z)$ which contradicts our initial assumption.

Conversely assume that $T_{\mathscr{E}}(0, 0_Y, 0_Z) \cap (\mathbb{R}^+ \times \{0_Y\} \times \{0_Z\}) = \varnothing$. It is easily seen that

$$T_{\widetilde{M}}(f(x_0), 0_Y, 0_Z) \subset \mathrm{cl\,Cone}(\widetilde{M} - (f(x_0), 0_Y, 0_Z)) \subset -\mathrm{cl\,Cone}\,\mathscr{E} = -T_{\mathscr{E}}(0, 0_Y, 0_Z)$$

which implies that $T_{\widetilde{M}}(f(x_0), 0_Y, 0_Z) \cap (\mathbb{R}^- \times \{0_Y\} \times \{0_Z\}) = \varnothing$. □

The following lemma will be useful in the proof of the strong duality theorem.

Lemma 2 *Let* $H = \{t \in \mathbb{R} \times Y \times Z \mid \langle a, t \rangle = 0\}$ *be an hyperplane of* $\mathbb{R} \times Y \times Z$. *The following statements are equivalent:*

1. *H separates the sets* $-(f, g, h)(\hat{S}) + (f(x_0), 0_Y, 0_Z)$ *and* $\mathbb{R}^+ \times C \times \{0_Z\}$.
2. *H separates the sets* $\mathscr{E}$ *and* $\mathbb{R}^+ \times C \times \{0_Z\}$.
3. *H separates the sets* $T_{\mathscr{E}}(0, 0_Y, 0_Z)$ *and* $\mathbb{R}^+ \times C \times \{0_Z\}$.
4. *H separates the sets* $T_{\mathscr{E}}(0, 0_Y, 0_Z)$ *and* $\mathbb{R}^+ \times \{0_Y\} \times \{0_Z\}$.

Proof Let $H^+ := \{t \in \mathbb{R} \times Y \times Z : \langle a, t \rangle \geq 0\}$ and $H^- = \{t \in \mathbb{R} \times Y \times Z : \langle a, t \rangle \leq 0\}$.

- In order to prove that *1* implies *2* let us assume that $-(f, g, h)(\hat{S}) + (f(x_0), 0_Y, 0_Z) \subset H^-$ and $\mathbb{R}^+ \times \{C\} \times \{0_Z\} \subset H^+$. By proving that $\mathscr{E} \subset H^-$ the conclusion follows. Let us suppose that here exists $\hat{t} \in \mathscr{E}$ such that $\langle a, \hat{t} \rangle > 0$ and, since $\mathscr{E} := (f(x_0), 0_Y, 0_Z) - (f, g, h)(\hat{S}) - \mathbb{R}_0^+ \times \{C\} \times \{0_Z\}$ it follows that $\hat{t} = t_1 - t_2$ for some $t_1 \in (f(x_0), 0_Y, 0_Z) - (f, g, h)(\hat{S})$ and $t_2 \in \mathbb{R}_0^+ \times C \times \{0_Z\}$.

From $\langle a, \hat{t}\rangle > 0$ we get $0 \le \langle a, t_2\rangle < \langle a, t_1\rangle \le 0$, where the first inequality follows from $\mathbb{R}_0^+ \times C \times \{0_Z\} \subset H^+$, and the third one follows from $(f(x_0), 0_Y, 0_Z) - (f, g, h)(\hat{S}) \subset H^-$. This is a contradiction, hence the claim is proved.

- In order to prove that *2* implies *3* let us assume that $\mathbb{R}^+ \times C \times \{0_Z\} \subset H^+$ and $\mathscr{E} \subset H^-$ from which it follows that $T_{\mathscr{E}}(0, 0_Y, 0_Z) = \operatorname{cl}\operatorname{Cone}\mathscr{E} \subset \operatorname{cl}\operatorname{Cone} H^- = H^-$ which proves the claim.
- Since $\mathbb{R}^+ \times \{0_Y\} \times \{0_Z\} \subset \mathbb{R}^+ \times C \times \{0_Z\}$, it easily follows that *3* implies *4*.
- In order to prove that *4* implies *1* let us assume that $T_{\mathscr{E}}(0, 0_Y, 0_Z) \subset H^-$ and $\mathbb{R}^+ \times \{0_Y\} \times \{0_Z\} \subset H^+$ and, by contradiction, assume that there exists $\hat{t} \in \mathbb{R}^+ \times C \times \{0_Z\}$ such that $\langle a, \hat{t}\rangle < 0$. For a given $\bar{t} \in \mathscr{E}$ and for every $\alpha \ge 0$ we get $\bar{t} - \alpha\hat{t} \in \mathscr{E} \subset T_{\mathscr{E}}(0, 0_Y, 0_Z)$. Hence

$$\lim_{\alpha\to+\infty} \langle a, \bar{t} - \alpha\hat{t}\rangle = +\infty$$

which contradicts that $\langle a, t\rangle \le 0$ for all $t \in T_{\mathscr{E}}(0, 0_Y, 0_Z)$. This means that $\mathbb{R}^+ \times C \times \{0_Z\} \subset H^+$, and since $(f(x_0), 0_Y, 0_Z) - (f, g, h)(\hat{S}) \subset \mathscr{E} \subset T_{\mathscr{E}}(0, 0_Y, 0_Z) \subset H^-$ the claim follows. □

We can now state the main result.

Theorem 5 *Let $x_0 \in S$. Then*

$$T_{\mathscr{E}}(0, 0_Y, 0_Z) \cap (\mathbb{R}^+ \times \{0_Y\} \times \{0_Z\}) = \varnothing \tag{14}$$

iff strong duality holds for (1) and x_0 is an optimal solution of (1).

Proof Let us suppose first that (14) is fulfilled. We first note that, since $\mathscr{E}$ is convex, then $T_{\mathscr{E}}(0, 0_Y, 0_Z)$ is a closed convex set and $\mathscr{E} \subset T_{\mathscr{E}}(0, 0_Y, 0_Z)$, which yields $\mathscr{E} \cap (\mathbb{R}^+ \times \{0_Y\} \times \{0_Z\}) = \varnothing$ or, equivalently, $\mathscr{E} \cap (\mathbb{R}^+ \times C \times \{0_Z\}) = \varnothing$ and therefore x_0 is an optimal solution for (1). By (14) it follows that there exists $h \in (\mathbb{R}^+ \times \{0_Y\} \times \{0_Z\}) \setminus T_{\mathscr{E}}(0, 0_Y, 0_Z)$. By Theorem 11 in the Appendix, there exists $a = (t, u, v) \in (\mathbb{R} \times Y^* \times Z^*) \setminus \{(0, 0_{Y^*}, 0_{Z^*})\}$ such that $\langle a, z\rangle \le 0 < \langle a, h\rangle$ for all $z \in T_{\mathscr{E}}(0, 0_Y, 0_Z)$.

Since $\mathscr{E} \subset T_{\mathscr{E}}(0, 0_Y, 0_Z)$ it follows that $\langle a, z\rangle \le 0$ for all $z \in \mathscr{E}$.

Now we can prove that $\langle a, z\rangle \ge 0$ for all $z \in \mathbb{R}^+ \times C \times \{0_Z\}$. Indeed, assume that there exists $\hat{z} \in \mathbb{R}^+ \times C \times \{0_Z\}$ such that $\langle a, \hat{z}\rangle < 0$ and consider an arbitrary $\bar{z} \in \mathscr{E}$. It follows that $\bar{z} - \alpha\hat{z} \in \mathscr{E}$ for all $\alpha \ge 0$, hence

$$\lim_{\alpha\to+\infty} \langle a, \bar{z} - \alpha\hat{z}\rangle = +\infty$$

which is a contradiction. Hence by Lemma 2 (*3* ⇒ *1*) we get

$$\langle a, z\rangle \ge 0 \quad \forall z \in \mathbb{R}^+ \times C \times \{0_Z\} \text{ and } \langle a, z\rangle \le 0 \quad \forall z \in -(f, g, h)(\hat{S}) + (f(x_0), 0_Y, 0_Z)$$

so the hyperplane $H = \{z \in \mathbb{R} \times Y \times Z | \langle a, z\rangle = 0\}$ separates the sets $\mathbb{R}^+ \times C \times \{0_Z\}$ and $-(f, g, h)(\hat{S}) + (f(x_0), 0_Y, 0_Z)$. Hence we get the following inequalities:

$$tr + \langle u, y\rangle \geq 0 \quad \forall r \in \mathbb{R}^+,\ \forall y \in C \tag{15}$$

$$t(f(x_0) - f(x)) + \langle u, -g(x) - y\rangle + \langle v, -h(x)\rangle \leq 0 \quad \forall x \in \hat{S},\ \forall y \in C. \tag{16}$$

Inequality (15) implies that $u \in C^*$ and $t \geq 0$.

If one assumes that $t = 0$ then $\langle a, f\rangle = 0$ for all $f \in \mathbb{R}^+ \times \{0_Y\} \times \{0_Z\}$, but this contradicts the fact that there exists $h \in \mathbb{R}^+ \times \{0_Y\} \times \{0_Z\}$ such that $\langle a, h\rangle > 0$. Hence t must be strictly positive.

Taking $y = 0_Y$, $x = x_0$ and $u_0 = \frac{u}{t} \in C^*$, $v_0 = \dfrac{v}{t} \in Z^*$ and substituting in (16) we get $\langle u_0, g(x_0)\rangle \geq 0$, and since $g(x_0) \in -C$ and $u_0 \in C^*$ it follows $\langle u_0, g(x_0)\rangle = 0$. Hence

$$f(x_0) = \min_{x \in \hat{S}} f(x) + \langle u_0, g(x)\rangle + \langle v_0, h(x)\rangle$$

which means that the Primal and the Dual Problem have the same optimal values and (u_0, v_0) is an optimal solution of the Dual.

To prove the necessary part, suppose that strong duality holds and x_0 is an optimal solution for (1). Let $(u_0, v_0) \in C^* \times Z^*$ be an optimal solution of the Dual Problem. Thus,

$$f(x) - f(x_0) + \langle u_0, g(x)\rangle + \langle v_0, h(x)\rangle \geq 0 \qquad \forall x \in \hat{S},$$

hence the hyperplane $H = \{(r, y, z) \in \mathbb{R} \times Y \times Z | r + \langle u_0, y\rangle + \langle v_0, z\rangle = 0\}$ separates the sets $-(f, g, h)(\hat{S}) + (f(x_0), 0_Y, 0_Z)$ and $\mathbb{R}^+ \times \{0_Y\} \times \{0_Z\}$ which means $-(f, g, h)(\hat{S}) + (f(x_0), 0_Y, 0_Z) \subset H^-$ and $\mathbb{R}^+ \times \{0_Y\} \times \{0_Z\} \subset H^+$. Note that since $\mathbb{R}^+ \times C \times \{0_Z\} \subset H^+$ then statement *1* of Lemma 2 holds. Therefore, statement *4* of Lemma 2 holds too, i.e.,

$$T_{\mathscr{E}}(0, 0_Y, 0_Z) \subset H^- \text{ and } \mathbb{R}^+ \times \{0_Y\} \times \{0_Z\} \subset H^+.$$

Since $\mathbb{R}^+ \times \{0_Y\} \times \{0_Z\} \subset H^+ \setminus H$, then (14) is fulfilled.

From Theorems 5 and 3 we immediately obtain the following result.

Theorem 6 *Let $x_0 \in \hat{S}$. Then there exists $(\overline{u}, \overline{v}) \in (C^* \times Z^*)$ such that $(x_0, \overline{u}, \overline{v})$ is a saddle point of the Lagrange functional L iff (14) is fulfilled and $x_0 \in S$.*

The reader who is interested in further developments of these results can refer to [4] or to the other papers cited in the introduction. As an application of the previous analysis we investigate in the following section a property of the so called Nash-Rosen equilibria.

4 Application to Generalized Nash Equilibrium Problems in Infinite Dimensional Spaces

In his seminal paper [25], Rosen introduced a new class of Nash equilibria (since then known as Rosen equilibria) which have been proved very useful in several applications. More recently, [10], Rosen equilibria have been investigated in the light of variational inequalities theory. In this section we show as the recent advancements in duality theory described in Sect. 3

4.1 The Setting of the Game

We describe the setting of our Nash game. For simplicity we deal with two players (the case of N players being easily deduced).

Assume that X_1 and X_2 are two Banach spaces, and denote by $X = X_1 \times X_2$ the product space and by $u = (u^1, u^2)$ the generic element of X, that is u^1 and u^2 are the variables respectively controlled by the first and the second player. Let also $K \subset X$ be a non empty, convex set, J_1 and $J_2 : X \to \mathbb{R}$ two functionals such that $J_1(\cdot, u^2)$ is convex and Gâteaux differentiable for every $u^2 \in X_2$ and $J_2(u^1, \cdot)$ is convex and Gâteaux differentiable for every $u^1 \in X_1$. Any of these functions is called the utility function of the player i or the payoff function or the loss function depending on the particular application in which the GNEP arises.

For every $u = (u^1, u^2) \in X$, the feasible strategies' sets of the two players are of the following kind:

$$K_1(u) = \{v^1 \in X_1 : (v^1, u^2) \in K\} \subset X_1$$

and

$$K_2(u) = \{v^2 \in X_2 : (u^1, v^2) \in K\} \subset X_2.$$

Notice that if $u \in K$ then the above sets are non empty ($u^i \in K_i(u)$) and convex. This class of strategies' sets, introduced by Rosen in [25], is often referred to as the *jointly convex case* or GNEPs *with coupled constraints* motivated by the fact that the feasible sets are linked through a shared or common constraint.

The goal of each player i, given the strategy of the rival, is to choose a strategy which minimizes the function J_i on its feasible set. The following definition describes the aim of the game: to find an *equilibrium* point for both players, that is a vector $(\bar{u}^1, \bar{u}^2)$ such that no player can decrease his utility function by changing unilaterally $\bar{u}^i$ to any other feasible point.

Definition 6 We say that $\bar{u} = (\bar{u}^1, \bar{u}^2)$ is a generalized Nash equilibrium or a solution of the generalized Nash equilibrium problem (in short GNEP) if $\bar{u} \in K$ and the following conditions hold:

$$\begin{cases} J_1(\bar{u}^1, \bar{u}^2) = \min_{u^1 \in K_1(\bar{u})} J_1(u^1, \bar{u}^2), \\ J_2(\bar{u}^1, \bar{u}^2) = \min_{u^2 \in K_2(\bar{u})} J_2(\bar{u}^1, u^2) \end{cases} \tag{17}$$

We recall that if Y is a Banach space, a function $I : Y \to \mathbb{R}$ is said to be Gâteaux differentiable in $\bar{u} \in Y$ if there exists $\varphi \in Y^\star$ (the topological dual space of Y) such that

$$\lim_{\lambda \to 0^+} \frac{I(\bar{u} + \lambda u) - I(\bar{u})}{\lambda} = \varphi(u) \qquad \forall\, u \in Y.$$

The functional φ is called the Gâteaux derivative of I and denoted by $\varphi \equiv DI(\bar{u})$.

Remark 1 By well known results of convex analysis, (see e.g. Theorem 3.8 of [19]), $\bar{u} = (\bar{u}^1, \bar{u}^2)$ is a solution of GNEP iff $\bar{u} \in K$ and

$$\begin{cases} D_1 J_1(\bar{u}^1, \bar{u}^2)(u^1 - \bar{u}^1) \geq 0 & \forall\, u^1 \in K_1(\bar{u}), \\ D_2 J_2(\bar{u}^1, \bar{u}^2)(u^2 - \bar{u}^2) \geq 0 & \forall\, u^2 \in K_2(\bar{u}) \end{cases} \tag{18}$$

where D_1 and D_2 stand for the Gâteaux derivative of $J_1(\cdot, \bar{u}^2)$ and $J_2(\bar{u}^1, \cdot)$ respectively.

Denote by $\Gamma : X \to X_1^\star \times X_2^\star$ the mapping

$$\Gamma(u^1, u^2) = \begin{pmatrix} D_1 J_1(u^1, u^2) \\ D_2 J_2(u^1, u^2) \end{pmatrix}. \tag{19}$$

With the above notation, it is clear that (18) are equivalent to

$$\Gamma(\bar{u})^T (u - \bar{u}) \geq 0 \qquad \forall u \in K_1(\bar{u}) \times K_2(\bar{u}).$$

Since the convex sets $K_i(\bar{u})$ depend on the solution, one obtains that a GNEP can be reformulated as a quasi-variational inequality. Following [10], the nature of the feasible sets of the strategies of the two players allows to reduce the problem to a variational inequality. Solving the variational inequality associated to Γ and the convex set K (in short, VI(Γ,K)), means finding a point $\bar{u} = (\bar{u}^1, \bar{u}^2) \in K$ such that

$$\Gamma(\bar{u})^T (u - \bar{u}) \geq 0 \qquad \forall u \in K. \tag{20}$$

Analogously to [10], we have the following

Theorem 7 *Every solution of the variational inequality* VI(Γ,K) *is a solution of* GNEP.

Proof Let $\bar{u} = (\bar{u}^1, \bar{u}^2) \in K$ be a solution of (20) where Γ is as in (19). If $u^1 \in K_1(\bar{u})$, then $u = (u^1, \bar{u}^2) \in K$ and from the definition of Γ, we get

$$0 \leq \Gamma(\bar{u})^T(u - \bar{u}) = D_1 J_1(\bar{u}^1, \bar{u}^2)(u^1 - \bar{u}^1)$$

which is the first of the (18). In a similar way we get the second inequality of (18). □

A solution of the GNEP that is also a solution of VI(Γ,K) is usually referred to as a *variational equilibrium.*

4.2 Lagrange Multipliers Rule

In the previous section we have proved that a solution of the GNEP can be obtained as a solution of the VI(Γ,K). By adopting this reduction method we can lose solutions of the GNEP. In the present section we investigate which kind of solutions are preserved for a special constraints set. As in the finite dimensional case, we can prove that a solution of the GNEP is a variational equilibrium if and only if the shared constraints have the same multipliers. We underline that our result holds under any constraints qualification condition.

We will assume also that Y is a Banach space ordered by a convex cone C, $g : X \to Y$ is a convex, continuously Gâteaux differentiable mapping and

$$K = \{u \in X : g(u) \in -C\}.$$

If $f : X \to \mathbb{R}$ and $\bar{u} \in K$, we say that $\bar{u}$ is a solution of the minimal problem $(P_{f,K})$ if

$$f(\bar{u}) = \min_K f.$$

Our main result is the following:

Theorem 8

(i) Let $\bar{u}$ be a solution of the VI(Γ,K) *such that a suitable constraints qualification condition (for the* VI(Γ,K)*) holds at $\bar{u}$. Then, $\bar{u}$ is a solution of GNEP such that both players share the same Lagrange multiplier.*

(ii) Let $\bar{u}$ be a solution of GNEP such that a constraints qualification condition (for the GNEP) holds at $\bar{u}$ and both players share the same Lagrange multiplier. Then, $\bar{u}$ is a solution of the VI(Γ,K).

Proof

(i) Assume that $\bar{u}$ is a solution of the VI(Γ,K). Then, if $f : X \to \mathbb{R}$ is the function defined by

$$f(u) = \Gamma(\bar{u})^T(u - \bar{u}), \tag{21}$$

f is convex, Gâteaux differentiable with derivative given by $Df(u)(z) = \Gamma(\bar{u})^T(z)$ for all $z \in X$ and for all $u \in X$ and

$$f(\bar{u}) = \min_K f = 0.$$

Under a suitable constraints qualification condition, there exists $\bar{w} \in C^\star$ such that

$$0 = Df(\bar{u}) + \bar{w} \circ Dg(\bar{u}) = \Gamma(\bar{u})^T + \bar{w} \circ Dg(\bar{u}), \tag{22}$$

and

$$\langle \bar{w}, g(\bar{u}) \rangle_{Y^\star, Y} = 0. \tag{23}$$

Since $g \in C^1(X, Y)$, $Dg(\bar{u})u = D_1 g(\bar{u})u^1 + D_2 g(\bar{u})u^2$, (22) can be rewritten as

$$D_1 J_1(\bar{u})u^1 + D_2 J_2(\bar{u})u^2 + \bar{w}(D_1 g(\bar{u})u^1) + \bar{w}(D_2 g(\bar{u})u^2) = 0 \quad \forall (u^1, u^2) \in X$$

and for the arbitrariness of $(u^1, u^2) \in X$, (22) and (23) read as

(α) $D_1 J_1(\bar{u}) + \bar{w} \circ D_1 g(\bar{u}) = 0, \qquad D_2 J_2(\bar{u}) + \bar{w} \circ D_2 g(\bar{u}) = 0,$
(β) $\langle \bar{w}, g(\bar{u}) \rangle_{Y^\star, Y} = 0.$

If $g_1 : X_1 \to Y$ is the mapping $g_1(u^1) = g(u^1, \bar{u}^2)$, then the set $K_1(\bar{u})$ can be written as $K_1(\bar{u}) = \{u^1 : g_1(u^1) \in -C\}$ and analogously, if $g_2 : X_2 \to Y$ is defined by $g_2(u^2) = g(\bar{u}^1, u^2)$, then $K_2(\bar{u}) = \{u^2 : g_2(u^2) \in -C\}$. One has also that $Dg_i(\bar{u}^i) = D_i g(\bar{u})$ and $g_i(\bar{u}) = g(\bar{u})$, $i = 1, 2$.

Then, (α) and (β) can be rewritten as

$$D_1 J_1(\bar{u}) + \bar{w} \circ Dg_1(\bar{u}^1) = 0,$$

$$D_2 J_2(\bar{u}) + \bar{w} \circ Dg_2(\bar{u}^2) = 0,$$

$$\langle \bar{w}, g_1(\bar{u}) \rangle_{Y^\star, Y} = \langle \bar{w}, g_2(\bar{u}) \rangle_{Y^\star, Y} = 0.$$

This means that $\bar{u}$ satisfies the Lagrange multipliers rule for the GNEP, and $\bar{w}$ is the multiplier for both players. These conditions guarantee (see Corollary 5.15 of [19]) that $\bar{u}$ is a minimal solution of the problems $(P_{f,K})$ with $(f, K) = (J_1, K_1(\bar{u}))$ and $(f, K) = (J_2, K(\bar{u}))$ respectively, that is $\bar{u}$ is a solution of GNEP and both players share the same Lagrange multiplier.

(ii) Assume that $\bar{u}$ is a solution of GNEP and some constraints qualification holds at $\bar{u}$. If the two players share the same Lagrange multipliers, then

(α_1) $D_1 J_1(\bar{u}) + \bar{w} \circ Dg_1(\bar{u}^1) = 0,$
(β_1) $\langle \bar{w}, g_1(\bar{u}^1) \rangle_{Y^\star, Y} = 0.$

and

$$\begin{array}{ll} (\alpha_2) & D_2 J_2(\bar{u}) + \bar{w} \circ Dg_2(\bar{u}^2) = 0, \\ (\beta) & \langle \bar{w}, g_2(\bar{u}^2) \rangle_{Y^\star, Y} = 0. \end{array}$$

Then, it is clear that (α) and (β) are satisfied. From Corollary 5.15 of [19], we get that $\bar{u}$ is a minimal solution of problem $(P_{f,K})$ with f as in (21). This implies that $\bar{u}$ is a solution of the VI(Γ,K). □

4.3 The Role of Assumption S

We need first to state, in a convenient form for our purposes, an infinite dimensional Lagrange multipliers rule for convex optimization problems, proved recently in [9].

Let us recall that

$$K = \{u \in X : g(u) \in -C\},$$

and $\bar{u} \in K$. Denote by

$$\tilde{M} = \{(f(u) - f(\bar{u}) + \alpha, g(u) + z) : \ u \in X \setminus K, \alpha \geq 0, z \in C\} \subseteq \mathbb{R} \times Y.$$

Theorem 9 ([9], Theorem 3) *Let X be a normed space, Y a Banach space ordered by a convex cone C. Let $f : X \to \mathbb{R}$ be a convex, Gâteaux differentiable functional, $g : X \to Y$ a convex, Gâteaux differentiable mapping. Denote by*

$$K = \{u \in X : g(u) \in -C\}.$$

Assume that $\bar{u} \in K$ is a solution of the minimal problem $(P_{f,K})$:

$$f(\bar{u}) = \min_K f$$

and Assumption S is fulfilled at $\bar{u}$. Then, there exists $\bar{w} \in C^\star$ such that

$$Df(\bar{u}) + \bar{w} \circ Dg(\bar{u}) = 0 \tag{24}$$

and

$$\langle \bar{w}, g(\bar{u}) \rangle_{Y^\star, Y} = 0. \tag{25}$$

Conversely, if (24) and (25) hold, then $\bar{u}$ is the minimal solution of problem $(P_{f,K})$ and Assumption S is fulfilled at $\bar{u}$.

We will refer to $\bar{w}$ as the Lagrange multiplier.

Remark 2 Notice that conditions (24) and (25) of the above theorem are the counterpart in the infinite dimensional case of the well known KKT conditions.

We now formulate the *Assumption S* for the VI(Γ,K) and for the GNEP.

Assume that $\bar{u}$ is a solution of VI(Γ,K). As in the proof of Theorem 8, if $f : X \to \mathbb{R}$ is the function defined in (21), then

$$f(\bar{u}) = \min_{u \in K} f(u) = 0.$$

The set $\tilde{M}$ is defined by

$$\tilde{M}=\{(D_1 J_1(\bar{u})(u^1-\bar{u}^1)+D_2 J_2(\bar{u})(u^2-\bar{u}^2)+\alpha, g(u)+z) : u \in X\backslash K, \alpha \geq 0, z \in C\}$$

and

$$\begin{aligned} T_{\tilde{M}}(0, \theta_Y) = \{(l, u) \in \mathbb{R} \times Y :\ & \\ & l = \lim_k \lambda_k [D_1 J_1(\bar{u})(u_k^1 - \bar{u}^1) + D_2 J_2(\bar{u})(u_k^2 - \bar{u}^2) + \alpha_k], \\ & u = \lim_k \lambda_k [g(u_k) + z_k], \\ & \lim_k [D_1 J_1(\bar{u})(u_k^1 - \bar{u}^1) + D_2 J_2(\bar{u})(u_k^2 - \bar{u}^2) + \alpha_k] = 0, \\ & \lim_k [g(u_k) + z_k] = \theta_Y, \\ & \lambda_k > 0, \alpha_k \geq 0, u_k \in X \setminus K, z_k \in C\}. \end{aligned}$$

Definition 7 We say that *Assumption S* holds at $\bar{u}$ for the VI(Γ, K) if $(l, \theta_Y) \in T_{\tilde{M}}(0, \theta_Y)$ implies that $l \geq 0$.

Assume that $\bar{u}$ is a solution of GNEP. This means that $\bar{u}$ verifies the following:

$$\begin{cases} J_1(\bar{u}^1, \bar{u}^2) = \min_{u^1 \in K_1(\bar{u})} J_1(u^1, \bar{u}^2), \\ J_2(\bar{u}^1, \bar{u}^2) = \min_{u^2 \in K_2(\bar{u})} J_2(\bar{u}^1, u^2) \end{cases}$$

In this case we have two sets, $\tilde{M}_1$ and $\tilde{M}_2$ defined by

$$\tilde{M}_1 = \{(J_1(u^1, \bar{u}^2) - J_1(\bar{u}) + \alpha, g(u^1, \bar{u}^2) + z) : u^1 \in X_1 \setminus K_1(\bar{u}), \alpha \geq 0, z \in C\}$$

and

$$\tilde{M}_2 = \{(J_2(\bar{u}^1, u^2) - J_2(\bar{u}) + \alpha, g(\bar{u}^1, u^2) + z) : u^2 \in X_2 \setminus K_2(\bar{u}), \alpha \geq 0, z \in C\}.$$

The tangent cone to $\tilde{M}_1$ at $(0, \theta_Y)$ is the set

$$\begin{aligned} T_{\tilde{M}_1}(0,\theta_Y) = \{(l,u) \in \mathbb{R} \times Y :\ & \\ & l = \lim_k \lambda_k [J_1(u_k^1, \bar{u}^2) - J_1(\bar{u}) + \alpha_k], \\ & u = \lim_k \lambda_k [g(u_k^1, \bar{u}^2) + z_k], \\ & \lim_k [J_1(u_k^1, \bar{u}^2) - J_1(\bar{u}) + \alpha_k] = 0, \\ & \lim_k [g(u_k^1, \bar{u}^2) + z_k] = \theta_Y, \\ & \lambda_k > 0, \alpha_k \geq 0, u_k^1 \in X^1 \setminus K_1(\bar{u}), z_k \in C\}. \end{aligned}$$

An analogous frame holds for $T_{\tilde{M}_2}(0,\theta_Y)$.

Definition 8 We say that *Assumption S* holds at $\bar{u}$ for GNEP if $(l_1,\theta_Y) \in T_{\tilde{M}_1}(0,\theta_Y)$ implies that $l_1 \geq 0$ and $(l_2,\theta_Y) \in T_{\tilde{M}_2}(0,\theta_Y)$ implies that $l_2 \geq 0$.

Remark 3 One has that $T_{\tilde{M}_i}(0,\theta_Y) \subseteq T_{\tilde{M}}(0,\theta_Y)$.

Let us prove the claim for $i = 1$. Indeed, if $(l,u) \in T_{\tilde{M}_1}(0,\theta_Y)$, there exist sequences $\{\lambda_k\}, \{\alpha_k\}, \{u_k^1\}, \{z_k\}$ such that $\lambda_k > 0, \alpha_k \geq 0, u_k^1 \in X_1 \setminus K_1(\bar{u}), z_k \in C$ for every $k \in N$ and $l = \lim_k \lambda_k [J_1(u_k^1, \bar{u}^2) - J_1(\bar{u}) + \alpha_k]$, $u = \lim_k \lambda_k [g(u_k^1, \bar{u}^2) + z_k]$, $\lim_k [J_1(u_k^1, \bar{u}^2) - J_1(\bar{u}) + \alpha_k] = 0$, and $\lim_k [g(u_k^1, \bar{u}^2) + z_k] = \theta_Y$. From the convexity of $J_1(\cdot, \bar{u}^2)$, one has

$$J_1(u_k^1, \bar{u}^2) - J_1(\bar{u}) \geq D_1 J_1(\bar{u})(u_k^1 - \bar{u}^1).$$

If we define $\beta_k = J_1(u_k^1, \bar{u}^2) - J_1(\bar{u}) - D_1 J_1(\bar{u})(u_k^1 - \bar{u}^1) + \alpha_k \geq 0$ and $u_k = (u_k^1, \bar{u}^2) \in X \setminus K$ then, it is immediately seen that $(l,u) \in T_{\tilde{M}}(0,\theta_Y)$.

Remark 4 From the previous Remark we get that if *Assumption S* holds for VI(Γ,K) at $\bar{u}$, solution of VI(Γ,K), then it holds also for the GNEP at $\bar{u}$.

From Theorem 8, it follows at once

Corollary 2

(i) Let $\bar{u}$ be a solution of the VI(Γ,K) *such that Assumption S holds at $\bar{u}$. Then, $\bar{u}$ is a solution of GNEP such that both players share the same Lagrange multiplier.*
(ii) Let $\bar{u}$ be a solution of GNEP such that Assumption S holds at $\bar{u}$ and both players share the same Lagrange multiplier. Then, $\bar{u}$ is a solution of the VI(Γ,K).

The above framework can be successfully applied to GNEP in Lebesgue spaces. In this respect, the main task is the verification of the validity of assumption S. For the technical details which concern this aspect we refer the interested reader to [11].

5 Conclusion and Further Research Directions

In this article we first explained in detail the main ideas and theorems of the classical Lagrangian duality theory in infinite dimension and then focused on a recent approach based on the so called assumption S which does not require that the ordering cone used to describe the inequality constraints has empty interior. This new approach has been refined and improved by different authors, mainly from the theoretical point of view. As we have mentioned in the last section on Nash equilibrium problems, the verification of assumption S is usually the difficult part even for relatively simple constraints sets. In this respect, the investigation of classes of sets for which assumption S holds has still to be carried out in a systematic manner. However, given that this assumption is both necessary and sufficient for strong duality, the use of sufficient conditions (see e.g. [4]) which imply it seems to be a more promising research avenue.

Acknowledgements The work of F. Raciti was partially supported from University of Pisa (grant PRA-2017-05).

Appendix

Theorem 10 *Let S and T be convex subsets of a real topological vector space X with* $\operatorname{int} S \neq \varnothing$.

We get $\operatorname{int} S \cap T = \varnothing$ *iff there are a non null continuous linear functional $l : X^* \to \mathbb{R}$ and $\gamma \in \mathbb{R}$ such that*

$$\langle l, s\rangle \leq \gamma \leq \langle l, t\rangle \quad \forall s \in S, \forall t \in T \quad \textit{and} \quad \langle l, s\rangle < \gamma \quad \forall s \in \operatorname{int} S.$$

Theorem 11 *Let S be a closed and convex subset of a locally convex linear space. We get that $x \in X \setminus S$ iff it exists a linear and continuous functional $l \in X^* \setminus \{0_{X^*}\}$ such that:*

$$l(x) < \inf_{s \in S} l(s).$$

Moreover, if S is a cone we have that it exists $l \in C^* \setminus \{0_{X^*}\}$ such that:

$$l(x) < 0 \leq l(s), \ \forall s \in S.$$

References

1. J.M. Borwein, R. Goebel, Notions of relative interior in Banach spaces. J. Math. Sci. **115**(4), 2542–2553 (2003)
2. J.M. Borwein, A.S. Lewis, Partially finite convex programming, part I: quasi relative interior and duality theory. Math. Program. **57**, 15–48 (1992)
3. R.I. Bot, G. Wanka, An alternative formulation for a new closed cone constraint qualification. Nonlinear Anal. **64**(6), 1367–1381 (2006)
4. R.I. Bot, E.R. Csetnek, A. Moldovan, Revisiting some duality theorems via the quasirelative interior in convex optimization. J. Optim. Theory Appl. **139**, 67–84 (2008)
5. R.I. Bot, E.R. Csetnek, G. Wanka, Regularity conditions via quasi-relative interior in convex programming. SIAM J. Optim. **19**(1), 217–233 (2008)
6. P. Daniele, S. Giuffrè, General infinite dimensional duality theory and applications to evolutionary network equilibrium problems. Optim. Lett. **1**(3), 227–243 (2007)
7. P. Daniele, S. Giuffrè, G. Idone, A. Maugeri, Infinite dimensional duality and applications. Math. Ann. **339**(1), 221–239 (2007)
8. P. Daniele, S. Giuffrè, A. Maugeri, F. Raciti, Duality theory and applications to unilateral problems. J. Optim. Theory Appl. **162**, 718–734 (2014)
9. M.B. Donato, The infinite dimensional Lagrange multiplier rule for convex optimization problems. J. Funct. Anal. **261**, 2083–2093 (2011)
10. F. Facchinei, A. Fischer, V. Piccialli, On generalized Nash games and variational inequalities. Oper. Res. Lett. **35**, 159–164 (2007)
11. F. Faraci, F. Raciti, On generalized Nash equilibrium in infinite dimension: the Lagrange multipliers approach. Optimization **64**(2), 321–338 (2015)
12. F. Faraci, B. Jadamba, F. Raciti, On stochastic variational inequalities with mean value constraints. J. Optim. Theory Appl. **171**(2), 675–693 (2016)
13. F. Flores-Bazán, G. Mastroeni, Strong duality in cone constrained nonconvex optimization. SIAM J. Optim. **23**(1), 153–169 (2013)
14. F. Giannessi, *Constrained Optimization and Image Space Analysis, Vol. 1. Separation of Sets and Optimality Conditions*. Mathematical Concepts and Methods in Science and Engineering (Springer, New York, 2005)
15. M.S. Gowda, M. Teboulle, A comparison of constraint qualifications in infinite-dimensional convex programming. SIAM J. Control Optim. **28**, 925–935 (1990)
16. A. Grad, Quasi-relative interior-type constraints qualifications ensuring strong Lagrange duality for optimization problems with cone and affine constraints. J. Math. Anal. Appl. **364**, 86–95 (2010)
17. B. Jadamba, F. Raciti, Variational inequality approach to stochastic Nash equilibrium problems with an application to Cournot oligopoly. J. Optim. Theory Appl. **165**, 1050–1070 (2015)
18. B. Jadamba, F. Raciti, On the modelling of some environmental games with uncertain data. J. Optim. Theory Appl. **167**(3), 959–968 (2015)
19. J. Jahn, *Introduction to the Theory of Nonlinear Optimization* (Springer, Berlin, 1996)
20. V. Jeykumar, H. Wolkowicz, Generalizations of Slater's constraint qualification for infinite convex programs. Math. Progr. **57**, 85–101 (1992)
21. A.A. Khan, M. Sama, A new conical regularization for some optimization and optimal control problems: convergence analysis and finite element discretization. Numer. Funct. Anal. Opt. **34**(8), 861–895 (2013)
22. M.A. Limber, R.K. Goodrich, Quasi interiors, Lagrange multipliers, and L^p spectral estimation with lattice bounds. J. Optim. Theory Appl. **78**(1), 143–161 (1993)
23. A. Maugeri, F. Raciti, On general infinite dimensional complementarity problems. Optim. Lett. **2**, 71–90 (2008)
24. A. Maugeri, F. Raciti, Remarks on infinite dimensional duality. J. Glob. Optim. **46**, 581–588 (2010)

25. J.B. Rosen, Existence and uniqueness of equilibrium points for concave n person games. Econometrica **33**, 520–534 (1965)
26. C. Zălinescu, Convex analysis in general vector spaces (World Scientific, Singapore, 2002)
27. C. Zălinescu, Private communications: September 4th, 5th, 7th, October 10th (2007)

Stability Analysis of the Inverse Problem of Parameter Identification in Mixed Variational Problems

M. Cho, A. A. Khan, T. Malysheva, M. Sama, and L. White

1 Introduction

Mixed variational problems involving variable parameters emerge from a variety of applied models. In this work, we study the inverse problem of estimating such parameters from a measurement of the solution of a mixed variational problem. The primary impetus of this work stems from the elasticity imaging inverse problem which uses the discrepancy in the elasticity properties of healthy and unhealthy tissues to locate cancerous tumors. The sought parameters in this application are the Lamé parameters in a system of linear elasticity equations describing the response of a body/traction force applied to an elastic object. Most works model the human body as nearly incompressible, and the necessity to avoid the so-called locking effect in numerical computations leads to an identification problem in a mixed variational problem. We will give more details of this model shortly.

M. Cho · A. A. Khan (✉)
Center for Applied and Computational Mathematics, School of Mathematical Sciences, Rochester Institute of Technology, Rochester, NY, USA
e-mail: mxcsma1@rit.edu; aaksma@rit.edu

T. Malysheva
Department of Natural and Applied Sciences, University of Wisconsin-Green Bay, Green Bay, WI, USA
e-mail: malyshet@uwgb.edu

M. Sama
Departamento de Matemática Aplicada, Universidad Nacional de Educación a Distancia, Madrid, Spain
e-mail: msama@ind.uned.es

L. White
Department of Mathematics, University of Oklahoma, Norman, OK, USA
e-mail: lwhite@ou.edu

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_4

We conduct this study in a general framework, and consequently, our approach applies to other models as well. In fact, using our results, we will give new formulations to study identification problems for nearly incompressible Stoke's equations. The field of inverse problems is currently among the most vibrant and expanding branches of applied mathematics, and we refer the interested reader to [4, 5, 8, 9, 12, 14, 15, 17, 19–22, 24, 25, 27–29, 33].

To facilitate the discussion, consider the elliptic boundary value problem (BVP)

$$-\nabla \cdot (q\nabla u) = f \text{ in } \Omega, \; u = 0 \text{ on } \partial\Omega, \tag{1}$$

where Ω is a suitable domain in $\mathbb{R}^2$ or $\mathbb{R}^3$ and $\partial\Omega$ is its boundary. Interesting real-world problems lead to (1). For instance, in (1), $u = u(x)$ may represent the steady-state temperature at a given point x of a body; then q would be a variable thermal conductivity coefficient, and f the external heat source. The system (1) also models from underground steady state aquifers in which the parameter q is the aquifer transmissivity coefficient, u is the hydraulic head, and f is the recharge.

To study the inverse problem of estimating q from a measurement z of the solution u of (1), there are mainly two approaches, namely, either regarding (1) as a hyperbolic PDE in q or formulating an optimization problem to estimate q. Among the optimization-based techniques, the output least-squares (OLS) method minimizes

$$q \to \|u(q) - z\|^2, \tag{2}$$

where z is the data and $u(q)$ solves the variational form of (1) given by

$$\int_\Omega q\nabla u \cdot \nabla v = \int_\Omega f v, \quad \text{for all } v \in H_0^1(\Omega). \tag{3}$$

As a variant of the OLS method, Knowles [26] proposed minimizing

$$q \to \int q\nabla(u(q) - z) \cdot \nabla(u(q) - z), \tag{4}$$

where z is the measurement of u and $u(q)$ solves (3). Although OLS is typically nonconvex, Knowles [26] showed that (4) is convex. For identification in variational problems, in [18], an extension of (4) was proposed and its convexity was proved in an abstract setting. Motivated by [18], Jadamba et al. [23] introduced a new modified output least-squares (MOLS) for the elasticity imaging inverse problem. Besides the MOLS formulation, there are three other optimization formulations for the elasticity imaging inverse problem, namely, the output least squares (OLS) formulation, the energy output least-squares (EOLS) approach (see [13]), and the equation error approach (see [12]).

In this work, we investigate various aspects of the MOLS functional introduced in [23], and contrast them with the corresponding features of the OLS approach and the EOLS approach. The main contributions are:

1. We develop regularization frameworks for the identification of smooth as well as nonsmooth parameters. We collect existence results and give new optimality conditions for optimization problems with the regularized OLS, MOLS, and EOLS objectives.
2. We conduct a thorough study of stability aspects of the inverse problem under data perturbation. This includes a new stability result showing that a sequence of the regularized problems involving noisy data, approximates the regularized problem with the exact data, given that the noise diminishes. A related result shows that the family of regularized problems converges to the original problem if the regularization parameter converges to zero in a controlled manner.
3. We give new stability estimates for general inverse problems using the regularized OLS, MOLS, and EOLS formulations. These results reflect on the selection of an optimal regularization parameter for the continuous dependence on the data perturbation. We present applications of the theoretical results and conduct numerical experiments.

The contents of this paper are organized into seven sections. In Sect. 2 we give various bounds on the solution of a mixed variational problem. Section 3 describes three objective functionals, namely, the OLS, MOLS, and ELOS objectives in an abstract setting. In Sect. 3 we describe a general regularization framework. The regularized optimization problems are studied in a nonreflexive Banach space setting as well as in a Hilbert space setting. We give existence results and derive optimality conditions. In Sect. 4, we examine the impact of some perturbation in the data. Section 5 provides a thorough study of the various stability issues of the OLS, the MOLS, and the EOLS formulations. Section 6 presents some numerical examples. The paper concludes with some general remarks.

2 Problem Formulation and Preliminary Results

Let $\widehat{V}$ and Q be real Hilbert spaces, let B be a real Banach space, let $S \subset B$ be open, and convex, and let $A \subset S$ be closed, and convex. Let $a : B \times \widehat{V} \times \widehat{V} \to \mathbb{R}$ be a trilinear form symmetric in the last two arguments, let $b : \widehat{V} \times Q \to \mathbb{R}$ be a bilinear map, let $c : Q \times Q \to \mathbb{R}$ be a symmetric bilinear map, and let $m : \widehat{V} \to \mathbb{R}$ be a linear and continuous map. Assume that there are positive constants $\kappa_1, \kappa_2, \varsigma_1, \varsigma_2$, and κ_0 such that for every $\ell \in S$, $p, q \in Q$, and $\bar{u}, \bar{v} \in \widehat{V}$, we have

$$a(\ell, \bar{v}, \bar{v}) \geq \kappa_1 \|\bar{v}\|^2_{\widehat{V}}, \tag{5a}$$

$$|a(\ell, \bar{u}, \bar{v})| \leq \kappa_2 \|\ell\|_B \|\bar{u}\|_{\widehat{V}} \|\bar{v}\|_{\widehat{V}}, \tag{5b}$$

$$c(q, q) \geq \varsigma_1 \|q\|_Q^2, \tag{5c}$$

$$|c(p, q)| \leq \varsigma_2 \|p\|_Q \|q\|_Q, \tag{5d}$$

$$|b(\bar{v}, q)| \leq \kappa_0 \|\bar{v}\|_{\widehat{V}} \|q\|_Q. \tag{5e}$$

Consider the mixed variational problem: Given $\ell \in S$, find $(\bar{u}, p) \in \widehat{V} \times Q$ with

$$a(\ell, \bar{u}, \bar{v}) + b(\bar{v}, p) = m(\bar{v}), \quad \text{for every } \bar{v} \in \widehat{V}, \tag{6a}$$

$$b(\bar{u}, q) - c(p, q) = 0, \quad \text{for every } q \in Q. \tag{6b}$$

Our focus is on the inverse problem of identifying the parameter $\ell \in A$ for which a solution $(\bar{u}, p)$ of (6) is *closest* to a given measurement $(\bar{z}, \hat{z})$ of $(\bar{u}, p)$.

In view of the coercivity and the continuity of the forms $a(\cdot, \cdot, \cdot)$ and $c(\cdot, \cdot)$, the Lax-Milgram lemma ensures that for every $\ell \in S$, there exists a unique $u = u(\ell) = (\bar{u}(\ell), p(\ell)) \in V := \widehat{V} \times Q$ satisfying (6). Therefore, for every $\ell \in S$, the map $\ell \to (\bar{u}(\ell), p(\ell))$ is well-defined and single-valued. The following lemma gives some additional information on the parameter-to-solution map:

Lemma 1 *For any $\ell \in S$, the following estimates hold:*

$$\|\bar{u}(\ell)\|_{\widehat{V}} \leq \frac{\|m\|_{\widehat{V}^*}}{\kappa_1}, \tag{7a}$$

$$\|p(\ell)\|_Q \leq \frac{\|m\|_{\widehat{V}^*}}{\sqrt{\kappa_1 \varsigma_1}}, \tag{7b}$$

$$\|u(\ell)\|_V \leq \frac{\|m\|_{\widehat{V}^*}}{\sqrt{\kappa_1 \min(\kappa_1, \varsigma_1)}}. \tag{7c}$$

Proof Taking $v = (\bar{u}, p)$ in (6) and combining the resulting two equations, we obtain

$$\kappa_1 \|\bar{u}\|_{\widehat{V}^2}^2 + \varsigma_1 \|p\|_Q^2 \leq a(\ell, \bar{u}, \bar{u}) + c(p, p) \leq \|m\|_{\widehat{V}^*} \|\bar{u}\|_{\widehat{V}},$$

implying (7a), which leads to (7b), and finally by these bounds and the inequality

$$\min(\kappa_1, \varsigma_1)\|u\|_V^2 \leq \|m\|_{\widehat{V}^*} \|\bar{u}\|_{\widehat{V}} \leq \frac{\|m\|_{\widehat{V}^*}^2}{\kappa_1},$$

the estimate (7c) follows. The proof is complete. □

We now show that the parameter-to-solution map is Lipschitz continuous:

Lemma 2 *For $\ell_1, \ell_2 \in S$, let $u(\ell_1)$ and $u(\ell_2)$ be the unique solutions of the corresponding mixed variational problem. Then the following estimates hold:*

$$\|\bar{u}(\ell_1) - \bar{u}(\ell_2)\|_{\widehat{V}} \le \frac{\kappa_2}{\kappa_1^2}\|m\|_{\widehat{V}^*}\|\ell_1 - \ell_2\|_B, \tag{8a}$$

$$\|p(\ell_1) - p(\ell_2)\|_Q \le \frac{\kappa_2}{\kappa_1\sqrt{\varsigma_1\kappa_1}}\|m\|_{\widehat{V}^*}\|\ell_1 - \ell_2\|_B, \tag{8b}$$

$$\|u(\ell_1) - u(\ell_2)\|_V \le \frac{\kappa_2}{\kappa_1\sqrt{\kappa_1 \min(\kappa_1, \varsigma_1)}}\|m\|_{\widehat{V}^*}\|\ell_1 - \ell_2\|_B. \tag{8c}$$

Proof By the definitions of $u(\ell_1)$ and $u(\ell_2)$, we have

$$a(\ell_1, \bar{u}(\ell_1), \bar{v}) + b(\bar{v}, p(\ell_1)) = m(\bar{v}), \quad \text{for every } \bar{v} \in \widehat{V},$$
$$b(\bar{u}(\ell_1), q) - c(p(\ell_1), q) = 0, \quad \text{for every } q \in Q,$$

and

$$a(\ell_2, \bar{u}(\ell_2), \bar{v}) + b(\bar{v}, p(\ell_2)) = m(\bar{v}), \quad \text{for every } \bar{v} \in \widehat{V},$$
$$b(\bar{u}(\ell_2), q) - c(p(\ell_2), q) = 0, \quad \text{for every } q \in Q.$$

We set $v = u(\ell_1) - u(\ell_2)$ in the above equations, and rearrange them to get

$$\begin{aligned} a(\ell_1, \bar{u}(\ell_1) - \bar{u}(\ell_2), \bar{u}(\ell_1) - \bar{u}(\ell_2)) + b(\bar{u}(\ell_1) - \bar{u}(\ell_2), p(\ell_1) - p(\ell_2)) \\ = a(\ell_2 - \ell_1, \bar{u}(\ell_2), \bar{u}(\ell_1) - \bar{u}(\ell_2)) \\ b(\bar{u}(\ell_1) - \bar{u}(\ell_2), p(\ell_1) - p(\ell_2)) = c(p(\ell_1) - p(\ell_2), p(\ell_1) - p(\ell_2)), \end{aligned}$$

which implies

$$\begin{aligned} \kappa_1\|\bar{u}(\ell_1) - \bar{u}(\ell_2)\|_{\widehat{V}}^2 + \varsigma_1\|p(\ell_1) - p(\ell_2)\|_Q^2 &\le a(\ell_1, \bar{u}(\ell_1) - \bar{u}(\ell_2), \bar{u}(\ell_1) - \bar{u}(\ell_2)) \\ &+ c(p(\ell_1) - p(\ell_2), p(\ell_1) - p(\ell_2)) \\ &= a(\ell_2 - \ell_1, \bar{u}(\ell_2), \bar{u}(\ell_1) - \bar{u}(\ell_2)) \\ &\le \kappa_2\|\ell_1 - \ell_2\|_B\|\bar{u}(\ell_2)\|_{\widehat{V}}\|\bar{u}(\ell_1) - \bar{u}(\ell_2))\|_{\widehat{V}} \\ &\le \frac{\kappa_2}{\kappa_1}\|m\|_{\widehat{V}^*}\|\ell_1 - \ell_2\|_B\|\bar{u}(\ell_1) - \bar{u}(\ell_2))\|_{\widehat{V}}, \end{aligned}$$

proving (8a). This estimate further gives that

$$\varsigma_1\|p(\ell_1) - p(\ell_2)\|_Q^2 \le \frac{\kappa_2^2}{\kappa_1^3}\|m\|_{\widehat{V}^*}^2\|\ell_1 - \ell_2\|_B^2,$$

and hence establishing (8b). Finally, from

$$\min(\kappa_1, \varsigma_1)\|u(\ell_1) - u(\ell_2)\|_V^2 \le \frac{\kappa_2^2}{\kappa_1^3}\|m\|_{\widehat{V}^*}^2\|\ell_1 - \ell_2\|_B^2,$$

we derive (8c). The proof is complete. □

We now investigate the smoothness of the parameter-to-solution map:

Theorem 1 *For each $\ell \in S$, $u = u(\ell) = (\bar{u}(\ell), p(\ell))$ is infinitely differentiable at ℓ. The first derivative $\delta u = (\delta\bar{u}, \delta p) = (D\bar{u}(\ell)\delta\ell, Dp(\ell)\delta\ell)$ is the unique solution of the mixed variational problem:*

$$a(\ell, \delta\bar{u}, \bar{v}) + b(\bar{v}, \delta p) = -a(\delta\ell, \bar{u}, \bar{v}),\ \forall\, \bar{v} \in \widehat{V} \tag{9a}$$

$$b(\delta\bar{u}, q) - c(\delta p, q) = 0,\ \forall\, q \in Q. \tag{9b}$$

The second-order derivative

$$\delta^2 u = (\delta^2\bar{u}, \delta^2 p) = (D^2\bar{u}(\ell)(\delta\ell_1, \delta\ell_2), D^2 p(\ell)(\delta\ell_1, \delta\ell_2))$$

is the unique solution of the mixed variational problem:

$$a(\ell, \delta^2\bar{u}, \bar{v}) + b(\bar{v}, \delta^2 p) = -a(\delta\ell_2, D\bar{u}(\ell)\delta\ell_1, \bar{v}) - a(\delta\ell_1, D\bar{u}(\ell)\delta\ell_2, \bar{v}),\ \forall\, \bar{v} \in \widehat{V} \tag{10a}$$

$$b(\delta^2\bar{u}, q) - c(\delta^2 p, q) = 0,\ \forall\, q \in Q. \tag{10b}$$

Furthermore, the following estimates hold:

$$\|D\bar{u}(\ell)\| \le \frac{\kappa_2}{\kappa_1^2}\|m\|_{\widehat{V}^*}, \tag{11a}$$

$$\|Dp(\ell)\| \le \frac{\kappa_2}{\kappa_1\sqrt{\kappa_1\varsigma_1}}\|m\|_{\widehat{V}^*}, \tag{11b}$$

$$\|Du(\ell)\| \le \frac{\kappa_2}{\kappa_1\sqrt{\kappa_1\ \min(\kappa_1, \varsigma_1)}}\|m\|_{\widehat{V}^*}, \tag{11c}$$

$$\|D^2\bar{u}(\ell)\| \le \frac{2\kappa_2^2}{\kappa_1^3}\|m\|_{\widehat{V}^*}, \tag{11d}$$

$$\|D^2 p(\ell)\| \le \frac{2\kappa_2^2}{\kappa_1^2\sqrt{\kappa_1\varsigma_1}}\|m\|_{\widehat{V}^*}, \tag{11e}$$

$$\|D^2 u(\ell)\| \le \frac{2\kappa_2^2}{\kappa_1^2\sqrt{\kappa_1\ \min(\kappa_1, \varsigma_1)}}\|m\|_{\widehat{V}^*}, \tag{11f}$$

$$\|\delta^2\bar{u}\|_{\widehat{V}} \le \frac{2\kappa_2}{\kappa_1}\|\delta\ell\|_B\|D\bar{u}(\ell)\delta\ell\|_{\widehat{V}}, \tag{11g}$$

$$\|\delta^2 p\|_Q \le \frac{2\kappa_2}{\sqrt{\varsigma_1\kappa_1}}\|D\bar{u}(\ell)\delta\ell\|_{\widehat{V}}\|\delta\ell\|_B, \tag{11h}$$

$$\|\delta^2 u\|_V \le \frac{2\kappa_2}{\sqrt{\min(\kappa_1, \varsigma_1)\kappa_1}}\|\delta\ell\|_B\|D\bar{u}(\ell)\delta\ell\|_{\widehat{V}}, \tag{11i}$$

Proof The derivative formulae have been given in [23]. We now proceed to verify (11). We set $v = (\delta\bar{u}, \delta p)$ in (9) and combine the resulting inequalities to obtain

$$\begin{aligned}\kappa_1\|\delta\bar{u}\|^2_{\widehat{V}} + \varsigma_1\|\delta p\|^2_Q &\leq a(\ell,\delta\bar{u},\delta\bar{u}) + c(\delta p,\delta p)\\ &= -a(\delta\ell,\bar{u},\delta\bar{u}) \leq \kappa_2\|\delta\ell\|_B\|\delta\bar{u}\|_{\widehat{V}}\|\bar{u}\|_{\widehat{V}}\\ &\leq \frac{\kappa_2}{\kappa_1}\|m\|_{\widehat{V}^*}\|\delta\ell\|_B\|\delta\bar{u}\|_{\widehat{V}},\end{aligned}$$

where we used an estimate from Lemma 1. Therefore, we have

$$\|\delta\bar{u}\|_{\widehat{V}} \leq \frac{\kappa_2}{\kappa_1^2}\|m\|_{\widehat{V}^*}\|\delta\ell\|_B,$$

which implies that

$$\|\delta p\|_Q \leq \frac{\kappa_2}{\kappa_1\sqrt{\kappa_1\varsigma_1}}\|m\|_{\widehat{V}^*}\|\delta\ell\|_B,$$

and also

$$\|\delta u\|_V \leq \frac{\kappa_2}{\kappa_1\sqrt{\kappa_1\ \min(\kappa_1,\varsigma_1)}}\|m\|_{\widehat{V}^*}\|\delta\ell\|_B,$$

and the first three bounds follow from the above three inequalities.

For estimates for the second derivative, we take $\delta\ell_1 = \delta\ell_2 = \delta\ell$ and set $v = \delta^2 u$ in (10) to get

$$\begin{aligned}a(\ell,\delta^2\bar{u},\delta^2\bar{u}) + b(\delta^2\bar{u},\delta^2 p) &= -a(\delta\ell, D\bar{u}(\ell)\delta\ell,\delta^2\bar{u}) - a(\delta\ell, D\bar{u}(\ell)\delta\ell,\delta^2\bar{u}),\\ b(\delta^2\bar{u},\delta^2 p) - c(\delta^2 p,\delta^2 p) &= 0,\end{aligned}$$

and as before, after combining the above set of equations, we obtain

$$\begin{aligned}\kappa_1\|\delta^2\bar{u}\|^2_{\widehat{V}} + \varsigma_1\|\delta^2 p\|^2_Q &\leq a(\ell,\delta^2\bar{u},\delta^2\bar{u}) + c(\delta^2 p,\delta^2 p)\\ &\leq 2\kappa_2\|\delta\ell\|_B\|D\bar{u}(\ell)\delta\ell\|_{\widehat{V}}\|\delta^2\bar{u}\|_{\widehat{V}},\end{aligned}\tag{12}$$

implying

$$\|\delta^2\bar{u}\|_{\widehat{V}} \leq \frac{2\kappa_2^2}{\kappa_1^3}\|m\|_{\widehat{V}^*}\|\delta\ell\|^2_B,$$

and consequently

$$\|\delta^2 p\|_Q \leq \frac{2\kappa_2^2}{\kappa_1^2\sqrt{\kappa_1\varsigma_1}}\|m\|_{\widehat{V}^*}\|\delta\ell\|^2_B,$$

and also

$$\|\delta^2 u\|_{\widehat{V}} \leq \frac{2\kappa_2^2}{\kappa_1^2\sqrt{k_1}\min(\kappa_1, \varsigma_1)}\|m\|_{\widehat{V}^*}\|\delta\ell\|_B^2,$$

leading to the next three bounds.

For the remaining bounds, we note that (12) yields

$$\|\delta^2\bar{u}\|_{\widehat{V}} \leq \frac{2\kappa_2}{\kappa_1}\|\delta\ell\|_B\|D\bar{u}(\ell)\delta\ell\|_{\widehat{V}},$$

implying

$$\|\delta^2 p\|_Q^2 \leq \frac{2\kappa_2}{\varsigma_1}\|\delta\ell\|_B\|D\bar{u}(\ell)\delta\ell\|_{\widehat{V}}\|\delta^2\bar{u}\|_{\widehat{V}} \leq \frac{2^2\kappa_2^2}{\varsigma_1\kappa_1}\|D\bar{u}(\ell)\delta\ell\|_{\widehat{V}}^2\|\delta\ell\|_B^2,$$

and hence

$$\|\delta^2 p\|_Q \leq \frac{2\kappa_2}{\sqrt{\varsigma_1\kappa_1}}\|D\bar{u}(\ell)\delta\ell\|_{\widehat{V}}\|\delta\ell\|_B.$$

Moreover, from the inequality

$$\min(\kappa_1, \varsigma_1)\|\delta^2 u\|_V^2 \leq 2\kappa_2\|\delta\ell\|_B\|D\bar{u}(\ell)\delta\ell\|_{\widehat{V}}\|\delta^2\bar{u}\|_{\widehat{V}},$$

we have

$$\|\delta^2 u\|_V^2 \leq \frac{2^2\kappa_2^2}{\min(\kappa_1, \varsigma_1)\kappa_1}\|\delta\ell\|_B^2\|D\bar{u}(\ell)\delta\ell\|_{\widehat{V}}^2$$

or equivalently

$$\|\delta^2 u\|_V \leq \frac{2\kappa_2}{\sqrt{\min(\kappa_1, \varsigma_1)\kappa_1}}\|\delta\ell\|_B\|D\bar{u}(\ell)\delta\ell\|_{\widehat{V}}, \tag{13}$$

which establishes the last inequality. This completes the proof. □

3 Optimization Formulations

As mentioned above, our focus is on the inverse problem of estimating the coefficient ℓ in (6) so that the unique solution $u(\ell) = (\bar{u}(\ell), p(\ell))$ of (6) is closest in some norm to the given measurement $z = (\bar{z}, \hat{z})$ of $u(\ell)$. A common approach is to pose this as an optimization problem minimizing the OLS functional

$$J_O(\ell) := \frac{1}{2}\|u(\ell) - z\|_Z^2 = \frac{1}{2}\|\bar{u}(\ell) - \bar{z}\|_{\bar{Z}}^2 + \frac{1}{2}\|p(\ell) - \hat{z}\|_{\widehat{Z}}^2,$$

with $(\bar{z}, \hat{z}) \in Z := \bar{Z} \times \widehat{Z}$, where Z is a suitable observation space.

The OLS formulation for nonlinear inverse problems of parameter identification is typically nonconvex, and hence it is limited to characterizing local minima. Of course, a natural strategy to circumvent the difficulties associated to the nonconvexity of the OLS functional is to introduce an analog of (4).

In [23], the following modified output least-squares (MOLS) was introduced

$$J_M(\ell) := \frac{1}{2}a(\ell, \bar{u}(\ell)-\bar{z}, \bar{u}(\ell)-\bar{z})+b(\bar{u}(\ell)-\bar{z}, p(\ell)-\hat{z})-\frac{1}{2}c(p(\ell)-\hat{z}, p(\ell)-\hat{z}),$$

where $(\bar{z}, \hat{z}) \in V$ is the data.

For sake a completeness, we recall the following feature of MOLS (see [23]):

Theorem 2 *The modified output least-squares functional defined above is convex on the set A.*

Proof Using the derivative characterization (9), we have

$$\begin{aligned} DJ_M(\ell)(\hat{\ell}) &= \frac{1}{2}a(\hat{\ell}, \bar{u}(\ell) - \bar{z}, \bar{u}(\ell) - \bar{z}) - a(\hat{\ell}, \bar{u}(\ell), \bar{u}(\ell) - \bar{z}) \\ &= -\frac{1}{2}a(\hat{\ell}, \bar{u}(\ell) + \bar{z}, \bar{u}(\ell) - \bar{z}), \end{aligned} \tag{14}$$

which, using (9) again, yields

$$D^2 J_M(\ell)(\hat{\ell}, \hat{\ell}) = a(\ell, \delta\bar{u}, \delta\bar{u}) + c(\delta p, \delta p) \geq \kappa_1 \|\delta\bar{u}\|^2 + \varsigma_1 \|\delta p\|^2, \tag{15}$$

and the convexity follows. □

In [13], the following energy output least-squares was proposed

$$J_E(\ell) := \frac{1}{2}a(\ell, \bar{u}(\ell) - \bar{z}, \bar{u}(\ell) - \bar{z}) + \frac{1}{2}c(p(\ell) - \hat{z}, p(\ell) - \hat{z}),$$

where $(\bar{z}, \hat{z}) \in V$ is the data.

The idea of minimizing the energy of the underlying variational problem was fundamental to the convexity of (4). However, since the mixed variational formulation involves a coupled system of equations, the two ways of combining them result in MOLS and EOLS with different features; MOLS preserves convexity but loses positivity whereas EOLS retains positivity but is non-convex in general.

The inverse problem of identifying parameters is ill-posed, and some regularization is essential. Two frameworks have been developed for identification in (1), one for the identification of smooth coefficients and the second for the identification of discontinuous coefficients. In the following, we formulate two assumptions to recapture these two frameworks:

Assumption A The Banach space B is continuously embedded in a Banach space L. There is another Banach space $\widehat{B}$ that is compactly embedded in L. The set A is a subset of $B \cap \widehat{B}$, closed and bounded in B and also closed in L. For any bounded sequences $\{\bar{u}_k\} \subset \widehat{V}$, and $\{\ell_k\} \subset B$ with $\ell_k \to \ell$ in L, for any fixed $v \in \widehat{V}$, we have

$$a(\ell_k - \ell, \bar{u}_k, v) \to 0. \tag{16}$$

Moreover, $R : \widehat{B} \to \mathbb{R}$ is a positive, convex, $\|\cdot\|_L$-lower-semicontinuous map with

$$R(\ell) \geq \tau_1 \|\ell\|_{\widehat{B}} - \tau_2, \quad \text{for every } \ell \in A, \quad \text{for some } \tau_1 > 0,\ \tau_2 > 0. \tag{17}$$

Assumption B A Hilbert space $\widehat{H}$ is compactly embedded into the space B, $A \subset \widehat{H}$ is nonempty, closed, and convex, $R : \widehat{H} \to \mathbb{R}$ is convex, lower-semicontinuous, and there is $\alpha_1 > 0$ such that

$$R(\ell) \geq \alpha_1 \|\ell\|^2_{\widehat{H}}, \quad \text{for every } \ell \in A. \tag{18}$$

Given a regularization parameter $\kappa > 0$, a regularization map given through either Assumption A or B, consider the following optimization problems involving the OLS, MOLS, and EOLS functionals.

Find $\ell \in A$ by solving the regularized OLS based optimization problem:

$$J^\kappa_O(\ell) := \frac{1}{2}\|u(\ell) - z\|^2_Z + \kappa R(\ell). \tag{19}$$

Find $\ell \in A$ by solving the regularized MOLS based optimization problem:

$$\begin{aligned} J^\kappa_M(\ell) := &\frac{1}{2}a(\ell, \bar{u}(\ell) - \bar{z}, \bar{u}(\ell) - \bar{z}) + b(\bar{u}(\ell) - \bar{z}, p(\ell) - \hat{z}) \\ &- \frac{1}{2}c(p(\ell) - \hat{z}, p(\ell) - \hat{z}) + \kappa R(\ell). \end{aligned} \tag{20}$$

Find $\ell \in A$ by solving the regularized EOLS based optimization problem:

$$J^\kappa_E(\ell) := \frac{1}{2}a(\ell, \bar{u}(\ell) - \bar{z}, \bar{u}(\ell) - \bar{z}) + \frac{1}{2}c(p(\ell) - \hat{z}, p(\ell) - \hat{z}) + \kappa R(\ell). \tag{21}$$

We have the following existence result for the above problems:

Theorem 3 *Under Assumption A, optimization problems (19)–(21) have nonempty solution sets, where every minimizer of (20) is a global minimizer. Moreover, $\bar{\ell} \in A$ is a minimizer of (20), if and only if, it solves the variational inequality:*

$$-\frac{1}{2}a(\hat{\ell} - \bar{\ell}, \bar{u}(\bar{\ell}) + \bar{z}, \bar{u}(\bar{\ell}) - \bar{z}) \geq \kappa[R(\bar{\ell}) - R(\hat{\ell})], \quad \textit{for every } \hat{\ell} \in A. \tag{22}$$

Moreover, $\bar{\ell} \in A$ is a minimizer of (19), then it solves the variational inequality

$$\left\langle Du(\bar{\ell})(\hat{\ell} - \bar{\ell}), u(\bar{\ell}) - z \right\rangle_Z \geq \kappa[R(\bar{\ell}) - R(\hat{\ell})], \quad \textit{for every } \hat{\ell} \in A. \tag{23}$$

Furthermore, if $\bar{\ell} \in A$ is a minimizer of (21)*, then it solves the variational inequality*

$$-\frac{1}{2}a(\hat{\ell}-\bar{\ell}, \bar{u}(\ell)+\bar{z}, \bar{u}(\ell)-\bar{z}) + b(\delta u, p(\ell)-\hat{z})$$
$$- b(\bar{u}(\ell)-\bar{z}, \delta p) \geq \kappa[R(\bar{\ell}) - R(\hat{\ell})], \quad \textit{for every } \hat{\ell} \in A. \tag{24}$$

Proof The solvability of (20) can be found in [23] and the solvability of (21) can be found in [13]. Analogous arguments can be used to prove the solvability of (19). Since the MOLS objective and the regularizer R are both convex, variational inequality (22) is a necessary and sufficient optimality conditions which follows using (14). Analogously (23) is a necessary optimality condition of (19). Finally, (24) is a necessary optimality condition of (21) and the specific form follows from (see [13])

$$DJ_E(\ell)(\delta\ell) = -\frac{1}{2}a(\delta\ell, \bar{u}(\ell)+\bar{z}, \bar{u}(\ell)-\bar{z}) + b(\delta u, p(\ell)-\hat{z}) - b(\bar{u}(\ell)-\bar{z}, \delta p),$$

and the proof is complete. □

Let us briefly touch upon the rationale behind the above assumptions. Assumption A provides a theoretical framework to recover discontinuous coefficients and is motivated by the use of total variation regularization in inverse problems. Recall that the total variation of $f \in L^1(\Omega)$ is given by

$$\mathrm{TV}(f) = \sup\left\{-\int_\Omega f\nabla\cdot g \;:\; g=(g_1,g_2)\in C_0^1(\Omega;\mathbb{R}^2),\ |g(x)|\leq 1\ \forall x\in\Omega\right\},$$

where $|\cdot|$ represents the Euclidean norm of a vector. Clearly, if $f \in W^{1,1}(\Omega)$, then $\mathrm{TV}(f) = \int_\Omega |\nabla f|$.

If $f \in L^1(\Omega)$ satisfies $\mathrm{TV}(f) < \infty$, then f is said to have bounded variation, and $\mathrm{BV}(\Omega)$ is defined by $\mathrm{BV}(\Omega) = \left\{f \in L^1(\Omega) : \mathrm{TV}(f) < \infty\right\}$. The norm on $\mathrm{BV}(\Omega)$ is $\|f\|_{\mathrm{BV}(\Omega)} = \|f\|_{L^1(\Omega)} + \mathrm{TV}(f)$. The functional $\mathrm{TV}(\cdot)$ is a seminorm on $\mathrm{BV}(\Omega)$ and is often called the BV-seminorm.

We set $B = L^\infty(\Omega)$, $L = L^1(\Omega)$, $\widehat{B} = \mathrm{BV}(\Omega)$, and $R(a) = TV(a)$, and define

$$A_1 := \{a \in L^\infty |\ 0 < c_1 \leq a(x) \leq c_2,\ \text{almost everywhere}\},$$
$$A_2 := \{a \in L^\infty |\ 0 < c_1 \leq a(x) \leq c_2,\ \text{almost everywhere},\ \mathrm{TV}(a) \leq c_3 < \infty\},$$

where c_1, c_2, and c_3 are constants. It is known that $L^\infty(\Omega)$ is continuously embedded in $L^1(\Omega)$, $\mathrm{BV}(\Omega)$ is compactly embedded in $L^1(\Omega)$, and $TV(\cdot)$ is convex and lower-semicontinuous in $L^1(\Omega)$-norm (see [1, 16]). If we study the regularized optimization problems on the set A_1, then the regularizer ensures that the minimizing sequence remains bounded in $\widehat{B}$. However, if the set is A_2, then the regularizer does not play any such role because A_2 is already bounded in $\widehat{B}$. Assumption B, however, is suitable for the identification of smooth parameters. It is clearly satisfied on a suitable domain Ω by taking $B = L^{\infty}(\Omega)$ and $\widehat{H} = H_2(\Omega)$ and $R(\ell) = \|\ell\|^2_{\widehat{H}}$.

4 Asymptotic Stability of the MOLS Approach

To study the asymptotic stability of the MOLS approach under data perturbation, assume $z = (\bar{z}, \hat{z}) \in V$ is the exact data and $z_\delta = (\bar{z}_\delta, \hat{z}_\delta) \in V$ is the contaminated data such that $\|z_\delta - z\|_V \leq \delta$, where $\delta > 0$.

Consider the problem of finding $\ell_\delta \in A$ by minimizing

$$\begin{aligned} J_M^\delta(\ell) &:= \frac{1}{2}a(\ell, \bar{u}(\ell) - \bar{z}_\delta, \bar{u}(\ell) - \bar{z}_\delta) + b(\bar{u}(\ell) - \bar{z}_\delta, p(\ell) - \hat{z}_\delta) \\ &\quad - \frac{1}{2}c(p(\ell) - \hat{z}_\delta, p(\ell) - \hat{z}_\delta) + \kappa R(\ell). \end{aligned} \tag{25}$$

where $u(\ell)$ solves (6), $\kappa > 0$ is a fixed parameter, and R is given in (17).

The following result shows that for a fixed regularized parameter, the regularized problems with contaminated data converge to the regularized problem with the exact data if the noise decays suitably:

Theorem 4 *For any $\delta > 0$, (25) has a solution ℓ_δ. Furthermore, there exists a subsequence $\{\ell_\delta\}$ converging in $\|\cdot\|_L$ as $\delta \to 0$ to a solution $\tilde{\ell}$ of (20).*

Proof The existence of a solution ℓ_δ of (25) follows from Theorem 3. We shall show that $\{\ell_\delta\} \subset A$ is bounded in $\widehat{B}$. For any $\hat{\ell} \in A$, we have

$$\begin{aligned} J_M^\delta(\ell_\delta) &= \frac{1}{2}a(\ell_\delta, \bar{u}(\ell_\delta) - \bar{z}_\delta, \bar{u}(\ell_\delta) - \bar{z}_\delta) + b(\bar{u}(\ell_\delta) - \bar{z}_\delta, p(\ell_\delta) - \hat{z}_\delta) \\ &\quad - \frac{1}{2}c(p(\ell_\delta) - \hat{z}_\delta, p(\ell_\delta) - \hat{z}_\delta) + \kappa R(\ell_\delta) \\ &\leq \frac{1}{2}a(\hat{\ell}, \bar{u}(\hat{\ell}) - \bar{z}_\delta, \bar{u}(\hat{\ell}) - \bar{z}_\delta) + b(\bar{u}(\hat{\ell}) - \bar{z}_\delta, p(\hat{\ell}) - \hat{z}_\delta) \\ &\quad - \frac{1}{2}c(p(\hat{\ell}) - \hat{z}_\delta, p(\hat{\ell}) - \hat{z}_\delta) + \kappa R(\hat{\ell}) \\ &\leq \frac{1}{2}a(\hat{\ell}, \bar{u}(\hat{\ell}) - \bar{z}, \bar{u}(\hat{\ell}) - \bar{z}) + a(\hat{\ell}, \bar{u}(\hat{\ell}) - \bar{z}, \bar{z} - \bar{z}_\delta) + \frac{1}{2}a(\hat{\ell}, \bar{z} - \bar{z}_\delta, \bar{z} - \bar{z}_\delta) \\ &\quad + b(\bar{u}(\hat{\ell}) - \bar{z}, p(\hat{\ell}) - \hat{z}) + b(\bar{u}(\hat{\ell}) - \bar{z}, \hat{z} - \hat{z}_\delta) + b(\bar{z} - \bar{z}_\delta, p(\hat{\ell}) - \hat{z}) \\ &\quad + b(\bar{z} - \bar{z}_\delta, \hat{z} - \hat{z}_\delta) - \frac{1}{2}c(p(\hat{\ell}) - \hat{z}, p(\hat{\ell}) - \hat{z}) - c(p(\hat{\ell}) - \hat{z}, \hat{z} - \hat{z}_\delta) \\ &\quad - \frac{1}{2}c(z - z_\delta, z - z_\delta) + \kappa R(\hat{\ell}) \leq M, \end{aligned}$$

where M is a constant. Here we used (5) and the fact that z_δ is bounded. Furthermore, by using the fact that A is bounded in B and $u(\ell_\delta)$ is bounded independent of δ, we note that the term

$$\frac{1}{2}a(\ell_\delta, \bar{u}(\ell_\delta)-\bar{z}_\delta, \bar{u}(\ell_\delta)-\bar{z}_\delta)+b(\bar{u}(\ell_\delta)-\bar{z}_\delta, p(\ell_\delta)-\hat{z}_\delta)-\frac{1}{2}c(p(\ell_\delta)-\hat{z}_\delta, p(\ell_\delta)-\hat{z}_\delta)$$

is bounded below by a constant. Therefore, by the definition of $R(\cdot)$, we deduce that $\{\ell_\delta\}$ is bounded in $\|\cdot\|_{\widehat{B}}$ and therefore there is a subsequence of $\{\ell_\delta\}$ converging in $\|\cdot\|_L$ to some $\tilde{\ell} \in A$. We set $u_\delta = (\bar{u}(\ell_\delta), p(\ell_\delta))$. It can be shown that $u_\delta \to \tilde{u} = (\bar{u}(\tilde{\ell}), p(\tilde{\ell}))$ as $\delta \to 0$.

For an arbitrary $\ell \in A$, we have

$$\begin{aligned}
J_M^\kappa(\tilde{\ell}) &= \frac{1}{2}a(\tilde{\ell}, \bar{u}(\tilde{\ell}) - \bar{z}, \bar{u}(\tilde{\ell}) - \bar{z}) + b(\bar{u}(\tilde{\ell}) - \bar{z}, p(\tilde{\ell}) - \hat{z}) \\
&\quad - \frac{1}{2}c(p(\tilde{\ell}) - \hat{z}, p(\tilde{\ell}) - \hat{z}) + \kappa R(\tilde{\ell}) \\
&\le \liminf_{\delta\to 0} \Big\{ \frac{1}{2}a(\ell_\delta, \bar{u}(\ell_\delta) - \bar{z}, \bar{u}(\ell_\delta) - \bar{z}) + b(\bar{u}(\ell_\delta) - \bar{z}, p(\ell_\delta) - \hat{z}) \\
&\quad -\frac{1}{2}c(p(\ell_\delta) - \hat{z}, p(\ell_\delta) - \hat{z}) + \kappa R(\ell_\delta)\Big\} \\
&\le \limsup_{\delta\to 0} \Big\{ \frac{1}{2}a(\ell_\delta, \bar{u}(\ell_\delta) - \bar{z}_\delta, \bar{u}(\ell_\delta) - \bar{z}_\delta) + b(\bar{u}(\ell_\delta) - \bar{z}_\delta, p(\ell_\delta) - \hat{z}_\delta) \\
&\quad -\frac{1}{2}c(p(\ell_\delta) - \hat{z}_\delta, p(\ell_\delta) - \hat{z}_\delta) + \kappa R(\ell_\delta)\Big\} \\
&\le \limsup_{\delta\to 0} \Big\{ \frac{1}{2}a(\ell, \bar{u}(\ell) - \bar{z}_\delta, \bar{u}(\ell) - \bar{z}_\delta) + b(\bar{u}(\ell) - \bar{z}_\delta, p(\ell) - \hat{z}_\delta) \\
&\quad -\frac{1}{2}c(p(\ell) - \hat{z}_\delta, p(\ell) - \hat{z}_\delta) + \kappa R(\ell)\Big\} \\
&\le \frac{1}{2}a(\ell, \bar{u}(\ell) - \bar{z}, \bar{u}(\ell) - \bar{z}) + b(\bar{u}(\ell) - \bar{z}, p(\ell) - \hat{z}) \\
&\quad - \frac{1}{2}c(p(\ell) - \hat{z}, p(\ell) - \hat{z}) + \kappa R(\ell),
\end{aligned}$$

where we repeatedly used the fact that $\delta \to 0$. Since ℓ is arbitrarily, it follows that $\tilde{\ell}$ is a minimizer. □

It would be of interest to explore an analogue of the above result when $\kappa \to 0$. Unfortunately the above proof does not carry over to this case, and consequently, we develop a new proof relying on the equivalent variational inequality formulation.

Let $z = (\bar{z}, \hat{z}) \in V$ be the exact data and let $z_n := (\bar{z}_n, \hat{z}_n) \in V$ be the noisy data such that $\|z_n - z\| \leq \delta_n$, where $\delta_n > 0$. For $n \in \mathbb{N}$, consider the problem of finding $\ell_n \in A$ by solving

$$\min_{\ell \in A} J_M^{\delta_n}(\ell) := \frac{1}{2} a(\ell, \bar{u}(\ell) - \bar{z}_{\delta_n}, \bar{u}(\ell) - \bar{z}_{\delta_n}) + b(\bar{u}(\ell) - \bar{z}_{\delta_n}, p(\ell) - \hat{z}_{\delta_n}) \\ - \frac{1}{2} c(p(\ell) - \hat{z}_{\delta_n}, p(\ell) - \hat{z}_{\delta_n}) + \kappa_n R(\ell), \tag{26}$$

where $u(\ell) = (\bar{u}(\ell), p(\ell))$ solves (6), $\kappa_n > 0$, and the regularizer R is defined in Assumption A.

The following result proves the convergence of the regularization solutions:

Theorem 5 *Assume that the solution set $\mathscr{S}$ of the following optimization problem is nonempty:*

$$\min_{\ell \in A} J_M(\ell) := \frac{1}{2} a(\ell, \bar{u}(\ell) - \bar{z}, \bar{u}(\ell) - \bar{z}) + b(\bar{u}(\ell) - \bar{z}, p(\ell) - \hat{z}) - \frac{1}{2} c(p(\ell) - \hat{z}, p(\ell) - \hat{z}). \tag{27}$$

Assume that $\{\kappa_n, \delta_n, \delta_n \kappa_n^{-1}\} \to 0$ as $n \to \infty$. Then, for every $n \in \mathbb{N}$, problem (26) has a minimizer ℓ_n. Moreover, there is a subsequence $\{\ell_n\}$, converging in $\| \cdot \|_L$ as $n \to \infty$, to a minimizer $\tilde{\ell}$ of (27).

Proof Since $J_M^{\delta_n}$ is convex, a necessary and sufficient optimality condition for a minimizer $\ell_n \in A$ of (26) is the following variational inequality

$$\langle D J_M^{\delta_n}(\ell_n), \ell - \ell_n \rangle \geq \kappa_n \left[R(\ell_n) - R(\ell) \right], \quad \text{for every } \ell \in A, \tag{28}$$

where

$$\langle D J_M^{\delta_n}(\ell_n), \delta\ell \rangle = -\frac{1}{2} a(\delta\ell, \bar{u}_n + \bar{z}_n, \bar{u}_n - \bar{z}_n).$$

Also any $\bar{\ell} \in \mathscr{S}$ satisfies the following variational inequality:

$$\langle D J_M(\bar{\ell}), \ell - \bar{\ell} \rangle \geq 0, \quad \text{for every } \ell \in A. \tag{29}$$

Setting $\ell = \bar{\ell}$ in (28), $\ell = \ell_n$ in (29), and using the monotonicity

$$\langle D J_M(\bar{\ell}) - D J_M(\ell_n), \bar{\ell} - \ell_n \rangle \geq 0,$$

which holds due to the convexity of MOLS, we get

$$\langle D J_M^{\delta_n}(\ell_n) - D J_M(\ell_n), \bar{\ell} - \ell_n \rangle \geq \kappa_n \left[R(\ell_n) - R(\bar{\ell}) \right].$$

To obtain an upper bound on the left-hand side term of the above inequality, we note

$$
\begin{aligned}
&2\langle DJ_M^{\delta_n}(\ell_n) - DJ_M(\ell_n), \bar{\ell} - \ell_n\rangle \\
&\quad = a(\ell_n - \bar{\ell}, \bar{u}_n + \bar{z}_n, \bar{u}_n - \bar{z}_n) - a(\ell_n - \bar{\ell}, \bar{u}_n + \bar{z}, \bar{u}_n - \bar{z}) \\
&\quad = a(\ell_n - \bar{\ell}, \bar{u}_n + \bar{z}, \bar{u}_n - \bar{z}_n) + a(\ell_n - \bar{\ell}, \bar{z}_n - \bar{z}, \bar{u}_n - \bar{z}_n) \\
&\qquad + a(\ell_n - \bar{\ell}, \bar{u}_n + \bar{z}, \bar{z} - \bar{u}_n) \\
&\quad = a(\ell_n - \bar{\ell}, \bar{u}_n + \bar{z}, \bar{z} - \bar{z}_n) + a(\ell_n - \bar{\ell}, \bar{z}_n - \bar{z}, \bar{u}_n - \bar{z}_n) \\
&\quad = a(\ell_n - \bar{\ell}, \bar{u}_n + \bar{z}, \bar{z} - \bar{z}_n) + a(\ell_n - \bar{\ell}, \bar{z}_n - \bar{u}_n, \bar{z} - \bar{z}_n) \\
&\quad = a(\ell_n - \bar{\ell}, \bar{z}_n + \bar{z}, \bar{z} - \bar{z}_n) \\
&\quad = a(\ell_n - \bar{\ell}, \bar{z}_n - \bar{z}, \bar{z} - \bar{z}_n) + a(\ell_n - \bar{\ell}, 2\bar{z}, \bar{z} - \bar{z}_n) \\
&\quad \le \kappa_2 \|\bar{\ell} - \ell_n\|_B \left[2\delta_n \|\bar{z}\|_V + \delta_n^2\right] \\
&\quad \le c\delta_n \|\bar{\ell} - \ell_n\|_B,
\end{aligned}
$$

where c is constant including κ_2, $\|\bar{z}\|_V$, and a fixed upper bound on δ_n.

Therefore, $\kappa_n \left[R(\ell_n) - R(\bar{\ell})\right] \le c\delta_n \|\bar{\ell} - \ell_n\|_B$, and A is bounded in B, we obtain that there is a constant $\tilde{c}$ such that $R(a_n) \le \tilde{c}$, ensuring that the sequence remains bounded in $\widehat{B}$ and consequently has a subsequence $\{\ell_n\}$ which converges, in $\|\cdot\|_L$, to some $\tilde{\ell} \in A$. We set $\tilde{u} = u(\tilde{\ell})$. It can be shown that $u_n \to \tilde{u} = \tilde{u}(\tilde{\ell})$ as $n \to \infty$.

Let $\ell \in A$ be arbitrary. Then,

$$
\begin{aligned}
J_M^{\kappa}(\tilde{\ell}) &= \frac{1}{2} a(\tilde{\ell}, \bar{u}(\tilde{\ell}) - \bar{z}, \bar{u}(\tilde{\ell}) - \bar{z}) + b(\bar{u}(\tilde{\ell}) - \bar{z}, p(\tilde{\ell}) - \hat{z}) \\
&\quad - \frac{1}{2} c(p(\tilde{\ell}) - \hat{z}, p(\tilde{\ell}) - \hat{z}) \\
&\le \lim_{n\to\infty} \left\{ \frac{1}{2} a(\ell_n, \bar{u}_n - \bar{z}, \bar{u}_n - \bar{z}) + b(\bar{u}_n - \bar{z}, p_n - \hat{z}) \right. \\
&\quad \left. - \frac{1}{2} c(p_n - \hat{z}, p_n - \hat{z}) \right\} \\
&\le \lim_{n\to\infty} \left\{ \frac{1}{2} a(\ell_n, \bar{u}_n - \bar{z}_n, \bar{u}_n - \bar{z}_n) + b(\bar{u}_n - \bar{z}_n, p_n - \hat{z}_n) \right. \\
&\quad \left. - \frac{1}{2} c(p_n - \hat{z}_n, p_n - \hat{z}_n) + \kappa_n R(\ell_n) \right\} \\
&\le \lim_{n\to\infty} \left\{ \frac{1}{2} a(\ell, \bar{u}(\ell) - \bar{z}_n, \bar{u}(\ell) - \bar{z}_n) + b(\bar{u}(\ell) - \bar{z}_n, p(\ell) - \hat{z}_n) \right. \\
&\quad \left. - \frac{1}{2} c(p(\ell) - \hat{z}_n, p(\ell) - \hat{z}_n) + \kappa_n R(\ell) \right\}
\end{aligned}
$$

$$= \frac{1}{2}a(\ell, \bar{u}(\ell) - \bar{z}, \bar{u}(\ell) - \bar{z}) + b(\bar{u}(\ell) - \bar{z}, p(\ell) - \hat{z})$$
$$- \frac{1}{2}c(p(\ell) - \hat{z}, p(\ell) - \hat{z}),$$

by similar calculations as in Theorem 4. Hence, $\tilde{\ell}$ is a minimizer of (20). □

5 Local Stability Estimates

Inspired by the approach developed by Alt [2, 3], Chavent [7], Colonius and Kunisch [10, 11], and White [31, 32], we shall now investigate the stability of OLS, MOLS, and EOLS to the data perturbation by using Tikhonov regularization. Throughout this section, we assume that the regularization space H is a Hilbert space continuously embedded in the parameter space B, that is, there is an embedding constant $\hat{c}$ such that $\|w\|_B \leq \hat{c}\|w\|_H$, for every $w \in H$. Let Y be a Banach space, let $C \subset Y$ be a pointed, closed, and convex cone inducing a partial ordering in Y, and let C^+ be the positive dual of C. The cone C is related to the set $A \subset H$ of admissible parameters which is contained in the open set $S \subset H$, and will be used to pose admissibility using explicit constraints. We assume that $g : S \to Y$ is twice differentiable map such that $A := \{\ell \in H | \, g(\ell) \in -C\}$. We also assume that the outcome space V is continuously embedded in the observation space Z. For the applications that we have in mind, it is enough to assume that $\|w\|_Z \leq \|w\|_V$, for every $w \in V$. Throughout this section, $\kappa > 0$ is a fixed regularization parameter.

Given $\widetilde{J} : S \times Z \to \mathbb{R}_+$ and $g : S \to Y$, for $z \in Z$, consider problem (P_z):

$$\min \widetilde{J}(\ell, z) \quad \text{subject to} \quad A := \{\ell \in H | \, g(\ell) \in -C\}.$$

We assume that for every $z = (\bar{z}, \hat{z}) \in Z := \bar{Z} \times \widehat{Z}$, there is at least one solution ℓ^z of problem (P_z). Conditions on g are vital for the selection of the constraint space Y. We denote by $\ell^z \in H$ a solution of (P_z), and explore how such solutions depend on the data z. Throughout this section, by $(\ell_0, z_0) \in H \times Z$, we denote the reference point where ℓ_0 is a solution of (P_{z_0}) and z_0 is the exact data. For simplicity, we assume that the constraint map g is twice continuously differentiable.

The well-known Karush-Kuhn-Tucker (KKT) condition for (P_{z_0}) will be used:

Theorem 6 *Let $\ell_0 \in A$ be a regular point, that is, for every $\ell \in A$, the constraint qualification*

$$0 \in \operatorname{int}\left(g(\ell_0) + Dg(\ell_0)(\ell - \ell_0) + C\right), \tag{30}$$

holds. Then there exists $\mu_0 \in C^+$ such that $\mu_0(g(\ell_0)) = 0$, $g(\ell_0) \in -C$, and for every $\ell \in A$, we have

$$D\widetilde{J}(\ell_0)(\ell - \ell_0) = -\mu_0(Dg(\ell_0)(\ell - \ell_0)). \tag{31}$$

Commonly used constraints sets in the identification problems are pointwise constraints and/or norm constraints. The following result (see [10]) shows that every point in such sets is a regular point:

Lemma 3 *Let Q be a real Hilbert space, and let $K \subset Q$ be a closed convex cone with vertex at zero inducing an ordering on Q such that $K = \{q \in Q | q \geq 0\}$. For $k \in K$ and $\gamma \in \mathbb{R}^+$, we define $G = (G_1, G_2) : Q \to Q \times \mathbb{R}$ by $G(q) = (G_1(q), G_2(q)) = (k - q, \|q\|^2 - \gamma^2)$. Let $\gamma > \|k\|$. Then every point of the set $\widetilde{Q} = \{q \in Q |\ G(q) \leq 0\}$ is a regular point.*

5.1 Stability of the Output Least-Squares Approach

We now return to the regularized OLS based optimization problem (P_z^1):

$$\begin{aligned} \min J_O(\ell, z) &:= \frac{1}{2} \|u(\ell) - z\|_Z^2 + \frac{\kappa}{2} \|\ell\|_H^2 \\ &= \frac{1}{2} \|\bar{u}(\ell) - \bar{z}\|_{\bar{Z}}^2 + \frac{1}{2} \left\|p(\ell) - \hat{z}\right\|_{\hat{Z}}^2 + \frac{\kappa}{2} \|\ell\|_H^2 \qquad (32) \\ &\text{subject to } \ell \in A := \{\ell \in H |\ g(\ell) \in -C\}. \end{aligned}$$

Using the chain rule, we get the following derivatives with respect to ℓ:

$$D_\ell J_O(\ell, z)(\delta\ell) = \langle Du(\ell)(\delta\ell), u(\ell) - z\rangle_Z + \kappa \langle \ell, \delta\ell\rangle_H , \qquad (33a)$$

$$\begin{aligned} D_\ell^2 J_O(\ell, z)(\delta\ell, \delta\ell) &= \left\langle D^2u(\ell)(\delta\ell, \delta\ell), u(\ell) - z\right\rangle_Z \\ &\quad + \|Du(\ell)(\delta\ell)\|_Z^2 + \kappa \|\delta\ell\|_H^2 . \qquad (33b) \end{aligned}$$

The following result establishes the local Hölder continuity of the OLS approach:

Theorem 7 *Let $\ell_0 \in A$ be a regular point. Then, for every κ satisfying*

$$\kappa > \frac{2\kappa_2^2 \hat{c}^2 \|m\|_{\widehat{V}^*}}{\kappa_1^2 \sqrt{\kappa_1} \min(\kappa_1, \varsigma_1)} \|u(\ell_0) - z_0\|_Z , \qquad (34)$$

there are neighborhoods U_{z_0} of z_0 and U_{ℓ_0} of ℓ_0 such that for every solution $\ell_z \in U_{\ell_0}$ of (P_z^1), we have

$$\|\ell_z - \ell_0\|_H \leq c \|z - z_0\|_Z^{\frac{1}{2}} ,$$

where $c > 0$ is independent of the data and $\hat{c} > 0$ is the embedding constant.

If $Z = V$, then the following condition on κ ensures the above conclusion:

$$\kappa > \frac{\kappa_2^2 \hat{c}^2}{\min\{\kappa_1, \varsigma_1\}\kappa_1} \|u(\ell_0) - z_0\|_V^2. \qquad (35)$$

Proof We will use Theorem 13 with $X = H$, $D = S \supset A$, and $W = Z$. Evidently, conditions (A1)–(A3) of Theorem 13 hold because J_O and g are twice continuously differentiable with respect to variables (ℓ, z). We only need to show that (65) holds. In other words, we need to show that there exists $\delta > 0$ such that for any feasible ℓ sufficiently close to ℓ_0, we have

$$J_O(\ell, z_0) - J_O(\ell_0, z_0) \geq \delta \|\ell - \ell_0\|^2. \tag{36}$$

By using the Taylor's expansion of J_O at (ℓ_0, z_0), we have

$$J_O(\ell, z_0) - J_O(\ell_0, z_0) = D_\ell J_O(\ell_0, z_0)(\Delta\ell) + \frac{1}{2} D_\ell^2 J_O(\ell_t, z_0)(\Delta\ell, \Delta\ell),$$

where $\Delta\ell = \ell - \ell_0$, $\ell_t = \ell_0 + t(\ell - \ell_0)$, and $t \in (0, 1)$. Using (31), we obtain

$$J_O(\ell, z_0) - J_O(\ell_0, z_0) = -\mu_0(Dg\,(\ell_0)\,(\Delta\ell)) + \frac{1}{2} D_\ell^2 J_O(\ell_t, z_0)(\Delta\ell, \Delta\ell). \tag{37}$$

Note that (30) means that $0 = g(\ell_0) + Dg(\ell_0)(\Delta\ell) + c$, for some $c \in C$, implying

$$0 = \mu_0(g(\ell_0)) + \mu_0(Dg(\ell_0)(\Delta\ell)) + \mu_0(c),$$

and hence $\mu_0(Dg\,(\ell_0)\,(\Delta\ell)) = -\mu_0(c) \leq 0$. It follows from (37) that

$$J_O(\ell, z_0) - J_O(\ell_0, z_0) \geq \frac{1}{2} D_\ell^2 J_O(\ell_t, z_0)(\Delta\ell, \Delta\ell). \tag{38}$$

We need a lower bound for the right-hand side term in the above inequality. For this, by (33), we get

$$\begin{aligned} D_\ell^2 J_O(\ell_t, z_0)(\Delta\ell, \Delta\ell) &= \left\langle D^2 u(\ell_t)(\Delta\ell, \Delta\ell), u(\ell_t) - z_0 \right\rangle_Z \\ &\quad + \|Du(\ell_t)(\Delta\ell)\|_Z^2 + \kappa\, \|\Delta\ell\|_H^2 \\ &\geq \left\langle D^2 u(\ell_t)(\Delta\ell, \Delta\ell), u(\ell_t) - z_0 \right\rangle_Z + \kappa\, \|\Delta\ell\|_H^2 \\ &\geq - \left\| D^2 u(\ell_t)(\Delta\ell, \Delta\ell) \right\|_Z \|u(\ell_t) - z_0\|_Z + \kappa\, \|\Delta\ell\|_H^2 . \end{aligned} \tag{39}$$

Using Theorem 1, we have

$$\begin{aligned} \left\| D^2 u(\ell_t)(\Delta\ell, \Delta\ell) \right\|_Z &\leq \|D^2 u(\ell_t)(\Delta\ell, \Delta\ell)\|_V \\ &\leq \frac{2\kappa_2^2 \|m\|_{\widehat{V}^*}}{\kappa_1^2 \sqrt{\kappa_1}\, \min(\kappa_1, \varsigma_1)} \|\Delta\ell\|_B^2 \\ &\leq \frac{2\kappa_2^2 \hat{c}^2 \|m\|_{\widehat{V}^*}}{\kappa_1^2 \sqrt{\kappa_1}\, \min(\kappa_1, \varsigma_1)} \|\Delta\ell\|_H^2 . \end{aligned} \tag{40}$$

Moreover, invoking Lemma 2, we obtain

$$
\begin{aligned}
\|u(\ell_t) - z_0\|_Z &\leq \|u(\ell_t) - u(\ell_0)\|_Z + \|u(\ell_0) - z_0\|_Z \\
&\leq \|u(\ell_t) - u(\ell_0)\|_V + \|u(\ell_0) - z_0\|_Z \\
&\leq \frac{\kappa_2 \|m\|_{\widehat{V}^*}}{\kappa_1 \sqrt{\kappa_1 \min(\kappa_1, \varsigma_1)}} t \, \|\Delta\ell\|_B + \|u(\ell_0) - z_0\|_Z \\
&\leq \frac{\hat{c}\kappa_2 \|m\|_{\widehat{V}^*}}{\kappa_1 \sqrt{\kappa_1 \min(\kappa_1, \varsigma_1)}} \, \|\Delta\ell\|_H + \|u(\ell_0) - z_0\|_Z \,. \qquad (41)
\end{aligned}
$$

Combining (39)–(41), we obtain

$$
D_\ell^2 J_O(\ell_t, z_0)(\Delta\ell, \Delta\ell) \geq \left(\kappa - \frac{2\kappa_2^2 \hat{c}^2 \|m\|_{\widehat{V}^*}}{\kappa_1^2 \sqrt{\kappa_1 \ \min(\kappa_1, \varsigma_1)}} \, \|u(\ell_0) - z_0\|_Z - \tilde{c} \, \|\Delta\ell\|_H \right) \|\Delta\ell\|_H^2 \,, \qquad (42)
$$

where $\tilde{c}$ is a constant, independent of the measured data, given by

$$
\tilde{c} := \frac{2\kappa_2^3 \hat{c}^3 \|m\|_{\widehat{V}^*}^2}{\kappa_1^4 \min(\kappa_1, \varsigma_1)} > 0.
$$

Due to the assumption (34), that is, due to the inequality

$$
\kappa > \frac{2\kappa_2^2 \hat{c}^2 \|m\|_{\widehat{V}^*}}{\kappa_1^2 \sqrt{\kappa_1 \ \min(\kappa_1, \varsigma_1)}} \, \|u(\ell_0) - z_0\|_Z \,,
$$

we can find a neighborhood U_{ℓ_0} of ℓ_0 such that

$$
\delta := \kappa - \frac{2\kappa_2^2 \hat{c} \|m\|_{\widehat{V}^*}}{\kappa_1^2 \sqrt{\kappa_1 \ \min(\kappa_1, \varsigma_1)}} \, \|u(\ell_0) - z_0\|_Z - \tilde{c} \, \|\Delta\ell\|_H > 0,
$$

and consequently

$$
D_\ell^2 J_O(\ell_t, z_0)(\Delta\ell, \Delta\ell) \geq \delta \, \|\Delta\ell\|_H^2 \,,
$$

for every $\ell \in U_{\ell_0}$, which, when combined with (38), implies that

$$
J_O(\ell, z_0) - J_O(\ell_0, z_0) \geq \frac{\delta}{2} \, \|\Delta\ell\|_H^2 \,, \qquad \text{for every } \ell \in U_{\ell_0},
$$

and hence condition (36) holds when the regularization parameter κ satisfies (34).

To prove (36) under (35) and the case $Z = V$, we write an analogue of (39) as follows

$$\begin{aligned}
&D^2_\ell J_O(\ell_t, z_0)(\Delta\ell, \Delta\ell)\\
&\quad= \left\langle D^2 u(\ell_t)(\Delta\ell, \Delta\ell), u(\ell_t) - z_0 \right\rangle_V + \|Du(\ell_t)(\Delta\ell)\|^2_V + \kappa \, \|\Delta\ell\|^2_H\\
&\quad\geq - \left\| D^2 u(\ell_t)(\Delta\ell, \Delta\ell) \right\|_V \|u(\ell_t) - z_0\|_V + \|Du(\ell_t)(\Delta\ell)\|^2_V + \kappa \, \|\Delta\ell\|^2_H \, ,
\end{aligned}$$

On the other hand, from (11), we can deduce the following inequality

$$\begin{aligned}
\|D^2 u(\ell_t)(\Delta\ell, \Delta\ell)\|_V &\leq \frac{2\kappa_2}{\sqrt{\min(\kappa_1, \varsigma_1)\kappa_1}} \|Du(\ell_t)(\Delta\ell)\|_V \|\Delta\ell\|_B\\
&\leq \frac{2\kappa_2 \hat{c}}{\sqrt{\min(\kappa_1, \varsigma_1)\kappa_1}} \|Du(\ell_t)(\Delta\ell)\|_V \|\Delta\ell\|_H
\end{aligned}$$

and subsequently the following chain of inequalities

$$\begin{aligned}
\|D^2 u(\ell_t)(\Delta\ell, \Delta\ell)\|_V \, \|u(\ell_t) - z_0\|_V \leq{}& \left\| D^2 u(\ell_t)(\Delta\ell, \Delta\ell) \right\|_V \big[\|u(\ell_t) - u(\ell_0)\|_V\\
&+ \|u(\ell_0) - z_0\|_V \big]\\
\leq{}& \left\| D^2 u(\ell_t)(\Delta\ell, \Delta\ell) \right\|_V \|u(\ell_t) - u(\ell_0)\|_V\\
&+ \left\| D^2 u(\ell_t)(\Delta\ell, \Delta\ell) \right\|_V \|u(\ell_0) - z_0\|_V\\
\leq{}& \left\| D^2 u(\ell_t)(\Delta\ell, \Delta\ell) \right\|_V \|u(\ell_t) - u(\ell_0)\|_V\\
&+ \frac{2\kappa_2 \hat{c}}{\sqrt{\min(\kappa_1, \varsigma_1)\kappa_1}} \|Du(\ell_t)(\Delta\ell)\|_V \|\Delta\ell\|_H \, \|u(\ell_0) - z_0\|_V\\
\leq{}& \frac{2\hat{c}^3 \kappa_2^3 \|m\|^2_{\hat{V}^*}}{\kappa_1^4 \min(\kappa_1, \varsigma_1)} \, \|\Delta\ell\|^3_H + \|Du(\ell_t)(\Delta\ell)\|^2_V\\
&+ \frac{\kappa_2^2 \hat{c}^2}{\min\{\kappa_1, \varsigma_1\}\kappa_1} \|\Delta\ell\|^2_H \|u(\ell_0) - z_0\|^2_V,
\end{aligned}$$

implying that

$$D^2_\ell J_O(\ell_t, z_0)(\Delta\ell, \Delta\ell) \geq \left(\kappa - \frac{\kappa_2^2 \hat{c}^2}{\min\{\kappa_1, \varsigma_1\}\kappa_1} \|u(\ell_0) - z_0\|^2_V - \tilde{c} \, \|\Delta\ell\|_H \right) \|\Delta\ell\|^2_H \, ,$$

where

$$\tilde{c} := \frac{2\hat{c}^3\kappa_2^3\|m\|_{\widehat{V}^*}^2}{\kappa_1^4 \min(\kappa_1, \varsigma_1)} > 0$$

is independent of the measured data. The remaining arguments to complete the proof are identical to the ones given for the case of (34). □

In essence, Theorem 7 conveys that the local Holder stability necessitates a lower bound on the regularization parameter which involves a fixed constant and the term $\|u(\ell_0) - z_0\|$ which speaks of the quality of the exact data. The condition $Z = V$ suggests that the data ought to be sufficiently regular. In this case, the lower bound on κ includes the term $\|u(\ell_0)-z_0\|_V^2$, and hence more regular data permits a smaller regularization parameter.

Local Lipschitz stability can be obtain using Theorem 14. However, the main challenge here is to show the bound on the multipliers. For this, we choose the regularization space to be $H := H^2(\Omega)$.

We consider the following regularized optimization problem (Q_z^1):

$$\begin{aligned} \min J_O(\ell, z) &:= \frac{1}{2}\,\|u(\ell) - z\|_Z^2 + \frac{\kappa}{2}\,\|\ell\|_H^2 \\ &\text{subject to } \ell \in A := \{\ell \in H \,|\, 0 < \alpha_0 \le \ell(x)\}. \end{aligned} \tag{43}$$

where $\kappa > 0$ is the regularization parameter and α_0 is a known constant. Evidently, the map $g : H \to H$ is given by $g(\ell) = \alpha_0 - \ell$. We choose C to be the cone of positive functions in H. For each z by ℓ^z we denote the corresponding solution to (Q_z^1).

By using the chain rule, we compute the derivative of the objective functional:

$$D_\ell J_O(\ell, z)(\delta\ell) = \langle Du(\ell)(\delta\ell), u(\ell) - z\rangle_Z + \kappa\,\langle \ell, \delta\ell\rangle_H \tag{44}$$

The following feature of the derivative of the OLS functional will be used shortly:

Lemma 4 *There exist neighborhoods $U_{\ell_0} \subset H$ of ℓ_0, $U_{z_0} \subset Z$ of z_0, respectively, and a constant $c > 0$ such that for every $z_1, z_2 \in U_{z_0}$, and $\ell^{z_1}, \ell^{z_2} \in U_{\ell_0}$, we have*

$$|D_\ell J_O(\ell^{z_1}, z_1)(\delta\ell) - D_\ell J_O(\ell^{z_2}, z_2)(\delta\ell)| \le c\left(\left\|\ell^{z_1} - \ell^{z_2}\right\|_H + \|z_1 - z_2\|_Z\right)\|\delta\ell\|_H\,. \tag{45}$$

Proof Using (44), we obtain

$$\begin{aligned} &D_\ell J_O(\ell^{z_1}, z_1)(\delta\ell) - D_\ell J_O(\ell^{z_2}, z_2)(\delta\ell) \\ &\qquad = \left\langle Du(\ell^{z_1})(\delta\ell), u(\ell^{z_1}) - z_1\right\rangle_Z - \left\langle Du(\ell^{z_2})(\delta\ell), u(\ell^{z_2}) - z_2\right\rangle_Z \\ &\qquad + \kappa\left\langle \ell^{z_1} - \ell^{z_2}, \delta\ell\right\rangle_H \end{aligned}$$

$$= \left\langle Du(\ell^{z_1})(\delta\ell) - Du(\ell^{z_2})(\delta\ell), u(\ell^{z_1}) - z_1 \right\rangle_Z$$
$$+ \left\langle Du(\ell^{z_2})(\delta\ell), u(\ell^{z_1}) - u(\ell^{z_2}) + z_2 - z_1 \right\rangle_Z + \kappa \left\langle \ell^{z_1} - \ell^{z_2}, \delta\ell \right\rangle_H ,$$

and hence

$$|D_\ell J_O(\ell^{z_1}, z_1)(\delta\ell) - D_\ell J_O(\ell^{z_2}, z_2)(\delta\ell)| \tag{46}$$
$$\le \left\| Du(\ell^{z_1})(\delta\ell) - Du(\ell^{z_2})(\delta\ell) \right\|_V \left\| u(\ell^{z_1}) - z_1 \right\|_Z$$
$$+ \left\| Du(\ell^{z_2})(\delta\ell) \right\|_V \left[\left\| u(\ell^{z_1}) - u(\ell^{z_2}) \right\|_V + \|z_1 - z_2\|_Z \right] + \kappa \left\| \ell^{z_1} - \ell^{z_2} \right\|_H \|\delta\ell\|_H . \tag{47}$$

Using the bounds from Theorem 1, we obtain

$$\left\| Du(\ell^{z_1})(\delta\ell) - Du(\ell^{z_2})(\delta\ell) \right\|_V \le c_1 \left\| \ell^{z_1} - \ell^{z_2} \right\|_H \|\delta\ell\|_H , \tag{48}$$

where

$$c_1 = \frac{2\kappa_2^2 \hat{c}^2}{\kappa_1^2 \sqrt{\kappa_1 \ \min(\kappa_1, \varsigma_1)}} \|m\|_{\widehat{V}^*}.$$

We also have the following bounds

$$\left\| Du(\ell^{z_2})(\delta\ell) \right\| \le c_2 \|\delta\ell\|_H , \tag{49}$$
$$\left\| u(\ell^{z_1}) - u(\ell^{z_2}) \right\|_V \le c_2 \left\| \ell^{z_1} - \ell^{z_2} \right\|_H , \tag{50}$$

where

$$c_2 = \frac{\hat{c}\kappa_2}{\kappa_1 \sqrt{\kappa_1 \ \min(\kappa_1, \varsigma_1)}} \|m\|_{\widehat{V}^*}.$$

By using (48)–(50), in (47), we obtain

$$\begin{aligned}
&|D_\ell J_O(\ell^{z_1}, z_1)(\delta\ell) - D_\ell J_O(\ell^{z_2}, z_2)(\delta\ell)| \\
&\qquad \le \left\| Du(\ell^{z_1})(\delta\ell) - Du(\ell^{z_2})(\delta\ell) \right\|_V \left\| u(\ell^{z_1}) - z_1 \right\|_Z \\
&\qquad + \left\| Du(\ell^{z_2})(\delta\ell) \right\|_V \left[\left\| u(\ell^{z_1}) - u(\ell^{z_2}) \right\|_V + \|z_1 - z_2\|_Z \right] \\
&\qquad + \kappa \left\| \ell^{z_1} - \ell^{z_2} \right\|_H \|\delta\ell\|_H \\
&\qquad \le c_1 \left\| u(\ell^{z_1}) - z_1 \right\|_Z \left\| \ell^{z_1} - \ell^{z_2} \right\|_H \|\delta\ell\|_H + c_2^2 \left\| \ell^{z_1} - \ell^{z_2} \right\|_H \|\delta\ell\|_H \\
&\qquad + c_2 \|z_1 - z_2\|_Z \|\delta\ell\|_H + \kappa \left\| \ell^{z_1} - \ell^{z_2} \right\|_H \|\delta\ell\|_H \\
&\qquad \le \Big(c_1 \left\| u(\ell^{z_1}) - z_1 \right\|_Z \left\| \ell^{z_1} - \ell^{z_2} \right\|_H + c_2^2 \left\| \ell^{z_1} - \ell^{z_2} \right\|_H \\
&\qquad + c_2 \|z_1 - z_2\|_Z + \kappa \left\| \ell^{z_1} - \ell^{z_2} \right\|_H \Big) \|\delta\ell\|_H .
\end{aligned}$$

Choosing two bounded neighborhoods, $U_{z_0} \subset Z$ of z_0, and $U_{\ell_0} \subset H$ of ℓ_0, and a constant $c_3 > 0$ such that $c_1 \|u(\ell^{z_1}) - z_1\| \leq c_3$, we write the above inequality as follows

$$\begin{aligned}
&\left|D_\ell J(\ell^{z_1}, z_1)(\delta\ell) - D_\ell J(\ell^{z_2}, z_2)(\delta\ell)\right| \\
&\quad\leq \left(\left(c_3 + c_2^2 + \kappa\right) \left\|\ell^{z_1} - \ell^{z_2}\right\|_H + c_2 \|z_1 - z_2\|_Z\right) \|\delta\ell\|_H \\
&\quad\leq c\left(\left\|\ell^{z_1} - \ell^{z_2}\right\|_H + \|z_1 - z_2\|_Z\right) \|\delta\ell\|_H ,
\end{aligned}$$

where c is a suitable constant. The proof is complete. □

We now establish the local Lipschitz continuity of the OLS approach:

Theorem 8 *Assume that either (34) holds or $Z = V$ and (35) holds. Then there are neighborhoods U_{z_0} of z_0 and U_{ℓ_0} of ℓ_0 such that for every solution $\ell_z \in U_{\ell_0}$ of (Q_z^1), there is a constant c such that*

$$\|\ell_z - \ell_0\|_H \leq c \|z - z_0\|_Z .$$

Proof We will employ Theorem 14. Since we have verified the assumptions (A1)–(A3) of this result in Theorem 7, we only need to show that the assumptions (A4) and (A5) hold.

To verify (A4), we define the Lagrangian functional $L : A \times Z \times Y^* \to \mathbb{R}$ by

$$L(\ell, z, \mu) = J_O(\ell, z) + \langle \mu, g(\ell)\rangle,$$

which due to the identity $D^2 g(\ell_0) = 0$, yields

$$\begin{aligned}
D_\ell^2 L(\ell_0, z_0, \mu)(\delta\ell, \delta\ell) &= D_\ell^2 J_O(\ell_0, z_0)(\delta\ell, \delta\ell) + \langle \mu, D^2 g(\ell_0)(\delta\ell, \delta\ell)\rangle \\
&\geq \delta \|\delta\ell\|_H^2 , \qquad (51)
\end{aligned}$$

where the existence of $\delta > 0$ follows from Theorem 7. Hence (A4) is verified.

Finally, to prove (A5), we first note that every element $\ell \in A$ is a regular point (see Lemma 3 or [31, Lemma 3.1]). Indeed, since the ordering cone C has a nonempty interior (see [31, Remark 3.5]), a Slater constraint qualification holds for each point $\ell \in C$. In fact,

$$g(\bar{\ell}) + Dg(\ell)(\ell - \bar{\ell}) = \alpha_0 - \bar{\ell} + \bar{\ell} - \ell = \alpha_0 - \ell \in -\operatorname{int}(C),$$

by taking $\ell = \alpha_0 + k_F$, where $k_F : D \to \mathbb{R}$ denotes the constant map $k_F(x) = k$ with $k > 0$. Therefore, independently of z, for every ℓ^z solution to (Q^z) there exists a (unique) multiplier $\mu_z \in C^+$ such that

$$D_\ell J_O(\ell^z, z)(\delta\ell) = -\langle \mu_z, Dg(\ell^z)(\delta\ell)\rangle = \langle \mu_z, \delta\ell\rangle = \mu_z(\delta\ell).$$

By Lemma 4, there are neighborhoods $U_{z_0} \subset Z$ of z_0, $U_{\ell_0} \subset H$ of ℓ_0 and a constant $c > 0$ so that

$$\begin{aligned} |\mu_z(\delta\ell) - \mu_{\bar{z}}(\delta\ell)| &= \left| D_\ell J_O(\ell^z, z)(\delta\ell) - D_\ell J_O(\ell^{\bar{z}}, \bar{z})(\delta\ell) \right| \\ &\leq c \left(\left\| \ell^{\bar{z}} - \ell^z \right\|_H + \|\bar{z} - z\|_Z \right) \|\delta\ell\|_H , \end{aligned}$$

and consequently, for every $z \in U_{z_0}$, $\ell^{\bar{z}} \in U_{\ell_0}$ we obtain

$$\|\mu_z - \mu_{\bar{z}}\|_{H^*} \leq c \left(\left\| \ell^{\bar{z}} - \ell^z \right\|_H + \|\bar{z} - z\|_Z \right),$$

and hence condition (A5) of Theorem 14 is also verified. The proof is complete. □

5.2 Stability of the Modified Output Least-Squares Approach

To study stability of the MOLS approach, we consider the regularized optimization problem (P_z^2) :

$$\begin{aligned} \min J_M(\ell, z) := \frac{1}{2} a(\ell, \bar{u}(\ell) - \bar{z}, \bar{u}(\ell) - \bar{z}) + b(\bar{u}(\ell) - \bar{z}, p(\ell) - \hat{z}) \\ - \frac{1}{2} c(p(\ell) - \hat{z}, p(\ell) - \hat{z}) + \frac{\kappa}{2} \|\ell\|_H^2 , \\ \text{subject to } \ell \in A := \{\ell \in H |\ g(\ell) \in -C\}. \end{aligned}$$

We have the following stability result concerning the Hölder continuity:

Theorem 9 *Let $\kappa > 0$ be arbitrary and let $\ell_0 \in A$ be a regular point. Then there are neighborhoods U_{z_0} of z_0, U_{ℓ_0} of ℓ_0 and a constant $c > 0$ such that for every solution $\ell_z \in U_{\ell_0}$ of (P_z^2), we have*

$$\|\ell_z - \ell_0\|_H \leq c \, \|z - z_0\|_Z^{\frac{1}{2}} .$$

Proof We will again apply Theorem 13. Evidently, conditions (A1)–(A3) hold as the maps J and g are twice continuously differentiable with respect to both variables (ℓ, z). To prove (65), we note that

$$\begin{aligned} D_\ell J_M(\ell, z)(\delta\ell) &= -\frac{1}{2} a(\delta\ell, \bar{u}(\ell) + \bar{z}, \bar{u}(\ell) - \bar{z}) + \kappa \, \langle \ell, \delta\ell \rangle_H , \\ D_\ell^2 J_M(\ell, z)(\delta\ell, \delta\ell) &= a(\ell, D\bar{u}(\ell)(\delta\ell), D\bar{u}(\ell)(\delta\ell)) + c(Dp(\ell)(\delta\ell), Dp(\ell)(\delta\ell)) \\ &\quad + \kappa \, \|\delta\ell\|_H^2 , \end{aligned}$$

As in Theorem 7, we have

$$
\begin{aligned}
J_M(\ell, z_0) - J_M(\ell_0, z_0) &\geq \frac{1}{2} D_\ell^2 J_M(\ell_t, z_0)(\Delta\ell, \Delta\ell) \\
&= \frac{1}{2} a(\ell_t, D\bar{u}(\ell_t)(\Delta\ell), D\bar{u}(\ell_t)(\Delta\ell)) + \frac{1}{2} c(Dp(\ell_t)(\Delta\ell), Dp(\ell_t)(\Delta\ell)) \\
&\quad + \frac{\kappa}{2} \|\Delta\ell\|_H^2 \\
&\geq \frac{\kappa_1}{2} \|D\bar{u}(\ell_t)(\Delta\ell)\|_{\widehat{V}}^2 + \frac{\varsigma_1}{2} \|Dp(\ell_t)(\Delta\ell)\|_{\widehat{V}}^2 + \frac{\kappa}{2} \|\Delta\ell\|_H^2 \geq \frac{\kappa}{2} \|\Delta\ell\|_H^2 ,
\end{aligned}
$$

and hence (65) holds for any $\kappa > 0$. The proof is complete. □

Following the same approach as for the case of the OLS functional, in our next result, we take $H := H^2(\Omega)$ to give an improved stability estimate. We consider the following problem (Q_z^2):

$$
\begin{aligned}
\min J_M(\ell, z) := &\frac{1}{2} a(\ell, \bar{u}(\ell) - \bar{z}, \bar{u}(\ell) - \bar{z}) + b(\bar{u}(\ell) - \bar{z}, p(\ell) - \hat{z}) \\
&- \frac{1}{2} c(p(\ell) - \hat{z}, p(\ell) - \hat{z}) + \frac{\kappa}{2} \|\ell\|_H^2 \\
&\text{subject to } \ell \in A := \{\ell \in H \mid 0 < \alpha_0 \leq \ell(x)\}.
\end{aligned}
$$

where $\kappa > 0$ is a fixed regularization parameter.

As before, for each z by ℓ^z we denote the corresponding solution to (Q_z^2). We again choose $Y = H$, and C as the cone of positive functions in H, and define $g : H \to H$ by $g(\ell) = \alpha_0 - \ell$.

The following technical result will be used shortly.

Lemma 5 *There exist neighborhoods $U_{z_0} \subset V$ of z_0, $U_{\ell_0} \subset H$ of ℓ_0, respectively, and a constant $c > 0$ such that for every ℓ^{z_1}, $\ell^{z_2} \in U_{\ell_0}$, z_1, $z_2 \in U_{z_0}$, we have*

$$
\left|D_\ell J_M(\ell^{z_1}, z_1)(\delta\ell) - D_\ell J_M(\ell^{z_2}, z_2)(\delta\ell)\right| \leq c \left(\left\|\ell^{z_1} - \ell^{z_2}\right\|_H + \|z_1 - z_2\|_V\right) \|\delta\ell\|_H . \tag{52}
$$

Proof Using the derivative characterization of the MOLS functional, we have

$$
\begin{aligned}
D_\ell J_M(\ell^{z_1}, z_1)(\delta\ell) - D_\ell J_M(\ell^{z_2}, z_2)(\delta\ell) &= \frac{1}{2} a(\delta\ell, \bar{u}(\ell^{z_2}) - z_2, \bar{u}(\ell^{z_2}) + z_2) \\
&\quad - \frac{1}{2} a(\delta\ell, \bar{u}(\ell^{z_1}) - z_1, \bar{u}(\ell^{z_1}) + z_1) + \kappa \left\langle \ell^{z_1} - \ell^{z_2}, \delta\ell \right\rangle_H \\
&= \frac{1}{2} a(\delta\ell, \bar{u}(\ell^{z_2}) - \bar{u}(\ell^{z_1}) + z_1 - z_2, \bar{u}(\ell^{z_2}) + z_2) \\
&\quad + \frac{1}{2} a(\delta\ell, \bar{u}(\ell^{z_1}) - z_1, \bar{u}(\ell^{z_2}) - \bar{u}(\ell^{z_1}) + z_2 - z_1) + \kappa \left\langle \ell^{z_1} - \ell^{z_2}, \delta\ell \right\rangle_H ,
\end{aligned}
$$

where we used the symmetry and the linearity of trilinear form a.

On a bounded neighborhood of z_0, the following two estimates hold:

$$\begin{aligned}|a(\delta\ell, \bar{u}(\ell^{z_2}) - \bar{u}(\ell^{z_1}) + z_1 - z_2, \bar{u}(\ell^{z_2}) + z_2)| &\leq c_1 \|\delta\ell\|_H \left[\|\bar{u}(\ell^{z_1}) - \bar{u}(\ell^{z_2})\|_V\right. \\ &\quad \left.+ \|z_1 - z_2\|_V\right] \\ |a(\delta\ell, \bar{u}(\ell^{z_1}) - z_1, \bar{u}(\ell^{z_2}) - \bar{u}(\ell^{z_1}) + z_2 - z_1)| &\leq c_2 \|\delta\ell\|_H \left[\|\bar{u}(\ell^{z_1}) - \bar{u}(\ell^{z_2})\|_V\right. \\ &\quad \left.+ \|z_1 - z_2\|_V\right],\end{aligned}$$

where c_1 and c_2 are two constants.

The above inequalities confirm that there exists a constant c_3 such that

$$\begin{aligned}&\left|D_\ell J(\ell^{z_1}, z_1)(\delta\ell) - D_\ell J(\ell^{z_2}, z_2)(\delta\ell)\right| \\ &\quad \leq c_3 \left(\|\bar{u}(\ell^{z_1}) - \bar{u}(\ell^{z_2})\|_V + \|z_1 - z_2\|_V\right) \|\delta\ell\|_H,\end{aligned}$$

and (52) follows using $\|\bar{u}(\ell^{z_2}) - \bar{u}(\ell^{z_1})\|_V \leq c_4 \|\ell^{z_1} - \ell^{z_2}\|_H$ (cf. Lemma 2). The proof is complete. □

The following is the Lipschitz continuity estimate:

Theorem 10 *There is a neighborhood U_{ℓ_0} of ℓ_0 such that for every solution $\ell^z \in U_{\ell_0}$ of (Q_z^2) and a constant $c > 0$ which is independent of the data, we have*

$$\left\|\ell^z - \ell_0\right\|_H \leq c \|z - z_0\|_V.$$

Proof We will again apply Theorem 14. As (A1)–(A3) have already been shown, we only need to show (A4)–(A5). We proceed by defining the Lagrangian functional $L : A \times Z \times Y^* \to \mathbb{R}$ by

$$L(\ell, z, \mu) = J_M(\ell, z) + \langle \mu, g(\ell) \rangle.$$

Since $D^2 g(\ell) = 0$, we have

$$D_\ell^2 L(\ell_0, z_0, \mu)(\delta\ell, \delta\ell) = D_\ell^2 J_M(\ell_0, z_0)(\delta\ell, \delta\ell) + \langle \mu, D^2 g(\ell)(\delta\ell, \delta\ell) \rangle \geq \delta \|\delta\ell\|_H^2,$$

where the coercivity condition follows by the arguments used in Theorem 7. Hence (A4) is verified.

We have already noticed that every point in A is a regular point. Therefore, independently of z, for every ℓ^z solution to (Q_2^z) there exists a (unique) multiplier $\mu_z \in C^+$ such that

$$D_\ell J_M(\ell^z, z)(\delta\ell) = -\langle \mu_z, Dg(\ell^z)(\delta\ell) \rangle = \mu_z(\delta\ell).$$

By applying Lemma 5 there exist neighborhoods $U_{z_0} \subset V$ of z_0, $U_{\ell_0} \subset H$ of ℓ_0 respectively and a constant $c > 0$ such that

$$\begin{aligned}\left|\mu_{z_1}(\delta\ell) - \mu_{z_2}(\delta\ell)\right| &= \left|D_\ell J_M(\ell^{z_1}, z_1)(\ell^{z_1})(\delta\ell) - D_\ell J_M(\ell^{z_2}, z_2)(\ell^{z_2})(\delta\ell)\right| \\ &\leq c\left(\left\|\ell^{z_1} - \ell^{z_2}\right\|_H + \|z_1 - z_2\|_V\right)\|\delta\ell\|_H\end{aligned}$$

which at once implies that

$$\left\|\mu_{z_1} - \mu_{z_2}\right\| \leq c\left(\left\|\ell^{z_1} - \ell^{z_2}\right\|_H + \|z_1 - z_2\|_V\right),$$

for every $z_1, z_2 \in U_{z_0}$, ℓ^{z_1}, $\ell^{z_2} \in U_{\ell_0}$. Therefore (A5) is also verified and the proof is complete. □

5.3 Stability of the Energy Output Least-Squares Approach

We now consider the EOLS based regularized optimization problem (P_z^3) :

$$\min J_E(\ell, z) := \frac{1}{2}a(\ell, \bar{u}(\ell) - \bar{z}, \bar{u}(\ell) - \bar{z}) + \frac{1}{2}c(p(\ell) - \hat{z}, p(\ell) - \hat{z}) + \frac{\kappa}{2}\|\ell\|_H^2$$
$$\text{subject to } \ell \in A := \{\ell \in H |\ g(\ell) \in -C\}.$$

We have the following stability result:

Theorem 11 *Let $\ell_0 \in A$ be a regular point. Then, for every regularization parameter κ satisfying*

$$\kappa > \frac{\kappa_2^2 \hat{c}^2 (\varsigma_2 + \kappa_0)^2}{\min\{\kappa_1, \varsigma_1\}\kappa_1^2}\|u(\ell_0) - z_0\|_V^2, \tag{53}$$

there are neighborhoods $U_{z_0} \subset V$ of z_0, $U_{\ell_0} \subset H$ of ℓ_0, and a constant $c > 0$ such that for every solution ℓ^z in U_{ℓ_0} of (P_z^3), we have

$$\|\ell_z - \ell_0\|_H \leq c\,\|z - z_0\|_V^{\frac{1}{2}}.$$

Proof We shall again use Theorem 13. Conditions (A1)–(A3) hold as J_E and g are twice continuously differentiable with respect to both variables (ℓ, z). Recall that

$$\begin{aligned}D_\ell J_E(\ell, z)(\delta\ell) = &-\frac{1}{2}a(\delta\ell, \bar{u}(\ell) - \bar{z}, \bar{u}(\ell) + \bar{z}) - b(\bar{u}(\ell) - \bar{z}, \delta p) \\ &+ c(\delta p, p(\ell) - \hat{z}) + \kappa\,\langle \ell, \delta\ell\rangle_H\end{aligned}$$

$$\begin{aligned} D^2_\ell J_E(\ell, z)(\delta\ell, \delta\ell) &= a(\ell, \delta\bar{u}, \delta\bar{u}) + c(\delta p, \delta p) - b(\delta\bar{u}, \delta p) - b(\bar{u}(\ell) - \bar{z}, \delta^2 p) \\ &\quad + c(\delta^2 p, p(\ell) - \hat{z}) + c(\delta p, \delta p) + \kappa \left\langle \delta\ell, \delta\ell \right\rangle_H , \end{aligned}$$

and since (9) yields $c(\delta p, \delta p) = b(\delta\bar{u}, \delta p)$, we have

$$\begin{aligned} D^2_\ell J_E(\ell, z)(\delta\ell, \delta\ell) &= a(\ell, \delta\bar{u}, \delta\bar{u}) + c(\delta p, \delta p) + c(\delta^2 p, p(\ell) - \hat{z}) \\ &\quad - b(\bar{u}(\ell) - \bar{z}, \delta^2 p) + \kappa \left\| \delta\ell \right\|^2_H \\ &\geq \kappa_1 \left\| \delta\bar{u} \right\|^2_{\widehat{V}} + \varsigma_1 \left\| \delta p \right\|^2_Q - (\varsigma_2 + \kappa_0) \left\| u(\ell) - z \right\|_V \left\| \delta^2 p \right\|_Q + \kappa \left\| \delta\ell \right\|^2_H \\ &\geq \kappa_1 \left\| \delta\bar{u} \right\|^2_{\widehat{V}} - (\varsigma_2 + \kappa_0) \left\| u(\ell) - z \right\|_V \left\| \delta^2 u \right\|_V + \kappa \left\| \delta\ell \right\|^2_H . \end{aligned}$$

Following the same reasoning as in the proof of Theorem 7, we have

$$\begin{aligned} & J_E(\ell, z_0) - J_E(\ell_0, z_0) \geq \frac{1}{2} D^2_\ell J_E(\ell_t, z_0)(\Delta\ell, \Delta\ell) \\ & \geq \frac{\kappa_1}{2} \left\| D\bar{u}(\ell_t)(\Delta\ell) \right\|^2_{\widehat{V}} - \frac{(\varsigma_2 + \kappa_0)}{2} \left\| u(\ell_t) - z \right\|_V \left\| D^2 u(\ell_t)(\Delta\ell, \Delta\ell) \right\|_V + \frac{\kappa}{2} \left\| \Delta\ell \right\|^2_H . \end{aligned}$$

Moreover, using (11), we can deduce the following identity

$$\left\| D^2 u(\ell_t)(\Delta\ell, \Delta\ell) \right\|_V \leq \frac{2\kappa_2 \hat{c}}{\sqrt{\min(\kappa_1, \varsigma_1)\kappa_1}} \| D\bar{u}(\ell_t)\Delta\ell \|_{\widehat{V}} \| \Delta\ell \|_H$$

and due the chain of inequalities

$$\begin{aligned} & (\varsigma_2 + \kappa_0) \left\| D^2 u(\ell_t)(\Delta\ell, \Delta\ell) \right\|_V \left\| u(\ell_t) - z_0 \right\|_V \\ & \qquad \leq (\varsigma_2 + \kappa_0) \left\| D^2 u(\ell_t)(\Delta\ell, \Delta\ell) \right\|_Z \left[\left\| u(\ell_t) - u(\ell_0) \right\|_V \right. \\ & \qquad \left. + \left\| u(\ell_0) - z_0 \right\|_V \right] \\ & \qquad \leq (\varsigma_2 + \kappa_0) \left\| D^2 u(\ell_t)(\Delta\ell, \Delta\ell) \right\|_V \left\| u(\ell_t) - u(\ell_0) \right\|_V \\ & \qquad + (\varsigma_2 + \kappa_0) \left\| D^2 u(\ell_t)(\Delta\ell, \Delta\ell) \right\|_Z \left\| u(\ell_0) - z_0 \right\|_V \\ & \qquad \leq (\varsigma_2 + \kappa_0) \left\| D^2 u(\ell_t)(\Delta\ell, \Delta\ell) \right\|_V \left\| u(\ell_t) - u(\ell_0) \right\|_V \\ & \qquad + \frac{2(\varsigma_2 + \kappa_0)\kappa_2 \hat{c}}{\sqrt{\min(\kappa_1, \varsigma_1)\kappa_1}\sqrt{\kappa_1}} \sqrt{\kappa_1} \| D\bar{u}(\ell_t)(\Delta\ell) \|_V \| \Delta\ell \|_H \left\| u(\ell_0) - z_0 \right\|_V \\ & \qquad \leq \frac{2\hat{c}^3 \kappa_2^3 \| m \|^2_{\widehat{V}^*} (\varsigma_2 + \kappa_0)}{\kappa_1^4 \min(\kappa_1, \varsigma_1)} \left\| \Delta\ell \right\|^3_H + \kappa_1 \left\| Du(\ell_t)(\Delta\ell) \right\|^2_{\widehat{V}} \end{aligned}$$

$$+\frac{\kappa_2^2\hat{c}^2(\varsigma_2+\kappa_0)^2}{\min\{\kappa_1,\varsigma_1\}\kappa_1^2}\|\Delta\ell\|_H^2\|u(\ell_0)-z_0\|_V^2,$$

we obtain

$$D_\ell^2 J_E(\ell_t,z_0)(\Delta\ell,\Delta\ell)\geq\left(\kappa-\frac{\kappa_2^2\hat{c}^2(\varsigma_2+\kappa_0)^2}{\min\{\kappa_1,\varsigma_1\}\kappa_1^2}\|u(\ell_0)-z_0\|_V^2-\tilde{c}\,\|\Delta\ell\|_H\right)\|\Delta\ell\|_H^2\,,$$

where

$$\tilde{c}:=\frac{2\hat{c}^3\kappa_2^3\|m\|_{\widehat{V}*}^2(\varsigma_2+\kappa_0)}{\kappa_1^4\min(\kappa_1,\varsigma_1)}>0.$$

The rest of proof can be completed by the same reasoning as used in Theorem 7. □

For the improved estimates, we take $H := H^2(\Omega)$ and consider problem (Q_z^3):

$$\min J_E(\ell,z):=\frac{1}{2}a(\ell,\bar{u}(\ell)-\bar{z},\bar{u}(\ell)-\bar{z})+\frac{1}{2}c(p(\ell)-\hat{z},p(\ell)-\hat{z})+\frac{\kappa}{2}\|\ell\|_H^2$$
$$\text{subject to } \ell\in A:=\{\ell\in H:\ 0<\alpha_0\leq\ell(x)\},$$

where $\kappa>0$ is the regularization parameter. Again, for each $z\in V$, the parameter ℓ^z denotes the solution to (Q_z^3)

The following technical result will be used shortly:

Lemma 6 *There exist neighborhoods $U_{z_0}\subset V$ of z_0, $U_{\ell_0}\subset H$ of ℓ_0, respectively, and a constant $c>0$ such that for every ℓ^{z_1}, $\ell^{z_2}\in U_{\ell_0}$, $z_1,z_2\in U_{z_0}$, we have*

$$\left|D_\ell J_E(\ell^{z_1},z_1)(\delta\ell)-D_\ell J_E(\ell^{z_2},z_2)(\delta\ell)\right|\leq c\left(\left\|\ell^{z_1}-\ell^{z_2}\right\|_H+\|z_1-z_2\|_V\right)\|\delta\ell\|_H\,. \tag{54}$$

Proof In previous result, we have seen that

$$D_\ell J_E(\ell,z)(\delta\ell)=-\frac{1}{2}a(\delta\ell,\bar{u}(\ell)-\bar{z},\bar{u}(\ell)+\bar{z})-b(\bar{u}(\ell)-\bar{z},\delta p)+c(\delta p,p(\ell)-\hat{z})$$
$$+\kappa\langle\ell,\delta\ell\rangle_H$$

Since the first term $-\frac{1}{2}a(\delta\ell,\bar{u}(\cdot)-\bar{z},\bar{u}(\cdot)+\bar{z})$ correspond with the derivative of functional by Lemma 5 we only have to verify the result for the term

$$D_\ell H(\ell^z,z):=-b(\bar{u}(\ell)-\bar{z},\delta p)+c(\delta p,p(\ell)-\hat{z})$$

In fact,

$$D_\ell H(\ell^{z_1},z_1)(\delta\ell)-D_\ell H(\ell^{z_2},z_2)(\delta\ell)=-b(\bar{u}(\ell^{z_1})-\bar{z}_1,Dp(\ell^{z_1})(\delta\ell))$$

$$\begin{aligned}
&+ c(Dp(\ell^{z_1})(\delta\ell), p(\ell^{z_1}) - \hat{z}_1) + b(\bar{u}(\ell^{z_2}) - \bar{z}_2, Dp(\ell^{z_2})(\delta\ell)) \\
&- c(Dp(\ell^{z_2})(\delta\ell), p(\ell^{z_2}) - \hat{z}_2) \\
&= b(\bar{u}(\ell^{z_2}) - \bar{u}(\ell^{z_1}) + \bar{z}_1 - \bar{z}_2, Dp(\ell^{z_2})(\delta\ell)) \\
&+ b(\bar{u}(\ell^{z_1}) - \bar{z}_1, Dp(\ell^{z_2})(\delta\ell) - Dp(\ell^{z_1})(\delta\ell)) \\
&+ c(Dp(\ell^{z_1})(\delta\ell) - Dp(\ell^{z_2})(\delta\ell), p(\ell^{z_1}) - \hat{z}_1) \\
&+ c(Dp(\ell^{z_2})(\delta\ell), p(\ell^{z_1}) - p(\ell^{z_2}) + \hat{z}_2 - \hat{z}_1) \\
&\leq \kappa_0 \left(\left\| \bar{u}(\ell^{z_2}) - \bar{u}(\ell^{z_1}) \right\|_{\widehat{V}} + \|\bar{z}_1 - \bar{z}_2\|_{\widehat{V}} \right) \left\| Dp(\ell^{z_2})(\delta\ell) \right\|_Q \\
&+ \kappa_0 \left\| \bar{u}(\ell^{z_1}) - \bar{z}_1 \right\|_{\widehat{V}} \left\| Dp(\ell^{z_2})(\delta\ell) - Dp(\ell^{z_1})(\delta\ell)) \right\|_Q \\
&+ \varsigma_2 \left\| Dp(\ell^{z_1})(\delta\ell) - Dp(\ell^{z_2})(\delta\ell) \right\|_Q \left\| p(\ell^{z_1}) - \hat{z}_1 \right\|_Q \\
&+ \varsigma_2 \left(\left\| p(\ell^{z_1}) - p(\ell^{z_2}) \right\|_Q + \left\| \hat{z}_2 - \hat{z}_1 \right\|_Q \right) \left\| Dp(\ell^{z_2})(\delta\ell) \right\|_Q
\end{aligned}$$

As for previous results, applying estimates of Lemma 2 and Theorem 1 it is easily seen that

$$\left| D_\ell H(\ell^{z_1}, z_1)(\delta\ell) - D_\ell H(\ell^{z_2}, z_2)(\delta\ell) \right| \leq c \left(\left\| \ell^{z_1} - \ell^{z_2} \right\|_H + \|z_1 - z_2\|_V \right) \|\delta\ell\|_H .$$

on a bounded neighborhood of z_0, for some appropriate constant $c > 0$. □

As a consequence of Theorem 11 and Lemma 6, and following the same reasoning as in Theorem 8, we have the following estimate.

Theorem 12 *Assume that ℓ_0 is regular and* (53) *holds. Then here are neighborhoods $U_{z_0} \subset V$ of z_0, $U_{\ell_0} \subset H$ of ℓ_0, and a constant $c > 0$ such that for every solution ℓ^z in U_{ℓ_0} of (Q_z^3), we have constant c such that*

$$\|\ell_z - \ell_0\|_H \leq c \|z - z_0\|_Z .$$

6 Computational Results

We now present two examples for the inverse problem of identifying a parameter μ on a two-dimensional domain $\Omega = (0, 1) \times (0, 1)$ with boundary $\partial\Omega = \Gamma_1 \times \Gamma_2$. The coefficients were identified in a finite dimensional space of dimension of 1522 on a mesh with 2901 triangles. Since we focus on the recovery of parameters in nearly incompressible materials, λ is taken as a large constant, typically $\lambda = 10^6$.

All experiments here are of a synthetic nature, and we used an adaptive mesh to obtain an accurate solution and then used it for the data z. The optimization was performed using the Newton method. For simplicity, the H^1 semi-norm regularization was used and the regularization parameter was chosen by trial and error.

6.1 Elasticity Imaging Inverse Problem

Given the domain $\varOmega$ as a subset of $\mathbb{R}^2$ or $\mathbb{R}^3$ and $\partial\varOmega = \varGamma_1 \cup \varGamma_2$ as its boundary, the following system models the response of an isotropic elastic body to the known body forces and boundary traction:

$$-\nabla \cdot \sigma = f \text{ in } \varOmega, \tag{55a}$$

$$\sigma = 2\mu\epsilon(u) + \lambda \text{div} u\, I, \tag{55b}$$

$$u = g \text{ on } \varGamma_1, \tag{55c}$$

$$\sigma n = h \text{ on } \varGamma_2. \tag{55d}$$

In (55), the vector-valued function $u = u(x)$ is the displacement of the elastic body, f is the applied body force, n is the unit outward normal, and $\epsilon(u) = \frac{1}{2}(\nabla u + \nabla u^{\mathrm{T}})$ is the linearized strain tensor. The resulting stress tensor σ in the stress-strain law (55b) is obtained under the condition that the elastic body is isotropic and the displacement is sufficiently small so that a linear relationship remains valid. Here μ and λ are the Lamé parameters which quantify the elastic properties of the object.

Our focus is on studying the elasticity imaging inverse problem of locating soft inclusions in an incompressible object, for example, cancerous tumor in the human body. From a mathematical standpoint, this inverse problem seeks μ from a measurement of the displacement vector u under the assumption that the parameter λ is very large. The fundamental idea behind the elasticity imaging inverse problem is that the stiffness of soft tissue can vary significantly based on its molecular makeup, and varying macroscopic/microscopic structure and such changes in stiffness are related to changes in tissue health. In other words, the elasticity imaging inverse problem mathematically mimics the practice of palpation by making use of the differing elastic properties of healthy and unhealthy tissue to identify tumors. In most of the existing literature on the elasticity imaging inverse problem, the human body is modeled as an incompressible elastic object. Although this assumption simplifies the identification process as there is only one parameter μ to identify, it significantly complicates the computational process as the classical finite element methods become entirely ineffective due to the so-called locking effect. One of the few techniques to handle this problem is by resorting to mixed finite element formulation. We explain this in the following. For the time being, in (55), we set $g = 0$. For this case, the space of test functions, denoted by V, is given by:

$$V = \{\bar{v} \in H^1(\varOmega) \times H^1(\varOmega) : \ \bar{v} = 0 \text{ on } \varGamma_1\}.$$

By using the Green's identity and the boundary conditions (55c) and (55d), we obtain the following weak form of the elasticity system (55): Find $\bar{u} \in V$ such that

$$\int_\varOmega 2\mu\epsilon(\bar{u}) \cdot \epsilon(\bar{v}) + \int_\varOmega \lambda(\text{div}\, \bar{u})(\text{div}\, \bar{v}) = \int_\varOmega f\bar{v} + \int_{\varGamma_2} \bar{v}h, \quad \text{for every } \bar{v} \in V. \tag{56}$$

The mixed finite elements approach then consists of introducing a pressure term $p \in Q = L^2(\Omega)$

$$p = \lambda(\operatorname{div} \bar{u}), \tag{57}$$

or equivalently,

$$\int_\Omega (\operatorname{div} \bar{u})q - \int_\Omega \frac{1}{\lambda} pq = 0, \quad \text{for every } q \in Q. \tag{58}$$

By using relation (57), the weak form (56) reads: Find $\bar{u} \in V$ such that

$$\int_\Omega 2\mu\epsilon(\bar{u}) \cdot \epsilon(\bar{v}) + \int_\Omega p(\operatorname{div} \bar{v}) = \int_\Omega f\bar{v} + \int_{\Gamma_2} \bar{v}h, \quad \text{for every } \bar{v} \in V. \tag{59}$$

The problem of finding $\bar{u} \in V$ satisfying (56) has now been reformulated as the problem of finding $(\bar{u}, p) \in V \times Q$ satisfying the mixed variational problems (58) and (59) (Fig. 1).

We now present a numerical example to identify a parameter μ in (55) where the top and bottom domain boundaries (Γ_1) are fixed with constant Dirichlet condition $g(x, y)$ and the left and right boundaries (Γ_2) have Neumann condition $h(x, y)$. The functions defining the coefficient, load, and boundary conditions are as follows:

$$\mu(x, y) = \left(1 - 0.12\cos(3\pi\sqrt{x^2 + y^2})\right)^{-1}, \quad f(x, y) = \begin{bmatrix} 1 + 0.1x^2 \\ 0.1(1 + y) \end{bmatrix},$$

$$g(x, y) = \begin{bmatrix} 0 \\ 0.1 \end{bmatrix} \text{ on } \Gamma_1, \qquad h(x, y) = \begin{bmatrix} 0.1x \\ 0.5 + y^2 \end{bmatrix} \text{ on } \Gamma_2.$$

6.2 *Identification in Stokes Equations*

We now consider Stokes equations

$$-\nabla \cdot (\mu\nabla u) + \nabla p = f \quad \text{in } \Omega, \tag{60a}$$

$$-\operatorname{div} u = 0 \quad \text{in } \Omega, \tag{60b}$$

where u can be considered as the velocity field of an incompressible fluid motion, and p is then the associated pressure, μ is the viscosity coefficient of the fluid. Here we consider homogeneous Dirichlet boundary condition for the velocity, i.e. $u|_{\partial\Omega} = 0$. By multiplying $v \in H_0^1(\Omega)$ to (60a) and $q \in L^2(\Omega)$ to the mass equation (60b), and applying integration by part for the momentum equation, we

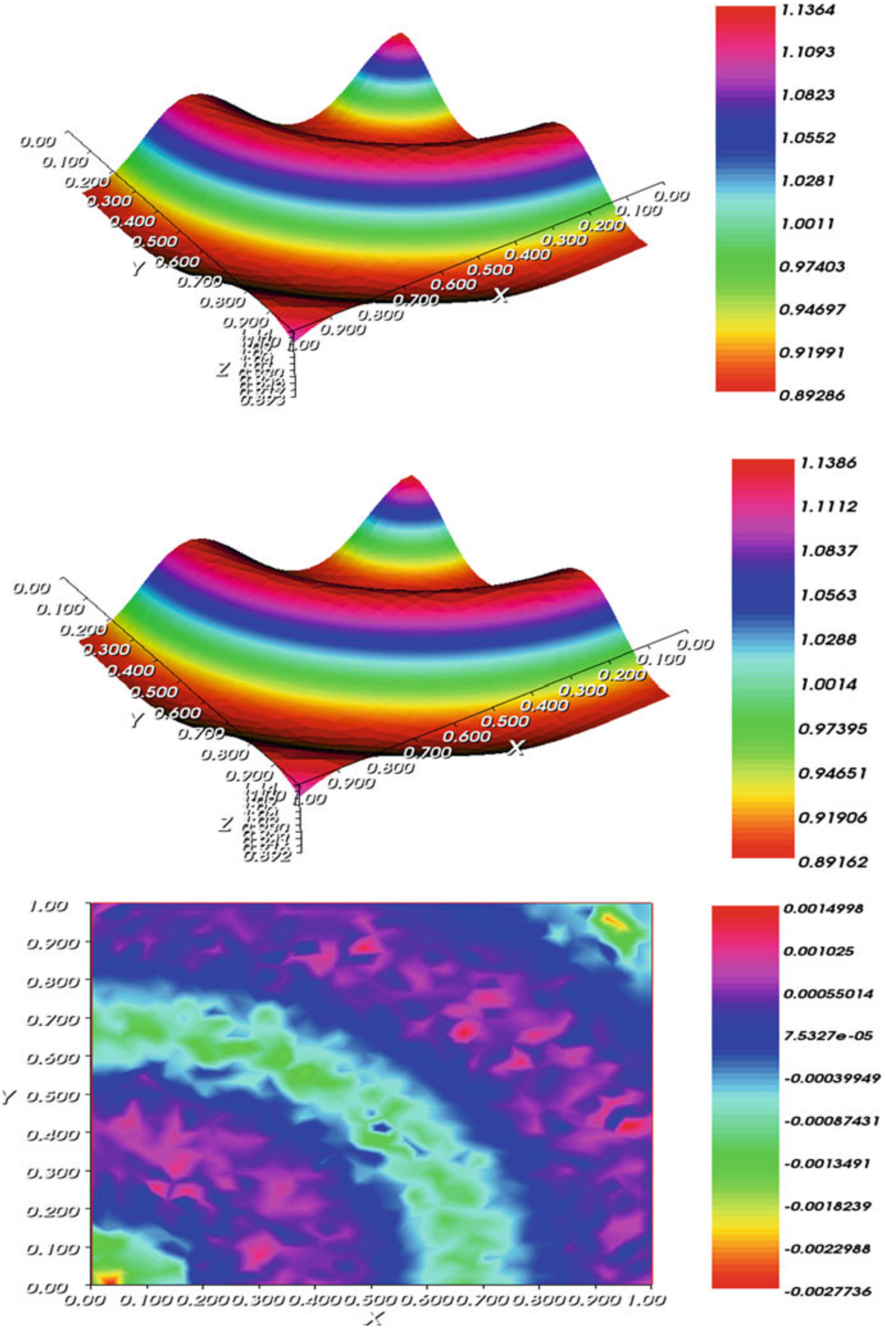


Fig. 1 Reconstruction for the elasticity imaging. The top figure shows the exact coefficient, the middle figure shows that estimated coefficient, and the bottom figure shows the error

obtain the following weak form of the Stokes equations (60): Find $u \in (H_0^1(\Omega))^2$ and a pressure $p \in L^2(\Omega)$ such that

$$\int_\Omega \mu \nabla u \cdot \nabla v - \int_\Omega p(\text{div } v) = \int_\Omega f v, \quad \text{for every } v \in (H_0^1(\Omega))^2 \tag{61}$$

$$- \int_\Omega (\text{div } u) q = 0 \qquad \text{for every } q \in L^2(\Omega). \tag{62}$$

The Stokes equations (60) can be introduced by the saddle point problem (6) with the following setting:

$$a(\mu, u, v) = \int_\Omega \mu \nabla u \cdot \nabla v, \; b(u, q) = - \int_\Omega (\text{div } u) q, \; c(p, q)$$
$$= \int_\Omega \frac{1}{\lambda} pq, \; m(v) = \int_\Omega f v,$$

where $c(p, q)$ is the penalization that removes the zero mean restriction on pressure. Figure 2 shows the numerical results for (60) with $\lambda = 10^6$ and

$$\mu(x, y) = 1 - \frac{1}{2} \sin\left[2\pi \left(x + \frac{1}{10}\right)\left(y + \frac{1}{10}\right)\right], \quad f(x, y) = \begin{bmatrix} -\frac{1}{5}x \\ \cos(\pi x) \end{bmatrix}.$$

6.3 Performance Analysis

We start all the methods with the same initial guess and under the same stopping criteria. Table 1 shows that the MOLS and EOLS functional require fewer iterations to converge to the solution. Also, we compare the minimum eigenvalues of the Hessian of the MOLS, the EOLS, and the OLS functional over some Newton algorithm iterations applied to the elasticity imaging inverse problem in Fig. 3. The minimum and the maximum of $\delta_{\min}$ of the Hessian matrix along with the L^2 error and a total number of algorithm iterations for all examples are given in Table 1.

6.4 Error Analysis for Decreasing λ

We also consider the behavior of the MOLS functional for various values of λ to study its general insusceptibility to the locking effect. Figure 4 shows that the $L_2(\Omega)$ norm of the residual error $||e|| = ||\mu_{\text{estimated}} - \mu_{\text{exact}}||$ are bounded and decreasing against finer mesh sizes.

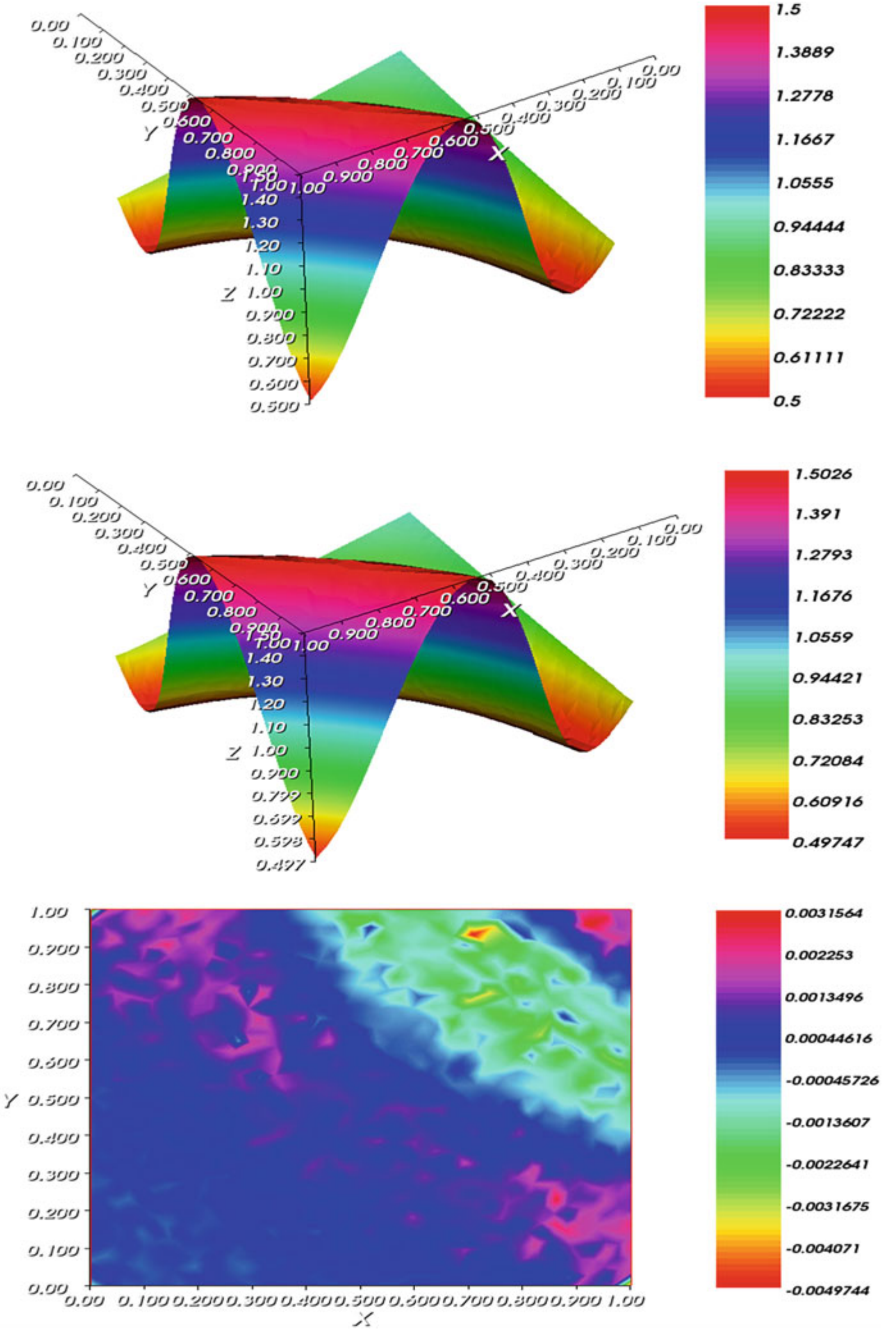


Fig. 2 Reconstruction for Stokes equations. The top figure shows the exact coefficient, the middle figure shows that estimated coefficient, and the bottom figure shows the error

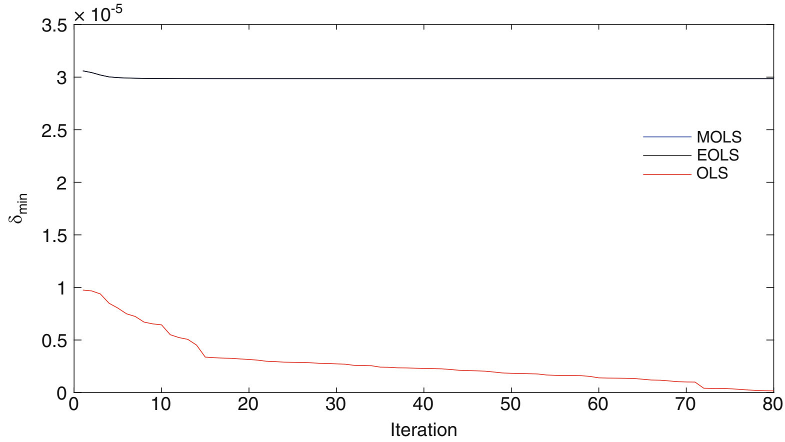


Fig. 3 Comparison of the minimum eigenvalue $\delta_{\min}$ of the Hessian for elasticity imaging

Table 1 Performance comparison for the MOLS, the EOLS, and the OLS approaches

Method	Iterations	L^2-error	Min. $\delta_{\min}$	Max. $\delta_{\max}$
Elasticity imaging: $\kappa = 10^{-5}$ *and* $\lambda = 10^6$				
MOLS	47	1.4235×10^{-6}	2.9855×10^{-5}	3.0598×10^{-5}
EOLS	49	1.1342×10^{-6}	2.9860×10^{-5}	3.1842×10^{-5}
OLS	88	3.7083×10^{-4}	1.6307×10^{-7}	9.7479×10^{-6}
Stokes equations: $\kappa = 10^{-6}$ *and* $\lambda = 10^6$				
MOLS	77	9.8614×10^{-7}	1.7541×10^{-6}	1.8284×10^{-6}
EOLS	79	9.7618×10^{-7}	1.7540×10^{-6}	1.8184×10^{-6}
OLS	134	2.7252×10^{-5}	4.1311×10^{-8}	2.0312×10^{-7}

7 Concluding Remarks

In this work, we performed a detailed study of various aspects of the convex MOLS functional, and in general, nonconvex OLS and EOLS functionals for the inverse problem of parameter identification in an abstract mixed variational problem. We developed a rigorous regularization framework and gave new stability results. Our numerical results showed the feasibility of our approach. There are many directions of research that we would like to pursue. One of our priorities is to extend our stability results estimates for nonquadratic regularization. We anticipate that the framework developed by Resmerita and Scherzer [30], which is based on the convexity arguments, can be quite useful. We would also like to conduct a thorough numerical experimentation to verify the stability estimates for varying regularity of the data. We are also keen on extending the developed framework for the identification of uncertain parameters (see [6]).

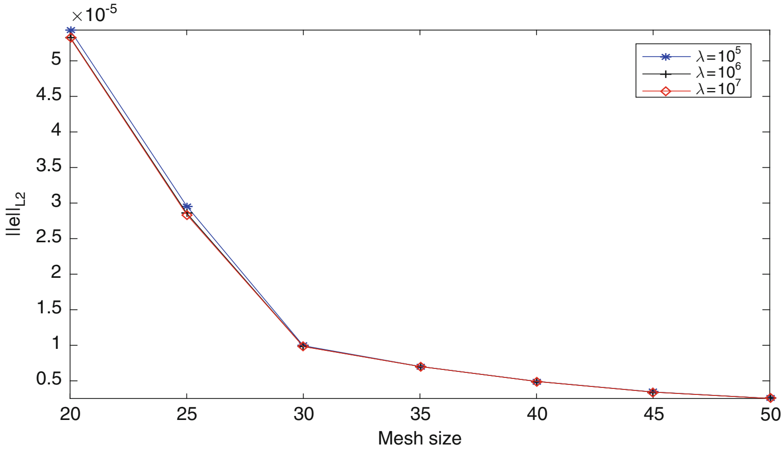


Fig. 4 Error comparison for increasing λ against mesh size

Acknowledgements The research of Akhtar Khan is supported by a National Science Foundation grant (Award No. 1720067). Miguel Sama's work is partially supported by Ministerio de Economìa y Competitividad (Spain), project MTM2015-68103-P.

Appendix: Tools from Stability and Optimization Theory

We collect a few results from abstract perturbation and optimization theory. Let X and Y be Banach spaces, let W be a normed space, and let D be an open subset of X. Let $K \subset Y$ be a pointed, closed, convex cone with apex at origin, and let K^+ be its dual. Let $f : D \times W \to \mathbb{R}$ and $g : D \times W \to Y$ be given single-valued maps. For $w \in W$, we consider the following perturbed optimization problem:

$$(P_w) \qquad \text{minimize } f(x, w) \text{ subject to } g(x, w) \in K. \tag{63}$$

For a fixed $w_0 \in W$, the above problem is the unperturbed problem. Let x_0 be a solution of the unperturbed problem. We make the following assumptions:

(A1) There is a neighborhood $N_1(w_0)$ of w_0 such that for all $w \in N_1(w_0)$, the mappings $f(\cdot, w)$ and $g(\cdot, w)$ are twice continuously differentiable on D.

(A2) There is a neighborhood $N_1(x_0)$ of w_0 and there are constants L_f, L'_f, L_g, L'_g such that for all $w_1, w_2 \in N_1(w_0)$ and for all $x_1, x_2 \in N_1(x_0)$, the following inequalities hold:

$$|f(x_1, w_1) - f(x_2, w_2)| \le L_f \left[\|x_1 - x_2\| + \|w_1 - w_2\|\right],$$

$$|f_x(x_1, w_1) - f_x(x_2, w_2)| \le L'_f \left[\|x_1 - x_2\| + \|w_1 - w_2\|\right],$$

$$|g(x_1, w_1) - g(x_2, w_2)| \leq L_g \left[\|x_1 - x_2\| + \|w_1 - w_2\|\right],$$
$$|g_x(x_1, w_1) - g_x(x_2, w_2)| \leq L'_g \left[\|x_1 - x_2\| + \|w_1 - w_2\|\right].$$

(A3) f_{xx} and $g_{x,x}$ are continuous at (x_0, w_0).

(A4) There exists a constant $\delta_0 > 0$ and a Lagrange multiplier λ_0 for x_0 so that

$$L_{x,x}(x_0, \lambda_0, w_0)(h, h) = (f_{x,x}(x_0, w_0) - \lambda_0 g_{x,x}(x_0, w_0))(h, h) \geq \delta_0 \|h\|^2,$$

for every $h = x - x_0$ with $x \in T_0 = \{x \in X |\ g(x_0, w_0) + g_x(x_0, w_0)(x - x_0) \in K\}$.

(A5) There are neighborhoods $N(w_0)$ of w_0 and $N(x_0)$ of x_0 so that if $w \in N(w_0)$ and x_w is a solution of (P_w) on $\Sigma(w) \cap N(x_0)$ then there are a constant k independent of w and a multiplier λ_w with

$$\|\lambda_w - \lambda_0\| \leq k(\|x_w - x_0\| + \|w - w_0\|). \tag{64}$$

We denote the feasible set by $\Sigma(w) := \{x \in D |\ g(x, w) \in K\}$.

The following result from Alt [3, Theorem 2.5]:

Theorem 13 *Let x_0 be a solution of (P_{w_0}). Suppose x_0 is a regular point and assumptions (A1)–(A3) hold. Suppose further that there is a neighborhood $N_2(x_0)$ of x_0 and a constant $\delta > 0$ such that for every $x \in \Sigma(w) \cap N_2(x_0)$, we have*

$$f(x, w_0) - f(x_0, w_0) \geq \delta \|x - x_0\|^2. \tag{65}$$

Then there is a $r > 0$ and a neighborhood $N(w_0)$ of w_0 so that for all $w \in N(w_0)$ the following holds: If x_w is a solution of P_w on $\Sigma(w) \cap B_r(x_0)$, then $x_w \in int B_r(x_0)$ and for a constant k, we have

$$\|x_w - x_0\| \leq k \|z - z_0\|^{1/2}.$$

We will also use the following interesting result by Alt [3, Theorem 3.5]:

Theorem 14 *Let assumptions (A1)–(A3) be fulfilled. Suppose x_0 is a regular solution of (P_{w_0}) and let λ_0 is a Lagrange multiplier for x_0 such that assumptions (A4)–(A5) hold. Then there is a $r > 0$ and a neighborhood $N(w_0)$ of w_0 such that for all $w \in N(w_0)$ the following holds: If x_w is a solution of (P_w), on $\Sigma(w) \cap B(x_0)$, then $x \in$ int $B_r(x_0)$ and for a constant k independent of w, we have*

$$\|x_w - x_0\| \leq k \|w - w_0\|.$$

References

1. R. Acar, C.R. Vogel, Analysis of bounded variation penalty methods for ill-posed problems. Inverse Prob. **10**(6), 1217–1229 (1994)
2. W. Alt, Stability of solutions for a class of nonlinear cone constrained optimization problems. II. Application to parameter estimation. Numer. Funct. Anal. Optim. **10**(11–12), 1065–1076 (1989)
3. W. Alt, Stability of solutions for a class of nonlinear cone constrained optimization problems, part 1: basic theory. Numer. Funct. Anal. Optim. **10**(11–12), 1053–1064 (1989)
4. C. Boehm, M. Ulbrich, A semismooth Newton-CG method for constrained parameter identification in seismic tomography. SIAM J. Sci. Comput. **37**(5), S334–S364 (2015)
5. R. Boiger, B. Kaltenbacher, An online parameter identification method for time dependent partial differential equations. Inverse Prob. **32**(4), 045006, 28 (2016)
6. J. Borggaard, H.-W. van Wyk, Gradient-based estimation of uncertain parameters for elliptic partial differential equations. Inverse Prob. **31**(6), 065008, 33 (2015)
7. G. Chavent, Local stability of the output least square parameter estimation technique. Math. Appl. Comput. **2**(1), 3–22 (1983)
8. M. Cho, B. Jadamba, R. Kahler, A.A. Khan, M. Sama, First-order and second-order adjoint methods for the inverse problem of identifying nonlinear parameters in PDEs, in *Industrial Mathematics and Complex Systems* (Springer, Berlin, 2017), pp. 1–16
9. C. Clason, L^∞ fitting for inverse problems with uniform noise. Inverse Prob. **28**(10), 104007, 18 (2012)
10. F. Colonius, K. Kunisch, Output least squares stability in elliptic systems. Appl. Math. Optim. **19**(1), 33–63 (1989)
11. F. Colonius, K. Kunisch, Stability of perturbed optimization problems with applications to parameter estimation. Numer. Funct. Anal. Optim. **11**(9–10), 873–915 (1990)
12. E. Crossen, M.S. Gockenbach, B. Jadamba, A.A. Khan, B. Winkler, An equation error approach for the elasticity imaging inverse problem for predicting tumor location. Comput. Math. Appl. **67**(1), 122–135 (2014)
13. M.M. Doyley, B. Jadamba, A.A. Khan, M. Sama, B. Winkler, A new energy inversion for parameter identification in saddle point problems with an application to the elasticity imaging inverse problem of predicting tumor location. Numer. Funct. Anal. Optim. **35**(7–9), 984–1017 (2014)
14. R.O. Evstigneev, M.Y. Medvedik, Y.G. Smirnov, Inverse problem of determining parameters of inhomogeneity of a body from acoustic field measurements. Comput. Math. Math. Phys. **56**(3), 483–490 (2016)
15. A. Gholami, A. Mang, G. Biros, An inverse problem formulation for parameter estimation of a reaction-diffusion model of low grade gliomas. J. Math. Biol. **72**(1–2), 409–433 (2016)
16. E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*. Monographs in Mathematics, vol. 80 (Birkhäuser Verlag, Basel, 1984)
17. M.S. Gockenbach, A.A. Khan, Identification of Lamé parameters in linear elasticity: a fixed point approach. J. Ind. Manag. Optim. **1**(4), 487–497 (2005)
18. M.S. Gockenbach, A.A. Khan, An abstract framework for elliptic inverse problems: part 1. An output least-squares approach. Math. Mech. Solids **12**(3), 259–276 (2007)
19. M.S. Gockenbach, A.A. Khan, An abstract framework for elliptic inverse problems. II. An augmented Lagrangian approach. Math. Mech. Solids **14**(6), 517–539 (2009)
20. M.S. Gockenbach, B. Jadamba, A.A. Khan, Numerical estimation of discontinuous coefficients by the method of equation error. Int. J. Math. Comput. Sci. **1**(3), 343–359 (2006)
21. M.S. Gockenbach, B. Jadamba, A.A. Khan, C. Tammer, B. Winkler, Proximal methods for the elastography inverse problem of tumor identification using an equation error approach, in *Advances in Variational and Hemivariational Inequalities*. Advances in Mechanics and Mathematics, vol. 33 (Springer, Cham, 2015), pp. 173–197

22. S. Guchhait, B. Banerjee, Constitutive error based material parameter estimation procedure for hyperelastic material. Comput. Methods Appl. Mech. Eng. **297**, 455–475 (2015)
23. B. Jadamba, A.A. Khan, G. Rus, M. Sama, B. Winkler, A new convex inversion framework for parameter identification in saddle point problems with an application to the elasticity imaging inverse problem of predicting tumor location. SIAM J. Appl. Math. **74**(5), 1486–1510 (2014)
24. B. Jadamba, A.A. Khan, A. Oberai, M. Sama, First-order and second-order adjoint methods for parameter identification problems with an application to the elasticity imaging inverse problem. Inverse Prob. Sci. Eng. **25**, 1768–1787 (2017)
25. S. Kindermann, L.D. Mutimbu, E. Resmerita, A numerical study of heuristic parameter choice rules for total variation regularization. J. Inverse Ill-Posed Prob. **22**(1), 63–94 (2014)
26. I. Knowles, Parameter identification for elliptic problems. J. Comput. Appl. Math. **131**(1–2), 175–194 (2001)
27. P. Kuchment, D. Steinhauer, Stabilizing inverse problems by internal data. II: non-local internal data and generic linearized uniqueness. Anal. Math. Phys. **5**(4), 391–425 (2015)
28. T. Liu, A wavelet multiscale-homotopy method for the parameter identification problem of partial differential equations. Comput. Math. Appl. **71**(7), 1519–1523 (2016)
29. A. Neubauer, T. Hein, B. Hofmann, S. Kindermann, U. Tautenhahn, Improved and extended results for enhanced convergence rates of Tikhonov regularization in Banach spaces. Appl. Anal. **89**(11), 1729–1743 (2010)
30. E. Resmerita, O. Scherzer, Error estimates for non-quadratic regularization and the relation to enhancement. Inverse Prob. **22**(3), 801–814 (2006)
31. L.W. White, Estimation of flexural rigidity in a Kirchhoff plate model. Appl. Math. Comput. **27**(4, Pt II), 337–359 (1988)
32. L.W. White, Stability of optimal output least squares estimators in certain beams and plate models. Appl. Anal. **39**(1), 15–33 (1990)
33. Y. Xu, J. Zou, Convergence of an adaptive finite element method for distributed flux reconstruction. Math. Comput. **84**(296), 2645–2663 (2015)

Nonlinear Duality in Banach Spaces and Applications to Finance and Elasticity

G. Colajanni, Patrizia Daniele, Sofia Giuffrè, and Antonino Maugeri

1 The Strong Duality in the Infinite-Dimensional Setting

The duality theory we intend to study can be summarized as follows.

Let $f : S \to \mathbb{R}$, $g : S \to Y$, $h : S \to Z$ be three mappings, where S here and in what follows is a convex subset of a real normed space X, Y is a real normed space ordered by a convex cone C, Z is a real normed space and consider the optimization problem:

$$\begin{cases} f(x_0) = \min\limits_{x \in \mathbb{K}} f(x) \\ x_0 \in \mathbb{K} = \{x \in S : g(x) \in -C,\ \ h(x) = \theta_Z\}, \end{cases} \tag{1}$$

where θ_Z is the zero element in the space Z.

The Lagrange dual problem is:

$$\max_{u \in C^*,\, v \in Z^*} \inf_{x \in S} \left[f(x) + \langle u, g(x) \rangle + \langle v, h(x) \rangle \right], \tag{2}$$

where

$$C^* := \left\{ u \in Y^* : \langle u, y \rangle \geq 0,\ \ \forall y \in C \right\}$$

is the dual cone of C and Z^* is the dual space of Z.

G. Colajanni · P. Daniele (✉) · A. Maugeri
Department of Mathematics and Computer Science, University of Catania, Catania, Italy
e-mail: colajanni@dmi.unict.it; daniele@dmi.unict.it; maugeri@dmi.unict.it

S. Giuffrè
D.I.I.E.S. "Mediterranea" University of Reggio Calabria, Reggio Calabria, Italy
e-mail: sofia.giuffre@unirc.it

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_5

Then, we say that the strong duality holds for problems (1) and (2) if and only if problems (1) and (2) admit a solution and the optimal values coincide.

The already classical results by Rockafellar [47], Holmes [36], Borwein and Lewis [3] give sufficient conditions in order that the strong duality between problems (1) and (2) holds.

All these conditions use concepts such as:

- the core:

$$Core\, C := \{x \in C : Cone\,(C + \{\, x\}) = X\}\,;$$

- the intrinsic core:

$$Intrinsic\ Core\ C := \left\{c \in C : \forall c' \in \ \mathrm{aff}\,(C) \setminus \{c\},\ \text{we have } (c, c') \cap C \neq \emptyset\right\},$$

where aff (C) is the affine hull of C and $(a, b) := \{(1-t)a + tb : t \in (0, 1)\}\,;$
- strong quasi-relative interior of C :

$$sqri\ \ C := \{x \in C : Cone\ \ (C - \{x\}) \text{ is a closed linear subspace of } X\}\,.$$

Such concepts (see [3, 36, 39, 47]) require the nonemptiness of the ordering cone, which defines the cone constraints in convex optimization and variational inequalities. However, the ordering cone of almost all the known problems, stated in infinite dimensional spaces, has the interior (and all the above generalized interior concepts) empty. Hence, the above interior conditions cannot be used to guarantee the strong duality. This is the case, for example, of optimization problems or variational inequalities connected with evolutionary financial network equilibrium problems, the obstacle problem, the elastic-plastic torsion problem, the infinite-dimensional bilevel problem, which use non-negative cones of Lebesgue or Sobolev spaces (see [1, 8, 10–14, 22, 24, 25, 27, 31, 34, 35, 37, 46, 49]).

Only recently, in [16, 42, 43] the authors introduced new conditions called S, S', NES, which turn out to be necessary and sufficient conditions for the strong duality and really useful in the applications. These conditions do not require the nonemptiness of the interior of the ordering cone. This new strong duality theory was then refined in [13, 17, 19, 37, 45].

Now we present in detail these new conditions.

1.1 Assumption S

Let us first recall that for a subset $C \subseteq X$ and $x \in X$ the tangent cone to C at x is defined as

$$T_C(x) = \{y \in X :\ y = \lim_{n\to\infty} \lambda_n(x_n - x),\ \lambda_n > 0,\ x_n \in C,\ \lim_{n\to\infty} x_n = x\}.$$

If $x \in clC$ (the closure of C) and C is convex, we have

$$T_C(x) = clcone(C - \{x\}),$$

where the $coneA = \{\lambda x : \ x \in A, \ \lambda \in \mathbb{R}^+\}$ denotes the cone hull of a general subset A of the space.

Definition 1 Given the mappings f, g, h and the set K as above, we say that *Assumption S* is fulfilled at a point $x_0 \in K$ if and only if

$$T_{\widetilde{M}}(0, \theta_Y, \theta_Z) \cap (\mathbb{R}^{--} \times \theta_Y \times \theta_Z) = \emptyset$$

where

$$\widetilde{M} = \{(f(x) - f(x_0) + \alpha, g(x) + y, h(x)) : \ x \in S \setminus K, \ \alpha \geq 0, \ y \in C\},$$

$$\mathbb{R}^{--} = \{\lambda \in \mathbb{R} : \lambda < 0\}.$$

Now we recall the main theorem on strong duality based on *Assumption S* (see [13, 16, 17, 19, 45]).

Theorem 1 *Assume that the functions $f : S \longrightarrow \mathbb{R}$, $g : S \longrightarrow Y$ are convex and that $h : S \longrightarrow Z$ is an affine-linear mapping. Assume that the* Assumption *S is fulfilled at the optimal solution $x_0 \in K$ of the problem (1). Then also problem (2) is solvable and if $\overline{u} \in C^*$, $\overline{v} \in Z^*$ are optimal solutions to (2), we have*

$$\langle \overline{u}, g(x_0) \rangle = 0 \tag{3}$$

and the optimal values of the two problems coincide; namely

$$f(x_0) = \min_{x \in \mathbb{K}} f(x) = f(x_0) + \langle \bar{u}, g(x_0) \rangle + \langle \bar{v}, h(x_0) \rangle$$

$$= \max_{\substack{u \in C^* \\ v \in Z^*}} \inf_{x \in S} \left[f(x) + \langle u, g(x) \rangle + \langle v, h(x) \rangle \right].$$

Moreover it is seen in [4] that *Assumption S* is also a necessary condition for the strong duality.

An important consequence of the strong duality is the usual relationship between a saddle point of the so-called Lagrange functional

$$L(x, u, v) = f(x) + \langle u, g(x) \rangle + \langle v, h(x) \rangle, \quad \forall x \in S, \ \forall u \in C^*, \ \forall v \in Z^*,$$

and the solution to (1) and (2). Indeed, we have the following theorem (see [16] and [23]).

Theorem 2 *Let the assumptions of Theorem 1 be fulfilled. Then, $x_0 \in \mathbb{K}$ is an optimal solution to (1) if and only if there exist $\bar{u} \in C^*$, $\bar{v} \in Z^*$ such that $(x_0, \bar{u}, \bar{v})$ is a saddle point of the Lagrange functional, namely:*

$$L(x_0, u, v) \leq L(x_0, \bar{u}, \bar{v}) \leq L(x, \bar{u}, \bar{v}), \quad \forall x \in S, \ \forall u \in C^*, \ \forall v \in Z^*$$

and

$$\langle \bar{u}, g(x_0) \rangle = 0.$$

1.2 Assumption S′

Assumption S′ requires additional hypotheses on the mappings f, g, h and works on directional derivatives. Sometimes it is easier to use with respect to Assumption S.

Let us assume that f, g, h have directional derivative at $x_0 \in K$ in every direction $x - x_0$ with arbitrary $x \in S$.

Definition 2 We say that *Assumption S′* is fulfilled at the point $x_0 \in K$ if and only if

$$T_{M'}(0, \theta_Y, \theta_Z) \cap (\mathbb{R}^{--} \times \{\theta_Y\} \times \{\theta_Z\}) = \emptyset,$$

where

$$M' = \{(f'(x_0)(x - x_0) + \alpha, g(x_0) + g'(x_0)(x - x_0) + y, h'(x_0)(x - x_0)) :$$

$$x \in S \setminus K, \ \alpha \geq 0, \ y \in C\}.$$

The next theorem holds (see [42]).

Theorem 3 *Let X and Z be real normed spaces, let Y be a real normed space ordered by a closed convex cone C. Let S be a convex subset of X and let $f : S \longrightarrow \mathbb{R}$ be a given convex functional, let $g : S \longrightarrow Y$ be a convex mapping and let $h : S \longrightarrow Z$ an affine-linear mapping. Assume that f, g have a directional derivative at $x_0 \in K$ solution to problem (1) in every direction $x - x_0$ with arbitrary $x \in S$. Then, the strong duality holds if and only if* Assumption *S' is fulfilled.*

1.3 Strong Duality in the Case of Nonlinear Equality Constrains

Let us assume that h is no longer an affine-linear mapping, but, for instance, a convex one, since it depends on the sign of v. Then the constraint set $\mathbb{K}$ is no longer convex. As a consequence, the usual optimality conditions for the convex functions on convex sets cannot be applied. Moreover, if we consider the Lagrange functional

$$L(x, \bar{u}, \bar{v}) = f(x) + \langle \bar{u}, g(x)\rangle + \langle \bar{v}, h(x)\rangle,$$

where $\bar{u} \in C^*$ and $\bar{v} \in Z^*$, even if h is convex as well as g, L is not, in general, a convex functional. In order to overcome these difficulties, some strong duality results have been elaborated under Assumption S′, but introducing additional conditions (see [44] Theorem 2).

Theorem 4 *Let $(X, \|\cdot\|_X)$, $(Y, \|\cdot\|_Y)$, $(Z, \|\cdot\|_Z)$ be real Banach spaces with an ordering closed convex cone X^+, C and D, respectively. Let S be an open convex subset of X, let $f : S \to \mathbb{R}$ be a convex functional such that f is Fréchet-differentiable at a minimal point $x_0 \in \mathbb{K}$, let $g : S \to Y$ and $h : S \to Z$ be convex functions with respect to the cones C and D, respectively, both Gâteaux-differentiable at x_0. Assume also:*

(i) $f'(x_0)(x) \leq 0, \forall x \in S \cap X^+$;
(ii) $g'(x_0)(x) \in -C \ \forall x \in S \cap X^+$;
(iii) $h'(x_0)(S \cap X^+) = D$;
(iv) $\lim\limits_{\substack{\|\lambda(x-x_0)\|_X \to +\infty \\ x\in\mathbb{K}}} \|h'(x_0)(\lambda(x - x_0)\|_Z = +\infty.$

If Assumption S′ *is fulfilled at x_0, then the strong duality holds.*

Example 1 Let us consider the following problem (see [44]):

$$\min_{u\in\mathbb{K}} \int_0^T \left(u_2^2(t) - u_1(t)\right) dt,$$

where

$$\mathbb{K} = \left\{u \in L^2([0, T], \mathbb{R}^2) : u(t) \geq 0 \text{ and } u_1^2(t) + u_2^2(t) = 1 \text{ a.e.}\right\}.$$

We set:

$$\begin{array}{lll} f : L^2([0, T], \mathbb{R}^2) \to \mathbb{R} & \text{defined as} & f(u) = \displaystyle\int_0^T \left(u_2^2(t) - u_1(t)\right) dt; \\ g : L^2([0, T], \mathbb{R}^2) \to L^2([0, T], \mathbb{R}^2) & \text{defined as} & g(u) = -u; \\ h : L^2([0, T], \mathbb{R}^2) \to L^1([0, T], \mathbb{R}) & \text{defined as} & h(u) = u_1^2(t) + u_2^2(t) - 1. \end{array}$$

We note that f attains its minimum value in $\mathbb{K}$ in correspondence of the couple of constant functions $u_0 = (1, 0)$. Now, we verify that all assumptions of Theorem 4 are satisfied. Indeed:

(i) $f'(u_0)(u) = \displaystyle\int_0^T -u_1(t)\, dt \leq 0, \forall (u_1, u_2) \in L^2([0, T], \mathbb{R}^2_+)$;

(ii) $g'(u_0)(u) = (-u_1, -u_2) \in -C$, $\forall (u_1, u_2) \in L^2([0, T], \mathbb{R}^2_+)$;
(iii) $h'(u_0)(u) = 2u_1 \geq 0$, $\forall (u_1, u_2) \in L^2([0, T], \mathbb{R}^2_+)$;
(iv) $\lim\limits_{\substack{\|\lambda(u-u_0)\|_{L^2}\to+\infty \\ u\in\mathbb{K}}} \|h'(u_0)(\lambda(u-u_0)\|_{L^1} = \lim\limits_{\|\lambda(u_1-1,u_2)\|_{L^2}\to+\infty} \|2\lambda(u_1-1)\|_{L^1} = +\infty$.

In order to have the strong duality, it remains to prove that also Assumption S′ holds true. Let

$$(\lambda, \theta_{L^2([0,T],\mathbb{R}^2}, 0) \in T_{M'}(\lambda, \theta_{L^2([0,T],\mathbb{R}^2}, 0).$$

We need to verify that:

$$\lambda = \lim_n \lambda_n \left(\int_0^T f'(u_0)(u_n - u_0)\, dt + \alpha_n \right) = \lim_n \lambda_n \left(\int_0^T -(u_1^n - 1)\, dt + \alpha_n \right) \geq 0,$$

taking into account that:

$$\theta_{L^2([0,T],\mathbb{R}^2)} = \lim_n \lambda_n \left(\varphi(u_0) + \varphi'(u_0)(u_n - u_0) + v_n \right)$$

$$= \lim_n \lambda_n \left(-1 - (u_1^n - 1) + v_1^n, -u_2^n + v_2^n \right) = 0,$$

and

$$\theta_{L^2([0,T],\mathbb{R}^2)} = \lim_n \lambda_n \left(2(u_1^n - 1) \right) = 0, \tag{4}$$

where $\lambda_n \geq 0$, $n \in \mathbb{N}$, $v_n \in L^2([0, T], \mathbb{R}^2)^+$, $u_n \in L^2([0, T], \mathbb{R}^2) \setminus \mathbb{K}$, $\alpha_n \geq 0$, $\forall n \in \mathbb{N}$.
From (4) it follows:

$$\lambda = \lim_n \lambda_n \left(\int_0^T (-u_1^n - 1)\, dt + \alpha_n \right) \geq 0.$$

Example 2 Now, we present an example where assumption (i) is not satisfied (see [44]).
Let us consider the problem:

$$\min_{u\in\mathbb{K}} \int_0^1 \left(\frac{1}{2} u_2^2(t) + u_1(t) \right) dt$$

where

$$\mathbb{K} = \left\{ u \in L^2([0, 1], \mathbb{R}^2) : u(t) \geq 0 \text{ and } u_1^2(t) + u_2^2(t) = 1, \text{ a.e. in } [0, 1] \right\}.$$

We set:

$$\begin{array}{lll} f: L^2([0,1],\mathbb{R}^2) \to \mathbb{R} & \text{defined as} & f(u) = \displaystyle\int_0^1 \left(\frac{1}{2}u_2^2(t) + u_1(t)\right) dt; \\ g: L^2([0,1],\mathbb{R}^2) \to L^2([0,1],\mathbb{R}^2) & \text{defined as} & g(u) = -u; \\ h: L^2([0,1],\mathbb{R}^2) \to L^1([0,1],\mathbb{R}) & \text{defined as} & h(u) = u_1^2(t) + u_2^2(t) - 1. \end{array}$$

We note that f attains its minimum value in $\mathbb{K}$ in correspondence of the couple of constant functions $u_0 = (1,0)$. Assumption (i) of Theorem 4 is not satisfied. Indeed:

(i) $f'(u_0)(u) = \displaystyle\int_0^1 u_1(t)\,dt \geq 0,\ \forall (u_1,u_2) \in L^2([0,1],\mathbb{R}^2_+)$;

(ii) $g'(u_0)(u) = (-u_1,-u_2) \in -C,\ \forall (u_1,u_2) \in L^2([0,1],\mathbb{R}^2_+)$;

(iii) $h'(u_0)(u) = 2u_1 \geq 0,\ \forall (u_1,u_2) \in L^2([0,1],\mathbb{R}^2_+)$;

(iv) $\displaystyle\lim_{\substack{\|\lambda(u-u_0)\|_{L^2}\to+\infty \\ u\in\mathbb{K}}} \|h'(u_0)(\lambda(u-u_0)\|_{L^1} = +\infty.$

As in the previous example, it is easy to show that also *Assumption S'* holds true. Since we have (see also formula (2.5) in [44]):

$$\big(f'(u_0) + \langle \bar{u}, g(u_0) + g'(u_0)\rangle + \langle \bar{v}, h'(u_0)\rangle\big)\, u = 0 \quad \forall u \in L^2([0,1],\mathbb{R}^2) \tag{5}$$

and (3) holds true, it follows, from an easy calculation, that the maximum over C^+ and $v \in Z^*$ is achieved when

$$u = (0, \bar{u}_2) \text{ and } v = -\frac{1}{2},$$

for some $\bar{u}_2 \geq 0$. Therefore,

$$\max_{\substack{u\in C^* \\ v\in Z^*}} \inf_{x\in L^2} [f(x) + \langle u, g(x)\rangle + \langle v, h(x)\rangle]$$

$$= \inf_{u\in L^2}\left[f(u) + \langle (0,\bar{u}_2), (-u_1,-u_2)\rangle + \langle -\frac{1}{2}, u_1^2 + u_2^2 - 1\rangle\right]$$

$$= \inf_{u\in L^2}\left[\int_0^1 \left(-\frac{1}{2}u_1^2(t) + u_1(t) - \bar{u}_2(t)u_2(t) + \frac{1}{2}\right) dt\right].$$

As we have seen, not all the assumptions of Theorem 4 are fulfilled. Hence, if strong duality holds, then we would have:

$$1 \leq \int_0^1 \left(-\frac{1}{2}u_1^2(t) + u_1(t) - \bar{u}_2(t)u_2(t) + \frac{1}{2}\right) dt \quad \forall (u_1,u_2) \in L^2([0,1],\mathbb{R}^2).$$

It is enough to choose, for instance, $(u_1, u_2) = (2, 0)$ to get a contradiction, since:

$$\int_0^1 \left(-\frac{1}{2}4 + 2 + \frac{1}{2}\right) dt = \int_0^1 \frac{1}{2}\, dt = \frac{1}{2}.$$

1.4 NES (Non Empty Subdifferential Condition)

This new necessary and sufficient condition is the one which requires a smaller number of assumptions on the functions. Recall that a subdifferential of a convex function $f : S \to \mathbb{R}$, where S is a subset of a real normed space X, at $\overline{x} \in S$ is defined by

$$\partial f(\overline{x}) = \{x^* \in X^* \ : \ f(x) - f(\overline{x}) \geq \langle x^*, x - \overline{x}\rangle, \forall x \in S\}.$$

For $y \in Y$, let us define a closed convex subset of Y as

$$D_y = (y - C)$$

with C the closed convex ordering cone of Y.

If on $Y \times Z$, we consider the norm $\|(y, z)\|_{Y \times Z\text{ł}} = \|y\|_Y + \|z\|_Z$, let us define

$$\varphi : Y \times Z \to \overline{\mathbb{R}}$$

by

$$\varphi(y, z) = \inf_{\substack{x \in S \\ g(x) \in D_y \\ h(x) = z}} f(x).$$

Definition 3 (Assumption NES) We say that the *Condition NES* is fulfilled for the triple f, g, h with respect to K if and only if

$$\partial\varphi(\theta_{Y \times Z}) \neq \emptyset. \tag{6}$$

Taking into account that $\varphi(\theta_{Y\times Z}) = \inf\limits_{\substack{x \in S \\ -g(x) \in C \\ h(x) = \theta_Z}} f(x) = \inf\limits_{x \in \mathbb{K}} f(x)$, (6) means that there exist $(y^*, z^*) \in Y^* \times Z^*$ such that

$$\varphi(y, z) - \varphi(\theta_{Y \times Z}) = \inf_{\substack{x \in S \\ g(x) \in D_y \\ h(x) = z}} f(x) - \inf_{x \in \mathbb{K}} f(x) \geq \langle y^*, y\rangle + \langle z^*, z\rangle, \quad \forall (y, z) \in Y \times Z.$$

Then, we have the following result (see [43] Theorem 3.2).

Theorem 5 *Let us assume that* $\inf_{x \in \mathbb{K}} f(x) \in \mathbb{R}$. *Then, the strong duality holds for problems* (1) *and* (2) *if and only if the Condition NES holds for* f, g, h.

Now, the following result easily follows.

Corollary 1 *Assume that* $f : S \to \mathbb{R}$, $g : S \to Y$ *are convex functions and let* $h : S \to Z$ *be an affine-linear mapping. Then* Assumption S *is fulfilled at the optimal solution* $x_0 \in \mathbb{K}$ *of problem* (1), *if and only if* Condition NES *holds for* f, g, *and* h *with respect to* $\mathbb{K}$.

Similarly, keeping in consideration the main result in [42], we have the following result.

Corollary 2 *Let* X *and* Z *be real normed spaces, let* Y *be a real normed space ordered by a closed convex cone* C. *Let* S *be a convex subset of* X *and let* $f : S \to \mathbb{R}$ *be a given convex functional, let* $g : S \to Y$ *be a convex mapping and let* $h : S \to Z$ *an affine-linear mapping. Assume that* f, g *have a directional derivative at* $x_0 \in K$ *solution to problem* (1) *in every direction* $x - x_0$ *with arbitrary* $x \in S$. *Then,* Assumption S′ *is fulfilled at* x_0 *if and only if* Condition NES *holds for* f, g, h *with respect to* $\mathbb{K}$.

As for *Assumptions S* and *S′*, also Condition NES is really useful in the applications as we can see in the next sections.

2 Applications to the General Financial Equilibrium Problem

In this chapter we apply *Assumption S*, which was introduced in Sect. 1.1, to a general equilibrium model of financial flows and prices (see also [15]).

2.1 Presentation of the Model

We consider a financial economy consisting of m sectors, for example households, domestic business, banks and other financial institutions, as well as state and local governments, with a typical sector denoted by i, and of n instruments, for example mortgages, mutual funds, saving deposits, money market funds, with a typical financial instrument denoted by j, in the time interval $[0, T]$. Let $s_i(t)$ denote the total financial volume held by sector i at time t as assets, and let $l_i(t)$ be the total financial volume held by sector i at time t as liabilities. Further, we allow markets of assets and liabilities to have different investments $s_i(t)$ and $l_i(t)$, respectively. Since we are working in the presence of uncertainty and of risk perspectives, the volumes $s_i(t)$ and $l_i(t)$ held by each sector cannot be considered stable with respect to time and may decrease or increase. For instance, depending on the crisis periods, a sector may decide not to invest on instruments and to buy goods as gold and silver. At time t, we denote the amount of instrument j held as an asset in sector i's portfolio by $x_{ij}(t)$ and the amount of instrument j held as a liability in sector i's portfolio by $y_{ij}(t)$. The assets and liabilities in all the sectors are grouped into the matrices $x(t),\ y(t) \in \mathbb{R}^{m\times n}$, respectively. At time t we denote the price of instrument j held as an asset and as a liability by $r_j(t)$ and by $(1 + h_j(t))r_j(t)$, respectively, where h_j is a nonnegative function defined into $[0, T]$ and belonging to $L^\infty([0, T], \mathbb{R})$. We introduce the term $h_j(t)$ because the prices of liabilities are generally greater than or equal to the prices of assets. In this manner we describe, in a more realistic way, the behaviour of the markets for which the liabilities are more expensive than the assets. We group the instrument prices held as an asset nd as a liability into the vectors $r(t) = [r_1(t), r_2(t), \dots, r_i(t), \dots, r_n(t)]^T$ and $(1 + h(t))r(t) = [(1 + h_1(t))r_1(t), (1 + h_2(t))r_2(t), \dots, (1 + h_i(t))r_i(t), \dots, (1 + h_n(t))r_n(t)]^T$, respectively. In our problem the prices of each instrument appear as unknown variables. Under the assumption of perfect competition, each sector will behave as if it has no influence on the instrument prices or on the behaviour of the other sectors, but on the total amount of the investments and the liabilities of each sector.

We choose as a functional setting the very general Lebesgue space

$$L^2([0, T], \mathbb{R}^p) = \left\{ f : [0, T] \to \mathbb{R}^p \text{ measurable} : \int_0^T \|f(t)\|_p^2 dt < +\infty \right\},$$

with the norm

$$\|f\|_{L^2([0,T],\mathbb{R}^p)} = \left(\int_0^T \|f(t)\|_p^2 dt \right)^{\frac{1}{2}}.$$

Then, the set of feasible assets and liabilities for each sector $i = 1, \dots, m$ becomes

$$P_i = \Bigg\{ (x_i(t), y_i(t)) \in L^2([0, T], \mathbb{R}^{2n}_+) : \\ \sum_{j=1}^n x_{ij}(t) = s_i(t), \quad \sum_{j=1}^n y_{ij}(t) = l_i(t) \text{ a.e. in } [0, T] \Bigg\}$$

and the set of all feasible assets and liabilities becomes

$$P = \left\{ (x(t), y(t)) \in L^2([0, T], \mathbb{R}^{2mn}) : (x_i(t), y_i(t)) \in P_i,\ i = 1, \dots, m \right\}.$$

Now, we introduce the ceiling and the floor price associated with instrument j, denoted by $\overline{r}_j$ and by $\underline{r}_j$, respectively, with $\overline{r}_j(t) > \underline{r}_j(t) \geq 0$, a.e. in $[0, T]$. The floor price $\underline{r}_j(t)$ is determined on the basis of the official interest rate fixed by the central banks, which, in turn, take into account the consumer price inflation. Then the equilibrium prices $r_j^*(t)$ cannot be less than these floor prices. The ceiling price

$\overline{r}_j(t)$ derives from the financial need to control the national debt arising from the amount of public bonds and of the rise in inflation. It is a sign of the difficulty on the recovery of the economy. However it should be not overestimated because it produced an availability of money.

In detail, the meaning of the lower and upper bounds is that to each investor a minimal price $\underline{r}_j$ for the assets held in the instrument j is guaranteed, whereas each investor is requested to pay for the liabilities in any case a minimal price $(1+h_j)\underline{r}_j$. Analogously each investor cannot obtain for an asset a price greater than $\overline{r}_j$ and as a liability the price cannot exceed the maximum price $(1+h_j)\overline{r}_j$.

We denote the given tax rate levied on sector i's net yield on financial instrument j, as τ_{ij}. Assume that the tax rates lie in the interval $[0, 1)$ and belong to $L^\infty([0, T], \mathbb{R})$. Therefore, the government in this model has the flexibility of levying a distinct tax rate across both sectors and instruments.

We group the instrument ceiling and floor prices into the column vectors $\overline{r}_j(t) = (\overline{r}_n(t))_{j=1,\dots,n}$, and $\underline{r}_j(t) = (\underline{r}_j(t))_{j=1,\dots,n}$, respectively, and the tax rates τ_{ij} into the matrix $\tau(t) \in L^2([0, T], \mathbb{R}^{m\times n})$.

The set of feasible instrument prices is:

$$\mathscr{R} = \{r \in L^2([0, T], \mathbb{R}^n) : \underline{r}_j(t) \le r_j(t) \le \overline{r}_j(t), \quad j = 1, \dots, n, \text{ a.e. in } [0, T]\},$$

where $\underline{r}$ and $\overline{r}$ are assumed to belong to $L^2([0, T], \mathbb{R}^n)$.

In order to determine for each sector i the optimal distribution of instruments held as assets and as liabilities, we consider, as usual, the influence due to risk-aversion and the optimality conditions of each sector in the financial economy, namely the desire to maximize the value of the asset holdings while minimizing the value of liabilities. An example of risk aversion is given by the well-known Markowitz quadratic function based on the variance-covariance matrix denoting the sector's assessment of the standard deviation of prices for each instrument (see [40, 41]). In our case, however, the Markowitz utility or other more general ones are assumed to be time-dependent in order to incorporate the adjustment in time which depends on the previous equilibrium states.

Then, we introduce the utility function $U_i(t, x_i(t), y_i(t), r(t))$, for each sector i, defined as follows:

$$U_i(t, x_i(t), y_i(t), r(t)) = u_i(t, x_i(t), y_i(t))$$

$$+\sum_{j=1}^{n} r_j(t)(1 - \tau_{ij}(t))[x_{ij}(t) - (1 + h_j(t))y_{ij}(t)],$$

where the term $-u_i(t, x_i(t), y_i(t))$ represents a measure of the risk of the financial agent and $r_j(t)(1 - \tau_{ij}(t))[x_i(t) - (1 + h_j(t))y_i(t)]$ represents the value of the difference between the asset holdings and the value of liabilities. We suppose that the sector's utility function $U_i(t, x_i(t), y_i(t))$ is defined on $[0, T] \times \mathbb{R}^n \times \mathbb{R}^n$, is measurable in t and is continuous with respect to x_i and y_i. Moreover we assume that $\dfrac{\partial u_i}{\partial x_{ij}}$ and $\dfrac{\partial u_i}{\partial y_{ij}}$ exist and that they are measurable in t and continuous with respect to x_i and y_i. Further, we require that $\forall i = 1, \ldots, m$, $\forall j = 1, \ldots, n$, and a.e. in $[0, T]$ the following growth conditions hold true:

$$|u_i(t, x, y)| \leq \alpha_i(t)\|x\|\|y\|, \quad \forall x, y \in \mathbb{R}^n, \tag{7}$$

and

$$\left|\frac{\partial u_i(t, x, y)}{\partial x_{ij}}\right| \leq \beta_{ij}(t)\|y\|, \quad \left|\frac{\partial u_i(t, x, y)}{\partial y_{ij}}\right| \leq \gamma_{ij}(t)\|x\|, \tag{8}$$

where α_i, β_{ij}, γ_{ij} are non-negative functions of $L^\infty([0, T], \mathbb{R})$. Finally, we suppose that the function $u_i(t, x, y)$ is concave.

In Sect. 2.5 we define a utility function of Markowitz type.

Now, we establish the equilibrium conditions for the prices which express the equilibration of the total assets, the total liabilities and the portion of financial transactions per unit F_j employed to cover the expenses of the financial institutions including possible dividends and manager bonus. Indeed, the equilibrium condition for the price r_j of instrument j is the following:

$$\sum_{i=1}^{m}(1 - \tau_{ij}(t))\left[x^*_{ij}(t) - (1 + h_j(t))y^*_{ij}(t)\right] + F_j(t)$$
$$\begin{cases} \geq 0 \text{ if } r^*_j(t) = \underline{r}_j(t) \\ = 0 \text{ if } \underline{r}_j(t) < r^*_j(t) < \overline{r}_j(t) \\ \leq 0 \text{ if } r^*_j(t) = \overline{r}_j(t) \end{cases} \tag{9}$$

where (x^*, y^*, r^*) is the equilibrium solution for the investments as assets and as liabilities and for the prices. In other words, the prices are determined taking into account the amount of the supply, the demand of an instrument and the charges F_j, namely if there is an actual supply excess of an instrument as assets and of the charges F_j in the economy, then its price must be the floor price. If the price of an instrument is positive, but not at the ceiling, then the market of that instrument must clear. Finally, if there is an actual demand excess of an instrument as liabilities in the economy, then the price must be at the ceiling.

Now, we can give different but equivalent equilibrium conditions, each of which is useful to illustrate particular features of the equilibrium.

Definition 4 A vector of sector assets, liabilities and instrument prices $(x^*(t), y^*(t), r^*(t)) \in P \times \mathscr{R}$ is an equilibrium of the dynamic financial model if and only if $\forall i = 1, \ldots, m$, $\forall j = 1, \ldots, n$, and a.e. in $[0, T]$, it satisfies the system of inequalities

$$-\frac{\partial u_i(t,x^*,y^*)}{\partial x_{ij}} - (1-\tau_{ij}(t))r_j^*(t) - \mu_i^{(1)*}(t) \geq 0, \tag{10}$$

$$-\frac{\partial u_i(t,x^*,y^*)}{\partial y_{ij}} + (1-\tau_{ij}(t))(1+h_j(t))r_j^*(t) - \mu_i^{(2)*}(t) \geq 0, \tag{11}$$

and equalities

$$x_{ij}^*(t)\Big[-\frac{\partial u_i(t,x^*,y^*)}{\partial x_{ij}} - (1-\tau_{ij}(t))r_j^*(t) - \mu_i^{(1)*}(t)\Big] = 0, \tag{12}$$

$$y_{ij}^*(t)\Big[-\frac{\partial u_i(t,x^*,y^*)}{\partial x_{ij}} + (1-\tau_{ij}(t))(1+h_j(t))r_j^*(t) - \mu_i^{(2)*}(t)\Big] = 0, \tag{13}$$

where $\mu_i^{(1)*}(t)$, $\mu_i^{(2)*}(t) \in L^2([0,T],\mathbb{R})$ are Lagrange multipliers, and verifies conditions (9) a.e. in $[0,T]$.

We associate with each financial volumes s_i and l_i held by sector i the functions $\mu_i^{(1)*}(t)$ and $\mu_i^{(2)*}(t)$, related, respectively, to the assets and to the liabilities and which represent the "equilibrium disutilities" per unit of sector i. Then, (10) and (12) mean that the financial volume invested in instrument j as assets x_{ij}^* is greater than or equal to zero if the j-th component $-\frac{\partial u_i(t,x^*,y^*)}{\partial x_{ij}} - (1-\tau_{ij}(t))r_j^*(t)$ of the disutility is equal to $\mu_i^{(1)*}(t)$, whereas if $-\frac{\partial u_i(t,x^*,y^*)}{\partial x_{ij}} - (1-\tau_{ij}(t))r_j^*(t) > \mu_i^{(1)*}(t)$, then $x_{ij}^*(t) = 0$. The same occurs for the liabilities.

The functions $\mu_i^{(1)*}(t)$ and $\mu_i^{(2)*}(t)$ are the Lagrange multipliers associated a.e. in $[0,T]$ with the constraints $\sum_{j=1}^{n} x_{ij}(t) - s_i(t) = 0$ and $\sum_{j=1}^{n} y_{ij}(t) - l_i(t) = 0$, respectively. They are unknown a priori, but this fact has no influence because we will prove in the following theorem that Definition 4 is equivalent to a variational inequality in which $\mu_i^{(1)*}(t)$ and $\mu_i^{(2)*}(t)$ do not appear (see [2] Theorem 2.1.).

Theorem 6 *A vector $(x^*, y^*, r^*) \in P \times \mathscr{R}$ is a dynamic financial equilibrium if and only if it satisfies the following variational inequality:*

Find $(x^*, y^*, r^*) \in P \times \mathscr{R}$:

$$\sum_{i=1}^{m}\int_0^T \Bigg\{\sum_{j=1}^{n}\Big[-\frac{\partial u_i(t,x_i^*(t),y_i^*(t))}{\partial x_{ij}} - (1-\tau_{ij}(t))r_j^*(t)\Big]$$
$$\times[x_{ij}(t) - x_{ij}^*(t)]$$

$$+\sum_{j=1}^{n}\Big[-\frac{\partial u_i(t,x_i^*(t),y_i^*(t))}{\partial y_{ij}}+(1-\tau_{ij}(t))r_j^*(t)(1+h_j(t))\Big]$$

$$\times[y_{ij}(t)-y_{ij}^*(t)]\Big\}dt$$

$$+\sum_{j=1}^{n}\int_0^T\sum_{i=1}^{m}\Big\{(1-\tau_{ij}(t))\Big[x_{ij}^*(t)-(1+h_j(t))y_{ij}^*(t)\Big]+F_j(t)\Big\}$$

$$\times\big[r_j(t)-r_j^*(t)\big]dt\geq 0,\qquad \forall(x,y,r)\in P\times\mathscr{R}. \tag{14}$$

Remark 1 We would like to explicitly remark that our definition of equilibrium conditions (Definition 4) is equivalent to the equilibrium definition given by a vector $(x^*, y^*, r^*)\in P\times\mathscr{R}$ satisfying

$$\max_{P_i}\int_0^T\Big\{u_i(t,x_i(t),y_i(t))+(1-\tau_i(t))r^*(t)\times[x_i(t)-(1+h(t))y_i(t)]\Big\}dt,$$

$\forall(x_i, y_i)\in P_i$, and (9). We prefer to use Definition 4, since it is expressed in terms of equilibrium disutilities.

Now, we would like to give an existence result. First of all, we remind some definitions. Let X be a reflexive Banach space and let $\mathbb{K}$ be a subset of X and X^* be the dual space of X.

Definition 5 A mapping $A:\mathbb{K}\to X^*$ is pseudomonotone in the sense of Brezis (B-pseudomonotone) iff

1. For each sequence u_n weakly converging to u (in short $u_n\rightharpoonup u$) in $\mathbb{K}$ and such that $\limsup_n\langle Au_n, u_n-v\rangle\leq 0$ it results that:
$$\liminf_n\langle Au_n, u_n-v\rangle\geq\langle Au, u-v\rangle,\quad \forall v\in\mathbb{K}.$$
2. For each $v\in\mathbb{K}$ the function $u\mapsto\langle Au, u-v\rangle$ is lower bounded on the bounded subset of $\mathbb{K}$.

Definition 6 A mapping $A:\mathbb{K}\to X^*$ is hemicontinuous in the sense of Fan (F-hemicontinuous) iff for all $v\in\mathbb{K}$ the function $u\mapsto\langle Au, u-v\rangle$ is weakly lower semicontinuous on $\mathbb{K}$.

The following existence result does not require any kind of monotonicity assumptions.

Theorem 7 *Let $\mathbb{K}\subset X$ be a nonempty closed convex bounded set and let $A:\mathbb{K}\subset E\to X^*$ be B-pseudomonotone or F-hemicontinuous. Then the variational inequality*

$$\langle Au, v-u\rangle\geq 0\quad\forall v\in\mathbb{K} \tag{15}$$

admits a solution.

2.2 The Duality for the Financial Equilibrium Problem

Now, in order to study the duality for the financial equilibrium problem, let us set:

$$\begin{aligned}
f(x,y,r) = \int_0^T \Bigg\{ & \sum_{i=1}^m \sum_{j=1}^n \left[-\frac{\partial u_i(t, x^*(t), y^*(t))}{\partial x_{ij}} - (1-\tau_{ij}(t)) r_j^*(t) \right] \\
& \times [x_{ij}(t) - x_{ij}^*(t)] \\
& + \sum_{i=1}^m \sum_{j=1}^n \left[-\frac{\partial u_i(t, x^*(t), y^*(t))}{\partial y_{ij}} + (1-\tau_{ij}(t))(1+h_j(t)) r_j^*(t) \right] \\
& \times [y_{ij}(t) - y_{ij}^*(t)] \\
& + \sum_{j=1}^n \left[\sum_{i=1}^m (1-\tau_{ij}(t)) \left[x_{ij}^*(t) - (1+h_j(t)) y_{ij}^*(t) \right] + F_j(t) \right] \\
& \times \left[r_j(t) - r_j^*(t) \right] \Bigg\} \, dt.
\end{aligned}$$

Then the Lagrange functional is

$$\begin{aligned}
& \mathscr{L}(x, y, r, \lambda^{(1)}, \lambda^{(2)}, \mu^{(1)}, \mu^{(2)}, \rho^{(1)}, \rho^{(2)}) = f(x,y,r) \\
& - \sum_{i=1}^m \sum_{j=1}^n \int_0^T \lambda_{ij}^{(1)}(t) x_{ij}(t)\, dt - \sum_{i=1}^m \sum_{j=1}^n \int_0^T \lambda_{ij}^{(2)} y_{ij}(t)\, dt \\
& - \sum_{i=1}^m \int_0^T \mu_i^{(1)}(t) \left(\sum_{j=1}^n x_{ij}(t) - s_i(t) \right) dt \qquad (16) \\
& - \sum_{i=1}^m \int_0^T \mu_i^{(2)}(t) \left(\sum_{j=1}^n y_{ij}(t) - l_i(t) \right) dt \\
& + \sum_{j=1}^n \int_0^T \rho_j^{(1)}(t)(\underline{r}_j(t) - r_j(t))\, dt + \sum_{j=1}^n \int_0^T \rho_j^{(2)}(t)(r_j(t) - \overline{r}_j(t))\, dt,
\end{aligned}$$

where $(x, y, r) \in L^2([0,T], \mathbb{R}^{2mn+n})$, $\lambda^{(1)}$, $\lambda^{(2)} \in L^2([0,T], \mathbb{R}_+^{mn})$, $\mu^{(1)}$, $\mu^{(2)} \in L^2([0,T],\ \mathbb{R}^m)$, $\rho^{(1)}, \rho^{(2)} \in L^2([0,T], \mathbb{R}_+^n)$ and $\lambda^{(1)}$, $\lambda^{(2)}$, $\rho^{(1)}$, $\rho^{(2)}$ are the Lagrange multipliers associated, a.e. in $[0,T]$, with the sign constraints $x_i(t) \geq 0$, $y_i(t) \geq 0$, $r_j(t) - \underline{r}_j(t) \geq 0$, $\overline{r}_j(t) - r_j(t) \geq 0$, respectively whereas the functions $\mu^{(1)}(t)$ and $\mu^{(2)}(t)$ are the Lagrange multipliers associated, a.e. in $[0,T]$, with the equality constraints $\sum_{j=1}^{n} x_{ij}(t) - s_i(t) = 0$ and $\sum_{j=1}^{n} y_{ij}(t) - l_i(t) = 0$, respectively. Hence, applying Theorem 1, the following result can be provided (see [2]):

Theorem 8 *Let* $(x^*, y^*, r^*) \in P \times \mathscr{R}$ *be a solution to variational inequality* (14) *and let us consider the associated Lagrange functional* (16). *Then, the strong duality holds and there exist* $\lambda^{(1)*}, \lambda^{(2)*} \in L^2([0,T], \mathbb{R}^{mn}_+)$, $\mu^{(1)*}, \mu^{(2)*} \in L^2([0,T], \mathbb{R}^m)$, $\rho^{(1)*}, \rho^{(2)*} \in L^2([0,T], \mathbb{R}^n_+)$ *such that* $(x^*, y^*, r^*, \lambda^{(1)*}, \lambda^{(2)*}, \mu^{(1)*}, \mu^{(2)*}, \rho^{(1)*}, \rho^{(2)*})$ *is a saddle point of the Lagrange functional, namely*

$$\begin{aligned}&\mathscr{L}(x^*, y^*, r^*, \lambda^{(1)}, \lambda^{(2)}, \mu^{(1)}, \mu^{(2)}, \rho^{(1)}, \rho^{(2)})\\ &\le \mathscr{L}(x^*, y^*, r^*, \lambda^{(1)*}, \lambda^{(2)*}, \mu^{(1)*}, \mu^{(2)*}, \rho^{(1)*}, \rho^{(2)*}) = 0\\ &\le \mathscr{L}(x, y, r, \lambda^{(1)*}, \lambda^{(2)*}, \mu^{(1)*}, \mu^{(2)*}, \rho^{(1)*}, \rho^{(2)*})\end{aligned} \tag{17}$$

$\forall (x, y, r) \in L^2([0,T], \mathbb{R}^{2mn+n})$, $\forall \lambda^{(1)}, \lambda^{(2)} \in L^2([0,T], \mathbb{R}^{mn}_+)$, $\forall \mu^{(1)}, \mu^{(2)} \in L^2([0,T], \mathbb{R}^m)$, $\forall \rho^{(1)}, \rho^{(2)} \in L^2([0,T], \mathbb{R}^n_+)$ *and, a.e. in* $[0,T]$,

$$-\frac{\partial u_i(t, x^*(t), y^*(t))}{\partial x_{ij}} - (1 - \tau_{ij}(t)) r_j^*(t) - \lambda_{ij}^{(1)*}(t) - \mu_i^{(1)*}(t) = 0,$$

$$\forall i = 1, \dots, m, \ \forall j = 1 \dots, n;$$

$$-\frac{\partial u_i(t, x^*(t), y^*(t))}{\partial y_{ij}} + (1 - \tau_{ij}(t))(1 + h_j(t)) r_j^*(t) - \lambda_{ij}^{(2)*}(t) - \mu_i^{(2)*}(t) = 0,$$

$$\forall i = 1, \dots, m, \ \forall j = 1 \dots, n;$$

$$\sum_{i=1}^{m} (1 - \tau_{ij}(t)) \Big[x_{ij}^*(t) - (1 + h_j(t)) y_{ij}^*(t) \Big] + F_j(t) + \rho_j^{(2)*}(t) = \rho_j^{(1)*}(t), \tag{18}$$

$$\forall j = 1, \dots, n;$$

$$\lambda_{ij}^{(1)*}(t) x_{ij}^*(t) = 0, \ \lambda_{ij}^{(2)*}(t) y_{ij}^*(t) = 0, \quad \forall i = 1, \dots, m, \ \forall j = 1, \dots, n \tag{19}$$

$$\mu_i^{(1)*}(t) \left(\sum_{j=1}^{n} x_{ij}^*(t) - s_i(t) \right) = 0, \quad \mu_i^{(2)*}(t) \left(\sum_{j=1}^{n} y_{ij}^*(t) - l_i(t) \right) = 0, \tag{20}$$

$$\forall i = 1, \dots, m$$

$$\rho_j^{(1)*}(t) (\underline{r}_j(t) - r_j^*(t)) = 0, \ \rho_j^{(2)*}(t) (r_j^*(t) - \overline{r}_j(t)) = 0, \quad \forall j = 1, \dots, n. \tag{21}$$

Formula (18) represents the Deficit Formula. Indeed, if $\rho_j^{(1)*}(t)$ is positive, then the prices are minimal and there is a supply excess of instrument j as an asset and of the charge $F_j(t)$, namely the economy is in deficit and, for this reason, $\rho_j^{(1)*}(t)$ is called *the deficit variable* and represents the deficit per unit.

Analogously if $\rho_j^{(2)*}(t)$ is positive, then the prices are maximal and there is a demand excess of instrument j as a liability, namely there is a surplus in the economy. For this reason $\rho_j^{(2)*}(t)$ is called *the surplus variable* and represents the surplus per unit.

From (18) it is possible to obtain the Balance Law

$$\sum_{i=1}^{m} l_i(t) = \sum_{i=1}^{m} s_i(t) - \sum_{i=1}^{m}\sum_{j=1}^{n} \tau_{ij}(t)\left[x_{ij}^*(t) - y_{ij}^*(t)\right]$$
$$-\sum_{i=1}^{m}\sum_{j=1}^{n}(1-\tau_{ij}(t))h_j(t)y_{ij}^*(t) + \sum_{j=1}^{n} F_j(t) - \sum_{j=1}^{n}\rho_j^{(1)*}(t) + \sum_{j=1}^{n}\rho_j^{(2)*}(t). \tag{22}$$

Finally, assuming that the taxes $\tau_{ij}(t)$, $i = 1,\dots,m$, $j = 1,\dots,n$, have a common value $\theta(t)$, and the increments $h_j(t)$, $j = 1,\dots,n$, have a common value $i(t)$, otherwise we can consider the average values (see Remark 7.1 in [2]), the significant Liability Formula follows

$$\sum_{i=1}^{m} l_i(t) = \frac{(1-\theta(t))\sum_{i=1}^{m} s_i(t) + \sum_{j=1}^{n} F_j(t) - \sum_{j=1}^{n}\rho_j^{(1)*}(t) + \sum_{j=1}^{n}\rho_j^{(2)*}(t)}{(1-\theta(t))(1+i(t))}.$$

2.3 The Viewpoints of the Sector and of the System

The financial problem can be considered from two different perspectives: one from the *Point of View of the Sectors*, which try to maximize the utility and a second point of view, that we can call *System Point of View*, which regards the whole equilibrium, namely the respect of the previous laws. For example, from the point of view of the sectors, $l_i(t)$, for $i = 1,\dots,m$, are liabilities, whereas for the economic system they are investments and, hence, the Liability Formula, from the system point of view, can be called "*Investments Formula*". The system point of view coincides with the dual Lagrange problem (the so-called "shadow market") in which $\rho_j^{(1)}(t)$ and $\rho_j^{(2)}(t)$ are the dual multipliers, representing the deficit and the surplus per unit arising from instrument j. Formally, the dual problem is given by

Find $(\rho^{(1)*}, \rho^{(2)*}) \in L^2([0,T], \mathbb{R}_+^{2n})$ such that

$$\sum_{j=1}^{n}\int_0^T (\rho_j^{(1)}(t)-\rho_j^{(1)*}(t))(\underline{r}_j(t)-r_j^*(t))dt \tag{23}$$

$$+\sum_{j=1}^{n}\int_0^T (\rho_j^{(2)}(t)-\rho_j^{(2)*}(t))(r_j^*(t)-\overline{r}_j(t))dt \leq 0,$$

$$\forall(\rho^{(1)},\rho^{(2)}) \in L^2([0,T],\mathbb{R}_+^{2n}).$$

Indeed, taking into account inequality (17), we get

$$-\sum_{i=1}^{m}\sum_{j=1}^{n}\int_0^T (\lambda_{ij}^{(1)}(t)-\lambda_{ij}^{(1)*}(t))x_{ij}^*(t)\,dt-\sum_{i=1}^{m}\sum_{j=1}^{n}\int_0^T (\lambda_{ij}^{(2)}-\lambda_{ij}^{(2)*})y_{ij}^*(t)\,dt$$

$$-\sum_{i=1}^{m}\int_0^T (\mu_i^{(1)}(t)-\mu_i^{(1)*}(t))\left(\sum_{j=1}^{n}x_{ij}^*(t)-s_i(t)\right)dt$$

$$-\sum_{i=1}^{m}\int_0^T (\mu_i^{(2)}(t)-\mu_i^{(2)*}(t))\left(\sum_{j=1}^{n}y_{ij}^*(t)-l_i(t)\right)dt$$

$$+\sum_{j=1}^{n}\int_0^T (\rho_j^{(1)}(t)-\rho_j^{(1)*}(t))(\underline{r}_j(t)-r_j^*(t))\,dt$$

$$+\sum_{j=1}^{n}\int_0^T (\rho_j^{(2)}(t)-\rho_j^{(2)*}(t))(r_j^*(t)-\overline{r}_j(t))\,dt \leq 0$$

$\forall\lambda^{(1)},\lambda^{(2)} \in L^2([0,T],\mathbb{R}_+^{mn})$, $\mu^{(1)},\mu^{(2)} \in L^2([0,T],\mathbb{R}^m)$, $\rho^{(1)},\rho^{(2)} \in L^2([0,T],\mathbb{R}_+^n)$.

Choosing $\lambda^{(1)}=\lambda^{(1)*}$, $\lambda^{(2)}=\lambda^{(2)*}$, $\mu^{(1)}=\mu^{(1)*}$, $\mu^{(2)}=\mu^{(2)*}$, we obtain the dual problem (23)

Note that, from the *System Point of View*, also the expenses of the institutions $F_j(t)$ are supported from the liabilities of the sectors.

Remark 2 Let us recall that from the Liability Formula we get the following index $E(t)$, called "Evaluation Index", that is very useful for the rating procedure:

$$E(t)=\frac{\sum_{i=1}^{m}l_i(t)}{\sum_{i=1}^{m}\tilde{s}_i(t)+\sum_{j=1}^{n}\tilde{F}_j(t)},$$

where we set

$$\tilde{s}_i(t)=\frac{s_i(t)}{1+i(t)},\quad \tilde{F}_j(t)=\frac{F_j(t)}{1+i(t)-\theta(t)-\theta(t)i(t)}.$$

From the Liability Formula we obtain

$$E(t)=1-\frac{\sum_{j=1}^{n}\rho_j^{(1)*}(t)}{(1-\theta(t))(1+i(t))\left(\sum_{i=1}^{m}\tilde{s}_i(t)+\sum_{j=1}^{n}\tilde{F}_j(t)\right)}$$

$$+\frac{\sum_{j=1}^{n}\rho_j^{(2)*}(t)}{(1-\theta(t))(1+i(t))\left(\sum_{i=1}^{m}\tilde{s}_i(t)+\sum_{j=1}^{n}\tilde{F}_j(t)\right)} \tag{24}$$

If $E(t)$ is greater or equal than 1, the evaluation of the financial equilibrium is positive (better if $E(t)$ is proximal to 1), whereas if $E(t)$ is less than 1, the evaluation of the financial equilibrium is negative.

2.4 The Contagion Problem

Let us note that in the balance law:

$$\sum_{i=1}^{m} l_i(t)-\sum_{i=1}^{m} s_i(t)+\sum_{i=1}^{m}\sum_{j=1}^{n}\tau_{ij}(t)\left[x_{ij}^*(t)-y_{ij}^*(t)\right]$$
$$+\sum_{i=1}^{m}\sum_{j=1}^{n}(1-\tau_{ij}(t))h_j(t)y_{ij}^*(t)-\sum_{j=1}^{n}F_j(t)=-\sum_{j=1}^{n}\rho_j^{(1)*}(t)+\sum_{j=1}^{n}\rho_j^{(2)*}(t),$$

if

$$\sum_{j=1}^{n}\rho_j^{(1)*}(t)>\sum_{j=1}^{n}\rho_j^{(2)*}(t), \tag{25}$$

namely the sum of all the deficit exceeds the sum of all the surplus, the balance of all the financial entities is negative (see also [18]). In this case we say that a negative contagion is determined and we can assume that the insolvencies of individual entities propagate through the entire system. It is sufficient that only one deficit $\rho_j^{(1)*}(t)$ is large to obtain, even if the other $\rho_j^{(2)*}(t)$ are lightly positive, a negative balance for the all system.

When condition (25) is verified, we get $E(t) \leq 1$ and, hence, also $E(t)$ is a significant indicator that the financial contagion happens.

In [20] a regularity result of $\rho_j^{(1)*}(t)$, $\rho_j^{(2)*}(t)$, has been proved. Let us set

$$F(t) = [F_1(t), F_2(t), \dots, F_n(t)]^T;$$

$$\nu = (x, y, r) = \left(\left(x_{ij}\right)_{\substack{i=1,\dots,m\\ j=1,\dots,n}}, \left(y_{ij}\right)_{\substack{i=1,\dots,m\\ j=1,\dots,n}}, \left(r_j\right)_{j=1,\dots,n} \right);$$

$$\begin{aligned} A(t,\nu) = \Bigg(&\left[-\frac{\partial u_i(t,x,y)}{\partial x_{ij}} - (1-\tau_{ij}(t))r_j(t) \right]_{\substack{i=1,\dots,m\\ j=1,\dots,n}}, \\ &\left[-\frac{\partial u_i(t,x,y)}{\partial y_{ij}} + (1-\tau_{ij}(t))(1+h_j(t))r_j(t) \right]_{\substack{i=1,\dots,m\\ j=1,\dots,n}}, \\ &\left[\sum_{i=1}^{m} (1-\tau_{ij}(t)) \left(x_{ij}(t) - (1+h_j(t))y_{ij}(t)\right) + F_j(t) \right]_{j=1,\dots,n} \Bigg); \end{aligned} \tag{26}$$

$$A : \mathscr{K} \to L^2([0,T], \mathbb{R}^{2mn+n}),$$

with

$$\mathscr{K} = P \times \mathscr{R}.$$

Let us note that $\mathscr{K}$ is a convex, bounded and closed subset of $L^2([0,T], \mathbb{R}^{2mn+n})$. Moreover assumption (8) implies that A is lower semicontinuous along line segments.

The following result holds true (see [20] Theorem 2.4):

Theorem 9 *Let $A \in C^0([0,T], \mathbb{R}^{2mn+n})$ be strongly monotone in x and y, monotone in r, namely, there exists α such that, for $t \in [0,T]$,*

$$\langle\langle A(t,\nu_1) - A(t,\nu_2), \nu_1 - \nu_2 \rangle\rangle \geq \alpha(\|x_1 - x_2\|^2 + \|y_1 - y_2\|^2), \tag{27}$$

$\forall \nu_1 = (x_1, y_1, r_1)$, $\nu_2 = (x_2, y_2, r_2) \in \mathbb{R}^{2mn+n}$.

Let $\underline{r}(t)$, $\overline{r}(t)$, $h(t)$, $F(t) \in C^0([0,T], \mathbb{R}^n_+)$, let $\tau(t) \in C^0([0,T], \mathbb{R}^{mn})$ and let $s, l \in C^0([0,T], \mathbb{R}^m)$, satisfying the following assumption (β):

- *there exists $\delta_1(t) \in L^2([0,T])$ and $c_1 \in \mathbb{R}$ such that, for a.a. $t \in [0,T]$:*

$$\|s(t)\| \leq \delta_1(t) + c_1;$$

- *there exists $\delta_2(t) \in L^2([0,T])$ and $c_2 \in \mathbb{R}$ such that, for a.a. $t \in [0,T]$:*

$$\|l(t)\| \leq \delta_2(t) + c_2.$$

Then the Lagrange variables, $\rho^{(1)*}(t)$, $\rho^{(2)*}(t)$, which represent the deficit and the surplus per unit, respectively, are continuous too.

2.5 An Example of a Markowitz-Type Risk Measure

We generalize and provide an evolutionary Markowitz-type measure of the risk proposed with a memory term. This function is effective, namely an existence theorem for the general financial problem holds (see [21]). In this way we cover a lack providing the existence of a significant evolutionary measure of the risk. The particular, but significant, example of utility function is:

$$u_i(x_i(t), y_i(t))$$

$$= \begin{bmatrix} x_i(t) \\ y_i(t) \end{bmatrix}^T Q^i \begin{bmatrix} x_i(t) \\ y_i(t) \end{bmatrix} + \int_0^t \begin{bmatrix} x_i(t-z) \\ y_i(t-z) \end{bmatrix}^T Q^i \begin{bmatrix} x_i(t-z) \\ y_i(t-z) \end{bmatrix} dz, \tag{28}$$

where Q^i denotes the sector i's assessment of the standard deviation of prices for each instrument j.

3 Applications to the Elastic-Plastic Torsion Problem

In this chapter we apply *Assumption S* to the elastic-plastic torsion problem.

3.1 Presentation of the Problem

The elastic-plastic torsion problem and its relationships with obstacle problem have been deeply investigated in years 1965–1980. Later on these studies have been resumed, with particular regards to existence and properties of Lagrange multipliers. The existence of Lagrange multipliers is strictly related to strong duality theory.

The problem arises from aerodynamics and has been formulated by R. Von Mises (see [53]): *the elastic-plastic torsion problem of a cylindrical bar with cross section Ω is to find a function $u(x)$ which vanishes on the boundary $\partial\Omega$ and, together with its first derivatives, is continuous on Ω; nowhere on Ω the gradient of u must have an absolute value (modulus) less than or equal to a given positive constant τ; whenever in Ω the strict inequality holds, the function u must satisfy the differential equation $\Delta u = -2\mu\theta$, where the positive constants μ and θ denote the shearing modulus and the angle of twist per unit length, respectively.*

From the Von Mises formulation it follows that the cross section Ω is divided into two regions: an elastic region $E = \{x \in \Omega : |Du(x)| < 1\}$ and a plastic region $P = \{x \in \Omega : |Du(x)| = 1\}$.

This problem is a free boundary one and a suitable tool for studying this kind of problems is the variational inequality theory. To this end, let us consider the following variational inequality:

Find $u \in K = \left\{ v \in H_0^{1,\infty}(\Omega) : |Dv| = \sum_{i=1}^{n} \left(\frac{\partial v}{\partial x_i} \right)^2 \leq 1 \text{ a.e. on } \Omega \right\}$ such that

$$\int_\Omega \sum_{i=1}^{n} \frac{\partial u}{\partial x_i} \left(\frac{\partial v}{\partial x_i} - \frac{\partial u}{\partial x_i} \right) dx \geq \int_\Omega F(v-u)dx \quad \forall v \in K, \tag{29}$$

with $\Omega \subset \mathbb{R}^n$ open bounded convex set with Lipschitz boundary $\partial\Omega$, $F \in L^p(\Omega)$, $p > 1$.

As it is well known, (29) admits a unique solution $u \in W^{2,p}(\Omega) \cap K$ (see [6, 7]).

In literature, in the planar case the existence and the properties of a smooth solution of the elastic-plastic torsion problem have been studied by Ting ([50–52]), whereas multidimensional case has been studied by Brezis in [5], who proved the existence of a Lagrange multiplier for (29), assuming $F = cost > 0$, namely, if u is the solution of variational inequality (29), then there exists a unique $\mu \in L^\infty(\Omega)$, $\mu \geq 0$ a.e. in Ω such that:

$$\begin{cases} \mu(1 - |Du|) = 0 \text{ a.e. in } \Omega \\ -\Delta u - \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left(\mu \frac{\partial u}{\partial x_i} \right) = F \text{ in the sense of } D^{'}(\Omega), \end{cases} \tag{30}$$

that is the solution of (29) solves the elastic-plastic torsion problem.

Conversely, if $u \in K$ and there exists μ satisfying (30), then it is easily proved that u is the solution of (29).

In virtue of this equivalence the variational inequality (29) is the elastic-plastic torsion problem formulated by Von Mises.

Moreover, in this case, the solution to elastic-plastic torsion problem coincides with the solution to obstacle problem and is nonnegative.

Only recently the relationship between problems (29) and (30) has been clarified in the case of general linear operators and nonlinear monotone operators. In this section we will describe these results, together with the study of radial solutions to the elastic-plastic torsion.

3.2 The Elastic-Plastic Torsion Problem for Linear Operators

First, we establish the existence of Lagrange multipliers associated to a general linear operator. In particular we prove that the Lagrange multipliers associated to the elastic-plastic torsion problem for linear operators always exist and, in general, they result as a Radon measure. This result is proved using the classical strong duality.

Moreover, the result may be generalized, namely, it is possible to prove that the L^p Lagrange multipliers exist if and only if *Assumption S* holds and this is a consequence of the new strong duality described in Sect. 1.1.

Let us now describe the problem in detail.

Let $\Omega \subset \mathbb{R}^n$ be an open bounded domain either convex or with boundary of class $C^{1,1}$. Let us consider the linear elliptic operator

$$\mathscr{L}u = -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_j}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) + \sum_{i=1}^{n} b_i \frac{\partial u}{\partial x_i} + cu \tag{31}$$

with associated bilinear form on $H_0^{1,\infty}(\Omega) \times H_0^{1,\infty}(\Omega)$ given by

$$a(u,v) = \int_\Omega \left(\sum_{i,j=1}^{n} a_{ij}\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i} + \sum_{i=1}^{n} b_i \frac{\partial u}{\partial x_i} v + cuv\right) dx,$$

where

$$\begin{cases} \displaystyle\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq a|\xi|^2 \text{ a.e. on } \Omega, \forall \xi \in \mathbb{R}^n \\ a > 0, a_{ij} \in C^1(\overline{\Omega}), b_i, c \in L^\infty(\Omega) \\ c > 0 \text{ such large that } a(u,u) \geq \alpha \|u\|^2_{H_0^{1,\infty}(\Omega)}, \ \alpha > 0, \ \forall u \in H_0^{1,\infty}(\Omega). \end{cases} \tag{32}$$

Let us consider the variational inequality:

$$\textit{Find } u \in K = \left\{ v \in H_0^{1,\infty}(\Omega) : \sum_{i=1}^{n}\left(\frac{\partial v}{\partial x_i}\right)^2 \leq 1, \text{ a.e. on } \Omega \right\} \textit{ such that:}$$

$$\int_\Omega \mathscr{L}u(v-u)\,dx \geq \int_\Omega F(v-u)\,dx, \quad \forall v \in K. \tag{33}$$

As it is well known, variational inequality (33) admits a unique solution $u \in K$ and, if $F \in L^p(\Omega)$, $p > 1$, $u \in W^{2,p}(\Omega) \cap K$ (see [6, 7]).

We are able to prove the existence of a Lagrange multiplier for variational inequality (33) as a Radon measure (see [28, 30]).

Theorem 10 *Under the above assumptions on Ω and $\mathscr{L}$, let $F \in L^p(\Omega)$, $p > 1$, and $u \in K \cap W^{2,p}(\Omega)$ be the solution to problem (33). Then there exists $\overline{\mu} \in (L^\infty(\Omega))^*$ such that*

$$\begin{cases} \langle \overline{\mu}, y \rangle \geq 0 \quad \forall y \in L^\infty(\Omega), \ y \geq 0 \quad a.e. \ in \ \Omega; \\ \langle \overline{\mu}, \left(\sum_{i=1}^{n} \left(\frac{\partial u}{\partial x_i} \right)^2 - 1 \right) \rangle = 0 \\ \int_\Omega (\mathscr{L} u - F) \varphi \, dx = \langle \overline{\mu}, -2 \sum_{i=1}^{n} \frac{\partial u}{\partial x_i} \frac{\partial \varphi}{\partial x_i} \rangle \quad \forall \varphi \in H_0^{1,\infty}(\Omega). \end{cases} \tag{34}$$

The theorem means that, if we consider a solution u of variational inequality (33), then conditions (34) are satisfied. Moreover it is possible to show that, as a consequence of conditions (34), the solution of variational inequality (33) is also a solution of the elastic-plastic torsion problem and vice versa.

We mention the paper [9], in which the authors prove the existence of a Lagrange multiplier as a positive Radon measure under different assumptions and using a different technique.

In order to prove Theorem 10 we use the strong duality property in the classical sense (see [26, 38]) and its consequence on the existence of saddle points of the Lagrange functional. We briefly recall them.

Theorem 11 (Classical Strong Duality Property) *Let S be a nonempty subset of a real linear space X; $(Y, \| \cdot \|)$ be a partially ordered real normed space with ordering cone C; $f : S \to \mathbb{R}$ be a given objective functional; $g : S \to Y$ be a given constraint mapping; let the composite mapping $(f, g) : S \to \mathbb{R} \times Y$ be convex-like with respect to product cone $\mathbb{R}_+ \times C$ in $\mathbb{R} \times Y$. Let the constraint set be given as $\mathbb{K} := \{v \in S : g(v) \in -C\}$ which is assumed to be nonempty. Let the ordering cone C have a nonempty interior $int(C)$. If the primal problem*

$$\begin{array}{c} \min \quad f(v) \\ v \in S \\ g(v) \in -C \end{array} \tag{35}$$

is solvable and the generalized Slater condition is satisfied, namely there is a vector $\hat{v} \in S$ with $g(\hat{v}) \in -int(C)$, then the dual problem

$$\max_{\mu \in C^*} \inf_{v \in S} [f(v) + \mu(g(v))] \tag{36}$$

is also solvable and the extremal values of the two problems are equal. Moreover, if u is the optimal solution to problem (35) and $\overline{\mu} \in C^*$ is a solution of the problem (36), it results

$$\overline{\mu}(g(u)) = 0. \tag{37}$$

Theorem 12 *Under the same assumptions as above, suppose the ordering cone C to be closed. Then a point $(u, \overline{\mu}) \in S \times C^*$ is a saddle point of the Lagrange functional L if and only if u is a solution of the primal problem (35), $\overline{\mu}$ is a solution of the dual problem (36) and the extremal values of the two problems are equal.*

Indeed, let $u \in K \cap W^{2,p}(\Omega)$ be the solution to (33). Let us rewrite the variational inequality (33) as the minimum problem

$$\min_{v \in K} f(v) = f(u) = 0 \tag{38}$$

where

$$f(v) = \int_\Omega (\mathscr{L}u - F)(v - u)\, dx.$$

Let us set $S = X = H_0^{1,\infty}(\Omega)$, $Y = L^\infty(\Omega)$,

$$f(v) : H_0^{1,\infty}(\Omega) \to \mathbb{R},$$

$$g(v) = \sum_{i=1}^{n} \left(\frac{\partial v}{\partial x_i}\right)^2 - 1 : H_0^{1,\infty}(\Omega) \to L^\infty(\Omega),$$

$$C = \{w \in L^\infty(\Omega) \ : \ w \geq 0\} = \{w \in L^\infty(\Omega) \ : \ w(x) \geq 0 \, a.a. \, x \in \Omega\}.$$

Since f and g are convex on the space $H_0^{1,\infty}(\Omega)$, then the composite mapping (f, g) is convex-like with respect to the product cone $\mathbb{R}_+ \times C$ in $\mathbb{R} \times Y$. Moreover, we are able to prove that $int(C) \neq \emptyset$ and that generalized Slater condition is verified. Then, every assumption of Theorem 11 is verified and, since the primal problem is solvable, it follows that the dual problem

$$\max_{\mu \in C^*} \inf_{v \in S} [f(v) + \mu(g(v))] \tag{39}$$

is also solvable and the extremal values of the two problems coincide. Moreover, if u is a solution of the problem (38) and $\overline{\mu} \in C^*$ is a solution of the problem (39), condition (37) holds, namely

$$\overline{\mu}(g(u)) = 0. \tag{40}$$

Finally, since the ordering cone C is closed, we may apply Theorem 12, from which it follows

$$\int_\Omega (\mathscr{L}u - F)\varphi \, dx = \langle \overline{\mu}, -2 \sum_{i=1}^{n} \frac{\partial u}{\partial x_i} \frac{\partial \varphi}{\partial x_i} \rangle. \tag{41}$$

In conclusion, since

$$C^* = \{\mu \in (L^\infty(\Omega))^* : \ \mu(y) \geq 0 \ \forall y \in C\}$$

$$= \{\mu \in (L^\infty(\Omega))^* : \ \mu(y) \geq 0 \ \forall y \in L^\infty(\Omega), \ y(x) \geq 0 \ a.a. \ x \ \in \Omega\},$$

from (40), (41) we obtain that, if u is a solution of (33), then of the primal problem (38), there exists $\overline{\mu} \in C^*$ solution of the dual problem (39) and the following conditions are satisfied:

$$\langle \overline{\mu}, y \rangle \geq 0 \quad \forall y \in L^\infty(\Omega), \ y \geq 0 \quad a.e. \ in \ \Omega;$$

$$\langle \overline{\mu}, \left(\sum_{i=1}^n \left(\frac{\partial u}{\partial x_i} \right)^2 - 1 \right) \rangle = 0$$

$$\int_\Omega (\mathscr{L}u - F)\varphi \, dx = \langle \overline{\mu}, -2 \sum_{i=1}^n \frac{\partial u}{\partial x_i} \frac{\partial \varphi}{\partial x_i} \rangle \quad \forall \varphi \in H_0^{1,\infty}(\Omega),$$

namely, Theorem 10 is proved.

Moreover, if $\overline{\mu} \in (L^\infty(\Omega))^*$, $\overline{\mu}$ can be expressed by a Radon's integral with respect to the finitely additive measure Ψ:

$$\overline{\mu}(v) = \int_\Omega v(x) \Psi(dx).$$

Ψ is finitely additive, has a bounded total variation and is absolutely continuous with respect to the Lebesgue measure, that is $m(B) = 0$ implies $\Psi(B) = 0$.

From this properties of $\overline{\mu}$ and conditions (34) it is possible to prove that the solution of variational inequality (33) is also a solution of the elastic plastic torsion problem and vice versa.

In order to obtain a regularization of this result, namely to obtain the existence of a Lagrange multiplier for variational inequality (33) as a L^∞ function, it is necessary to consider the convex set K in $H_0^1(\Omega)$, that is

$$K_\nabla = \left\{ v \in H_0^1(\Omega) : \sum_{i=1}^n \left(\frac{\partial v}{\partial x_i} \right)^2 \leq 1 \text{ a.e. on } \Omega \right\}.$$

But in this case the interior of the ordering cone, which defines the sign constraints, is empty, then it is not possible to apply the classical strong duality theory. It is necessary to apply the new strong duality principle described in Sect. 1.1 and, then, we obtain the following characterization in terms of *Assumption S* of the elastic-plastic torsion problem (see Theorem 3.4 in [19]).

Theorem 13 *Let $u \in K_\nabla \cap W^{2,p}(\Omega)$ be the solution to problem*

$$\int_\Omega \mathscr{L}u(v - u) \, dx \geq \int_\Omega F(v - u) \, dx, \quad \forall v \in K_\nabla. \tag{42}$$

Then there exists $\bar{\mu} \in L^{\infty}(\Omega)$ *such that*

$$\begin{cases} \bar{\mu} \geq 0 \\ \bar{\mu}\left(1-\sum_{i=1}^{n}\left(\frac{\partial u}{\partial x_i}\right)^2\right)=0 \ \text{ a.e. } \ \text{ in } \ \Omega \\ Lu-F=2\sum_{i=1}^{n}\frac{\partial}{\partial x_i}\left(\bar{\mu}\frac{\partial u}{\partial x_i}\right) \ \textit{in the sense of distributions} \end{cases} \tag{43}$$

if and only if the solution u of (42) *verifies Assumption S.*

The theorem means that, if the solution of (42) verifies *Assumption S*, then conditions (43) are satisfied, that is the solution of (42) is a solution of the elastic-plastic torsion problem; vice versa if $u \in W^{2,p}(\Omega)$ verifies (43), and, then, in particular, is a solution of the elastic-plastic torsion problem, then u solves (42) and verifies *Assumption S*.

The thesis is achieved rewriting the variational inequality (42) as the minimum problem

$$\min_{v \in K_{\nabla}} f(v) = f(u) = 0 \tag{44}$$

where

$$f(v) = \int_{\Omega} (\mathscr{L}u - F)(v-u)\, dx \tag{45}$$

with the settings

$$\begin{gathered} f(v): H_0^1(\Omega) \to \mathbb{R} \\ g(v) = \sum_{i=1}^{n}\left(\frac{\partial v}{\partial x_i}\right)^2 - 1 : H_0^1(\Omega) \to L^1(\Omega) \\ C = \{w \in L^1(\Omega) \ : \ w \geq 0\} \\ \widetilde{M} = \left\{(\psi(v)+\alpha, g(v)+w) : \ v \in H_0^1(\Omega) \setminus K_{\nabla}, \ \alpha \geq 0, \ w \in C\right\}. \end{gathered}$$

Assuming that *Assumption S* holds, from Theorem 1 it follows that there exists $\bar{\mu} \in C^* = \left\{\mu \in L^{\infty}(\Omega) : \ \int_{\Omega} \mu\, v\, dx \geq 0, \ \forall v \in L^1(\Omega)\right\}$ such that

$$\bar{\mu}\left(\sum_{i=1}^{n}\left(\frac{\partial u}{\partial x_i}\right)^2 - 1\right) = 0 \text{ a.e. in } \Omega. \tag{46}$$

Then the thesis is obtained using Theorem 2 and (46). Vice versa assuming that conditions (43) hold, it is possible to show that u verifies *Assumption S* and, finally, it is easy to verify that u is a solution to variational inequality (42).

3.3 The Elastic-Plastic Torsion Problem for Nonlinear Monotone Operators

Second, we are aimed at the investigation of the existence of Lagrange multipliers associated to the following nonlinear problem (see [7] for the existence and the regularity of solutions to (47)):

$$\text{Find } u \in K : \int_\Omega \sum_{i=1}^n a_i(Du)\left(\frac{\partial v}{\partial x_i} - \frac{\partial u}{\partial x_i}\right) dx \geq \int_\Omega F(v-u)dx, \quad \forall v \in K. \tag{47}$$

In particular, we are able to prove that the Lagrange multiplier is always a Radon measure when the operator is strictly monotone, whereas the Lagrange multiplier is a L^p function when the operator is strongly monotone (see [32]). The first result is proved using classical strong duality theory, whereas for the second one we apply the new strong duality theory described in Sect. 1.1.

From now on we assume that $\Omega \subset \mathbb{R}^n$ is an open bounded convex set with Lipschitz boundary $\partial\Omega$ and a is an operator of class C^2. In a first step we suppose that the operator is strictly monotone, namely

$$(a(P) - a(Q), P - Q) > 0 \quad \forall P, Q \in \mathbb{R}^n, P \neq Q. \tag{48}$$

In a first theorem we are able to prove the equivalence between elastic-plastic torsion problem and obstacle problem.

Theorem 14 *Under the above assumptions on Ω and a, if $a(0) = 0$ and $F \equiv const.$, the solution u of (47) coincides with the solution of*

$$\text{Find } u \in K_\delta : \int_\Omega \sum_{i=1}^n a_i(Du)\left(\frac{\partial v}{\partial x_i} - \frac{\partial u}{\partial x_i}\right) dx \geq \int_\Omega F(v-u)dx, \quad \forall v \in K_\delta, \tag{49}$$

where

$$K_\delta = \left\{ v \in W_0^{1,\infty}(\Omega) : |v(x)| \leq \delta(x) = dist(x, \partial\Omega) \text{ a.e. on } \Omega \right\}.$$

In a second theorem we prove the existence of Lagrange multipliers for problem (47) as a Radon measure.

Theorem 15 *Under the above assumptions on Ω and a, let $F \in L^p(\Omega)$, $p > 1$, and $u \in K$ be the solution to (47). Then there exists $\overline{\mu} \in (L^\infty(\Omega))^*$ such that*

$$\begin{cases} \langle \overline{\mu}, y \rangle \geq 0 \quad \forall y \in L^{\infty}(\Omega),\ y \geq 0 \quad \text{a.e. in } \Omega; \\ \langle \overline{\mu}, \left(\sum_{i=1}^{n} \left(\frac{\partial u}{\partial x_i} \right)^2 - 1 \right) \rangle = 0; \\ \int_{\Omega} \left\{ \sum_{i=1}^{n} a_i(Du) \frac{\partial \varphi}{\partial x_i} - F\varphi \right\} dx = \langle \overline{\mu}, -2 \sum_{i=1}^{n} \frac{\partial u}{\partial x_i} \frac{\partial \varphi}{\partial x_i} \rangle \quad \forall \varphi \in W_0^{1,\infty}(\Omega). \end{cases} \tag{50}$$

From conditions (50) it follows that, if u belongs to the elastic region E, $\overline{\mu} \equiv 0$ and then u is a solution of the elliptic equation $Au = F$ a.e. in Ω, where $A = -\sum_{i=1}^{n} \frac{\partial a_i(Du)}{\partial x_i}$ and, in particular, a solution of (47) solves the elastic-plastic torsion problem. Conversely it is easily proved that, if $u \in K$ satisfies conditions (50), then u solves variational inequality (47).

The proof of Theorem 15 is based on the following steps. First we rewrite variational inequality (47) as the minimum problem

$$\min_{v \in K} f(v) = f(u) = 0 \tag{51}$$

where

$$f(v) = \int_{\Omega} \left\{ \sum_{i=1}^{n} a_i(Du) \left(\frac{\partial v}{\partial x_i} - \frac{\partial u}{\partial x_i} \right) - F(v-u)dx \right\}. \tag{52}$$

Then, setting $S = X = W_0^{1,\infty}(\Omega)$, $Y = L^{\infty}(\Omega)$,

$$f(v) : W_0^{1,\infty}(\Omega) \to \mathbb{R}$$

$$g(v) = \sum_{i=1}^{n} \left(\frac{\partial v}{\partial x_i} \right)^2 - 1 : W_0^{1,\infty}(\Omega) \to L^{\infty}(\Omega),$$

as in the linear case (see [29, 30]) we are able to prove that the assumptions of Theorems 11 and 12 hold. Consequently, if u is a solution of (47), then of the problem (51), there exists $\overline{\mu} \in C^*$ solution of the dual problem

$$\max_{\mu \in C^*} \inf_{v \in S} [f(v) + \langle \mu, g(v) \rangle] \tag{53}$$

and $(u, \overline{\mu})$ is a saddle point of the so called Lagrange functional

$$L(v, \mu) = f(v) + \langle \mu, g(v) \rangle,\ \forall v \in W_0^{1,\infty}(\Omega), \forall \mu \in C^*,$$

namely

$$L(u, \mu) \leq L(u, \overline{\mu}) \leq L(v, \overline{\mu}),\ \forall v \in W_0^{1,\infty}(\Omega), \forall \mu \in C^*. \tag{54}$$

Via variational arguments we obtain that

$$\int_{\Omega} \left\{ \sum_{i=1}^{n} a_i(Du) \frac{\partial \varphi}{\partial x_i} - f\varphi \right\} dx = \langle \overline{\mu}, -2 \sum_{i=1}^{n} \frac{\partial u}{\partial x_i} \frac{\partial \varphi}{\partial x_i} \rangle \quad \forall \varphi \in W_0^{1,\infty}(\Omega). \tag{55}$$

Then from (37), (55), we obtain conditions (50).

If now we assume strong monotonicity assumption

$$(a(P) - a(Q), P - Q) > \nu \|P - Q\|^2 \quad \forall P, Q \in \mathbb{R}^n, P \neq Q, \tag{56}$$

we are able to prove the following regularization theorem concerning Lagrange multipliers.

Theorem 16 *Under the same assumptions on Ω as above, let a satisfy strong monotonicity assumption* (56)*, with $a(0) = 0$, let F be a positive constant and $u \in K \cap W^{2,p}(\Omega)$ be the solution to problem* (47)*. Then there exists $\bar{\mu} \in L^p(\Omega)$ such that*

$$\begin{cases} \bar{\mu} \geq 0 \text{ a.e. in } \Omega \\ \bar{\mu} \left(1 - \displaystyle\sum_{i=1}^{n} \left(\frac{\partial u}{\partial x_i} \right)^2 \right) = 0 \text{ a.e. in } \Omega \\ \displaystyle\sum_{i=1}^{n} \frac{\partial a_i(Du)}{\partial x_i} + F = \bar{\mu} \text{ a.e. in } \Omega. \end{cases} \tag{57}$$

Of course, as for the linear problem, it is easy to prove that, if $u \in K$ and there exists $\bar{\mu}$ satisfying (57), then u is also the solution to problem (47).

Let us notice that, if u is the solution to problem (47), in virtue of Theorem 14, it is the solution of problem (49). In particular, since $f \equiv const. > 0$, a is monotone and $a(0) = 0$, it is possible to prove that u is the solution of the problem

$$\text{Find } u \in K_1 : \int_{\Omega} \sum_{i=1}^{n} a_i(Du) \left(\frac{\partial v}{\partial x_i} - \frac{\partial u}{\partial x_i} \right) dx \geq \int_{\Omega} f(v - u)dx, \quad \forall v \in K_1, \tag{58}$$

where

$$K_1 = \left\{ v \in W_0^{1,\infty}(\Omega) : 0 \leq v(x) \leq \delta(x) = dist(x, \partial\Omega) \text{ a.e. on } \Omega \right\}.$$

Finally, we are able to prove that the elastic region coincides with the set where u does not touch the obstacle, namely

Theorem 17 *Under the same assumptions on Ω, a and f as in Theorem 16, setting*

$$I = \{x \in \Omega : u(x) = \delta(x)\},$$

$$\Lambda = \{x \in \Omega : u(x) < \delta(x)\}$$

it results

$$P = \{x \in \Omega : |Du| = 1\} = I,$$

$$E = \{x \in \Omega : |Du| < 1\} = \Lambda.$$

In order to prove Theorem 16, we should apply strong duality theory in the case $X = S = H_0^1(\Omega)$, but in this case, as we already observed, the ordering cone $C = \{w \in L^1(\Omega) \ : \ w(t) \geq 0 \text{ a.e. } \ \text{in } \Omega\}$ has an empty interior, then the classical strong duality theory cannot be applied.

It is necessary to use the new strong duality theory described in Sect. 1.1.

To this end, let us consider variational inequality (47) under assumption (56) and let $u \in K$ be the solution to (47). From the regularity results in [7] it follows that, if $F \in L^p(\Omega)$, $1 < p < \infty$, u belongs to $W^{2,p}(\Omega) \cap K$. In particular, if $p > n$, Du belongs to $C^{0,\alpha}(\overline{\Omega})$. From Theorem 14, it follows that u is a solution to problem (49). Since strong monotonicity holds and u is regular, it also solves the problem

$$\text{Find } u \in K_\delta^1 : \int_\Omega \sum_{i=1}^n a_i(Du) \left(\frac{\partial v}{\partial x_i} - \frac{\partial u}{\partial x_i} \right) dx \geq \int_\Omega f(v-u)dx, \quad \forall v \in K_\delta^1, \tag{59}$$

where

$$K_\delta^1 = \left\{ v \in H_0^1(\Omega) : |v(x)| \leq \delta(x) \text{ a.e. on } \Omega \right\}.$$

Moreover, since f is positive and a monotone, with $a(0) = 0$, u is also the solution to

$$\text{Find } u \in K_2 : \int_\Omega \sum_{i=1}^n a_i(Du) \left(\frac{\partial v}{\partial x_i} - \frac{\partial u}{\partial x_i} \right) dx \geq \int_\Omega f(v-u)dx, \quad \forall v \in K_2, \tag{60}$$

where

$$K_2 = \left\{ v \in H_0^1(\Omega) : 0 \leq v(x) \leq \delta(x) \text{ a.e. in } \Omega \right\}.$$

Now, we may rewrite problem (60) as an optimization problem. Let us set

$$f(v) = \int_\Omega (Au - F)\,(v-u)\,dx, \quad v \in K_2, \tag{61}$$

where

$$A = -\sum_{i=1}^n \frac{\partial a_i(Du)}{\partial x_i}.$$

As we already observed, $u \in K_2 \cap W^{2,p}(\Omega)$ is the solution of (60) and let us remark that

$$\min_{v \in K_2} f(v) = f(u) = 0. \tag{62}$$

We are able to prove that, assuming

$$X = Y = L^2(\Omega),\ C = C^* = \left\{ v \in L^2(\Omega) : v(x) \geq 0 \text{ a.e. in } \Omega \right\},\ g(v) = v - \delta,$$

the optimization problem (62) fulfills *Assumption S*. Then the strong duality and Theorem 2 hold, from which we get, via variational arguments, that there exists $\overline{\mu} \in C$ such that

$$Au - F + \overline{\mu} = 0 \ \text{ a.e. in } \Omega, \tag{63}$$

$$\overline{\mu}(x)(u(x) - \delta(x)) = 0 \ \text{ a.e. in } \Omega. \tag{64}$$

From Theorem 17, that is achieved using delicate tools of nonlinear partial differential equations, conditions (57) follow.

Remark 3 Another way to reach the strong duality is to verify Assumption NES. Indeed, in this particular setting, our map

$$\varphi : L^2(\Omega) \longrightarrow \overline{\mathbb{R}}$$

is defined by

$$\varphi(\alpha) = \inf_{\substack{v \in H_0^1(\Omega) \\ 0 \leq v \leq \delta + \alpha}} \int_\Omega (Au - F)(v - u)dx. \tag{65}$$

In particular, in virtue of (61),

$$\varphi(\theta_{L^2(\Omega)}) = \inf_{\substack{v \in H_0^1(\Omega) \\ 0 \leq v \leq \delta}} f(v) = \inf_{v \in K_2} f(v) = f(u) = 0, \tag{66}$$

then it results

$$\partial\varphi(\theta_{L^2(\Omega)}) = \left\{ \varphi^* \in L^2(\Omega) : \varphi(\alpha) \geq \langle \varphi^*, \alpha \rangle \ \forall \alpha \in L^2(\Omega) \right\}, \tag{67}$$

and we are able to prove that, setting $\mu = (A\delta - f) \cdot \chi_{\{x \in \Omega:\ u(x) = \delta(x)\}}$,

$$\mu \in \partial\varphi(\theta_{L^2(\Omega)}).$$

In any case, the strong duality holds.

3.4 Von Mises Functions

We now provide an example of the so-called "Von Mises functions", namely of solutions of the elastic-plastic torsion problem, associated to nonlinear monotone operators, which are not obtained by means of the obstacle problem in the case $F = constant$.

We consider an operator $a(p) : \mathbb{R}^n \to \mathbb{R}^n$, of class C^2, strictly monotone. Let $\Omega \subseteq \mathbb{R}^n$ with boundary $\partial\Omega \in C^{2,1} = W^{3,\infty}$, $P = \Gamma_\mu = \{x \in \Omega : \delta(x) = d(x, \partial\Omega) < \mu\}$, $E = \Omega \setminus P$.

As it is well known μ can be chosen in such a way that for every $x \in \Gamma_\mu$ there is a unique closest point from $\partial\Omega$ to x and $\delta(x)$ owns the same regularity of $\partial\Omega$ on Γ_μ. Then $\delta(x) \in W^{3,p}(P)$, $\forall p > 1$, and its trace $\delta_{/\partial P} \in W^{3-1/p,p}(\partial P)$.

Let

$$F(x) = \sum_{i=1}^{n} D_i a_i(D\delta) - \Delta\delta(x) \text{ a.e. in } P$$

and $w(x) \in W^{3,p}(E)$, $\forall p > 1$, the solution of

$$\begin{cases} \sum_{i=1}^{n} D_i a_i(Dw) = 0 & \text{a.e. in } E \\ w(x) = \delta(x) & on\ \partial E. \end{cases}$$

We can directly prove that, in E, $G(Dw) = |Dw|^2 - 1$ verifies

$$\sum_{i,j} \frac{\partial}{\partial x_i}\left[\frac{\partial a_i(Dw)}{\partial p_j}\,\frac{\partial}{\partial x_j} G(Dw)\right] \geq 0.$$

Then we may apply maximum principle to $G(Dw)$, from which it follows $|Dw| < 1$ in E.

The function $u(x) \in W^{2,p}(\Omega)$, $\forall p > 1$,

$$u(x) = \begin{cases} \delta(x) & x \in P \\ w(x) & x \in E \end{cases}$$

arises.

Setting

$$\tilde{F}(x) = \begin{cases} F(x) & x \in P \\ 0 & x \in E, \end{cases}$$

it results

$$\sum_{i=1}^{n} D_i a_i(Du) - \tilde{F}(x) = \begin{cases} \Delta\delta(x) & x \in P \\ 0 & x \in E, \end{cases}$$

namely

$$\sum_{i=1}^{n} D_i a_i(Du) - \tilde{F}(x) = \sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(\chi_P(x)\frac{\partial u}{\partial x_i}\right) \quad \text{a.e. in } \Omega.$$

Moreover

$$\chi_P(x)(|Du|^2 - 1) = 0,$$

and by means of maximum principle we can prove

$$|Du| \le 1,$$

that is u is a solution of an elastic-plastic torsion problem.

3.5 Radial Solutions

Finally, we search for radial solutions to the elastic-plastic torsion problem, assuming the free term to belong to $L^p(\Omega)$ (see [33]). In particular, for $n = 2$, we investigate the nature of the torsion and when the transition from the elastic case to the plastic one happens. We are able to find the explicit solution $u \in W^{2,p}(\Omega)$ and the Lagrange multiplier $\overline{\mu} \in L^p(\Omega)$ in the two admissible cases, namely, when the elastic and the plastic regions both exist and when the torsion is only elastic. Moreover, we characterize the free boundary and obtain a necessary and sufficient condition in order that the plastic region exists. Finally, we provide some examples.

To this aim, let us assume that Ω is the ball of $\mathbb{R}^n$ of radius 1 centered at the origin, and $F \in L^p(\Omega)$, $p > n$, is of radial type, namely $F(x) = f(|x|) = f(\rho)$, with $|x| = \rho$.

We search for solutions to (43) such that $u(x) \in W^{2,p}(\Omega)$ and $\overline{\mu}(x) \in L^p(\Omega)$ are of radial type, namely $\overline{\mu}(x) = \mu(|x|) = \mu(\rho)$, $u(x) = \varphi(|x|) = \varphi(\rho)$. In this case, since $u(x) = \varphi(\rho)$, $\frac{\partial u}{\partial x_i} = \varphi'(\rho)\frac{x_i}{\rho}$, $\Delta u = \varphi'' + \frac{n-1}{\rho}\varphi'(\rho)$, $|Du| = \varphi'(\rho)$, bearing in mind that $u \in K$, conditions (43) become

$$\begin{cases} |\varphi'(\rho)| = 1; \;\; \mu(\rho) \ge 0 \quad a.e. \; in \; [0, 1]; \\ \mu(\rho)\left(1 - |\varphi'(\rho)|\right) = 0 \quad a.e. \; in \; [0, 1]; \\ -\varphi''(\rho) - \dfrac{n-1}{\rho}\varphi'(\rho) - \displaystyle\sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(\mu \frac{\partial u}{\partial x_i}\right) = f(\rho). \end{cases} \tag{68}$$

Under the following assumptions: there exists $\overline{\rho} \in (0, 1)$ such that

$$\frac{\int_{C_{\overline{\rho}}(0)} F(x)dx}{|\partial C_{\overline{\rho}}(0)|} = 1, \tag{69}$$

where $C_{\overline{\rho}}(0)$ is the closed ball of radius $\overline{\rho}$ centered at the origin, namely

$$\int_0^{\overline{\rho}} \rho f(\rho) d\rho = \overline{\rho}, \tag{70}$$

and

$$\rho f(\rho) \geq 0 \textit{ is a nondecreasing function in } [0, 1], \tag{71}$$

we are able to prove the following result.

Theorem 18 *Under conditions* (69), (71)*, the region* $[0, \overline{\rho}]$ *is an elastic region and the region* $[\overline{\rho}, 1]$ *is a plastic region. Moreover, the solution* φ *to* (68) *is*

$$\varphi(\rho) = \begin{cases} 1 - \overline{\rho} + \displaystyle\int_\rho^{\overline{\rho}} \frac{1}{t} \int_0^t \sigma f(\sigma) d\sigma dt & \rho \in [0, \overline{\rho}] \\ 1 - \rho & \rho \in (\overline{\rho}, 1], \end{cases} \tag{72}$$

and it results to be $\varphi(\rho) \in W^{2,p}(0, 1)$ *and* $\mu(\rho) \in L^p(0, 1)$.

If Eq. (70) does not admit any solution $\overline{\rho} \in (0, 1)$, namely $\forall \rho \in (0, 1)$

$$\frac{1}{\rho} \int_0^\rho \sigma f(\sigma) d\sigma < 1 \quad \textit{or} \quad \frac{1}{\rho} \int_0^\rho \sigma f(\sigma) d\sigma > 1,$$

the plastic region does not exist. The case

$$\frac{1}{\rho} \int_0^\rho \sigma f(\sigma) d\sigma > 1 \quad \forall \rho \in (0, 1)$$

is not admissible, since it implies

$$\varphi'(\rho) < -1 \quad \forall \rho \in (0, 1).$$

Then, we are able to prove the following result.

Theorem 19 *Under condition* (71)*, if*

$$\frac{1}{\rho} \int_0^\rho \sigma f(\sigma) d\sigma < 1 \quad \forall \rho \in (0, 1),$$

then, [0,1] is an elastic region. Moreover, the solution φ *to* (68) *is*

$$\varphi(\rho) = \int_\rho^1 \frac{1}{t} \int_0^t \sigma f(\sigma) d\sigma dt \quad \forall \rho \in [0, 1]. \tag{73}$$

It results to be $\varphi(\rho) \in W^{2,p}(0, 1)$.

Example 3 Let us consider a first example, namely $F = const = k > 0$. In this case we obtain the same results as in [48], p. 15.

If we consider a first case, $F = k > 2$, the plastic region exists, since

$$\lim_{\rho \to 0^+} \int_{\rho}^{\overline{\rho}} k\sigma d\sigma = \frac{k}{2}\overline{\rho}^2,$$

namely, $\overline{\rho} = \frac{2}{k} < 1$ is the solution to (70).

Then, by (72) we get the continuous function

$$\varphi(\rho) = \begin{cases} \frac{k}{4}\left[(1-\rho^2) - (1-\frac{2}{k})^2\right] & in\ E = [0, \frac{2}{k}) \\ 1-\rho & in\ P = [\frac{2}{k}, 1]. \end{cases}$$

It is easily seen that $u \in W^{2,p}(\Omega)$.

Moreover, the Lagrange multiplier $\mu(\rho)$ exists and belongs to $L^p([0, 1])$:

$$\mu(\rho) = \begin{cases} k\frac{\rho}{2} - 1 \geq 0 \ in\ P = [\frac{2}{k}, 1] \\ 0 \qquad\qquad \in\ E = [0, \frac{2}{k}). \end{cases}$$

If we consider the other case $F = const = k$, $0 < k \leq 2$, the plastic region does not exist, since $\overline{\rho} = \frac{2}{k} \geq 1$ is the solution to (70).

Then, the torsion is all elastic and by (73) we get the continuous function

$$\varphi(\rho) = \frac{k}{4}(1-\rho^2). \tag{74}$$

$\varphi(\rho)$ as in (74) and $\mu = 0$ verify conditions (68) in [0, 1]. Moreover $u \in W^{2,p}(\Omega)$.

Let us now consider problem (68) with $f(\rho) = \dfrac{k}{\rho^\alpha}$, $0 < \alpha < 1$.

The condition $\alpha < 1$ ensures that $F(x) \in L^p(\Omega)$, $2 = n < p < \frac{2}{\alpha}$. Moreover, condition (71) is verified.

If we consider the case $k > 2 - \alpha$, the plastic region exists, since

$$\lim_{\rho \to 0^+} \int_{\rho}^{\overline{\rho}} \sigma f(\sigma)\, d\sigma = \frac{k}{2-\alpha}\overline{\rho}^{2-\alpha}$$

namely, $\overline{\rho} = \left(\frac{2-\alpha}{k}\right)^{\frac{1}{1-\alpha}} < 1$ is the solution to (70).

Then, by (72) we get the continuous function

$$\varphi(\rho) = \begin{cases} 1 - \left(\dfrac{2-\alpha}{k}\right)^{\frac{1}{1-\alpha}} + \dfrac{k}{(2-\alpha)^2}\left(\dfrac{2-\alpha}{k}\right)^{\frac{2-\alpha}{1-\alpha}} - \dfrac{k}{(2-\alpha)^2}\rho^{2-\alpha} & in\ E = [0, \overline{\rho}) \\ 1-\rho & in\ P = [\overline{\rho}, 1]. \end{cases}$$

It is easily seen that $u \in W^{2,p}(\Omega)$.

Moreover, the Lagrange multiplier $\mu(\rho)$ exists and belongs to $L^p([0, 1])$:

$$\mu(\rho) = \begin{cases} \dfrac{k}{2-\alpha}\rho^{1-\alpha} \; - \; 1 \geq 0 \; in \; P = (\overline{\rho}, 1] \\ 0 \qquad\qquad\qquad\quad in \; E = [0, \overline{\rho}]. \end{cases}$$

Finally, if we consider the other case $0 < k \leq 2 - \alpha$, the plastic region does not exist, since $\overline{\rho} = \left(\frac{2-\alpha}{k}\right)^{\frac{1}{1-\alpha}} \geq 1$ is the solution to (70).

Then, the torsion is all elastic and by (73) we get the continuous function

$$\varphi(\rho) = \frac{k}{(2-\alpha)^2}(1 - \rho^{2-\alpha}) \quad \forall \rho \in [0, 1]. \tag{75}$$

$\varphi(\rho)$ as in (75) and $\mu = 0$ verify conditions (68) in [0, 1]. Moreover, $u \in W^{2,p}(\Omega)$.

References

1. A. Barbagallo, A. Maugeri, Duality theory for a dynamic oligopolistic market equilibrium problem. Optimization **60**, 29–52 (2011)
2. A. Barbagallo, P. Daniele, S. Giuffrè, A. Maugeri, Variational approach for a general financial equilibrium problem: the deficit formula, the balance law and the liability formula. A path to the economy recovery. Eur. J. Oper. Res. **237**(1), 231–244 (2014)
3. J.M. Borwein, V. Jeyakumar, A.S. Lewis, M. Wolkowicz, Constrained approximation via convex programming. University of Waterloo. Preprint (1988)
4. R.I. Bot, E.R. Csetnek, A. Moldovan, Revisiting some duality theorems via the quasirelative interior in convex optimization. J. Optim. Theory Appl. **139**(1), 67–84 (2008)
5. H. Brezis, Moltiplicateur de Lagrange en Torsion Elasto-Plastique. Arch. Rational Mech. Anal. **49**, 32–40 (1972)
6. H. Brezis, Problèmes Unilatéraux. J. Math. Pures Appl. **51**, 1–168 (1972)
7. H. Brezis, G. Stampacchia, Sur la régularité de la solution d'inéquations elliptiques. Bull. Soc. Math. Fr. **96**, 153–180 (1968)
8. V. Caruso, P. Daniele, A network model for minimizing the total organ transplant costs. Eur. J. Oper. Res. (2017). https://doi.org/10.1016/j.ejor.2017.09.040
9. V. Chiadó-Piat, D. Percivale, Generalized Lagrange multipliers in elastoplastic torsion, J. Differ. Equ. **114**, 570–579 (1994)
10. M.G. Cojocaru, P. Daniele, A. Nagurney, Projected dynamical systems and evolutionary variational inequalities via Hilbert spaces and applications. J. Optim. Theory Appl. **127**, 549–563 (2005)
11. P. Daniele, *Dynamic Networks and Evolutionary Variational Inequalities* (Edward Elgar Publishing, Cheltenham, 2006)
12. P. Daniele, Evolutionary variational inequalities and applications to complex dynamic multi-level models. Transp. Res. Part E **46**, 855–880 (2010)
13. P. Daniele, S. Giuffrè, General infinite dimensional duality and applications to evolutionary network equilibrium problems. Optim. Lett. **1**, 227–243 (2007)
14. P. Daniele, S. Giuffrè, Random variational inequalities and the random traffic equilibrium problem. J. Optim. Theory Appl. **167**(1), 363–381 (2015)

15. P. Daniele, S. Giuffrè, S. Pia, Competitive financial equilibrium problems with policy interventions. J. Ind. Manag. Optim. **1**(1), 39–52 (2005)
16. P. Daniele, S. Giuffrè, G. Idone, A. Maugeri, Infinite dimensional duality and applications. Math. Ann. **339**, 221–239 (2007)
17. P. Daniele, S. Giuffrè, A. Maugeri, Remarks on general infinite dimensional duality with cone and equality constraints. Commun. Appl. Anal. **13**(4), 567–578 (2009)
18. P. Daniele, S. Giuffrè, M. Lorino, A. Maugeri, C. Mirabella, Functional inequalities and analysis of contagion in the financial networks, in *Handbook of Functional Equations – Functional Inequalities*, ed. by Th.M. Rassias. Optimization and Its Applications, vol. 95 (Springer, Berlin, 2014), pp. 129–146
19. P. Daniele, S. Giuffrè, A. Maugeri, F. Raciti, Duality theory and applications to unilateral problems. J. Optim. Theory Appl. **162**(3), 718–734 (2014)
20. P. Daniele, S. Giuffrè, M. Lorino, Functional inequalities, regularity and computation of the deficit and surplus variables in the financial equilibrium problem. J. Glob. Optim. **65**, 575–596 (2016)
21. P. Daniele, M. Lorino, C. Mirabella, The financial equilibrium problem with a Markowitz-type memory term and adaptive, constraints. J. Optim. Theory Appl. **171**, 276–296 (2016)
22. P. Daniele, A. Maugeri, A. Nagurney, Cybersecurity investments with nonlinear budget constraints: analysis of the marginal expected utilities, in *Operations Research, Engineering, and Cyber Security*, ed. by N.J. Daras, T.M. Rassias. Springer Optimization and Its Applications, vol. 113 (Springer, Berlin, 2017), pp. 117–134
23. M.B. Donato, The infinite dimensional Lagrange multiplier rule for convex optimization problems. J. Funct. Anal. **261**(8), 2083–2093 (2011)
24. M.B. Donato, A. Maugeri, M. Milasi, C. Vitanza, Duality theory for a dynamic Walrasian pure exchange economy. Pac. J. Optim. **4**, 537–547 (2008)
25. S. Giuffrè, Strong solvability of boundary value contact problems. Appl. Math. Optim. **51**(3), 361–372 (2005)
26. S. Giuffrè, Elements of duality theory, in *Topics in Nonlinear Analysis and Optimization*, ed. by Q.H. Ansari (World Education, Delhi, 2012), pp. 251–267
27. S. Giuffrè, S. Pia, Weighted traffic equilibrium problem in non pivot Hilbert spaces with long term memory, in *AIP Conference Proceedings Rodi*, September 2010, vol. 1281, pp. 282–285
28. S. Giuffrè, A. Maugeri, New results on infinite dimensional duality in elastic-plastic torsion. Filomat **26**(5), 1029–1036 (2012)
29. S. Giuffrè, A. Maugeri, Lagrange multipliers in elastic-plastic torsion, in *AIP Conference Proceedings Rodi*, September 2013, vol. 1558, pp. 1801–1804
30. S. Giuffrè, A. Maugeri, A measure-type Lagrange multiplier for the elastic-plastic torsion. Nonlinear Anal. **102**, 23–29 (2014)
31. S. Giuffrè, G. Idone, A. Maugeri, Duality theory and optimality conditions for generalized complementary problems. Nonlinear Anal. **63**, e1655–e1664 (2005)
32. S. Giuffrè, A. Maugeri, D. Puglisi, Lagrange multipliers in elastic-plastic torsion problem for nonlinear monotone operators. J. Differ. Equ. **259**(3), 817–837 (2015)
33. S. Giuffrè, A. Pratelli, D. Puglisi, Radial solutions and free boundary of the elastic-plastic torsion problem. J. Convex Anal. **25**(2), 529–543 (2018)
34. J. Gwinner, F. Raciti, Random equilibrium problems on networks. Math. Comput. Model. **43**, 880–891 (2006)
35. J. Gwinner, F. Raciti, On a class of random variational inequalities on random sets. Numer. Funct. Anal. Optim. **27**, 619–636 (2006)
36. R.B. Holmes, *Geometric Functional Analysis* (Springer, Berlin, 1975)
37. G. Idone, A. Maugeri, Generalized constraints qualification and infinite dimensional duality. Taiwan. J. Math. **13**, 1711–1722 (2009)
38. J. Jahn, *Introduction to the Theory of Nonlinear Optimization*, 3rd edn. (Springer, Berlin, 2007)
39. V. Jeyakumar, H. Wolkowicz, Generalizations of slater constraint qualification for infinite convex programs. Math. Program. **57**, 85–101 (1992)
40. H.M. Markowitz, Portfolio selection. J. Financ. **7**, 77–91 (1952)

41. H.M. Markowitz, *Portfolio Selection: Efficient Diversification of Investments* (Wiley, New York, 1959)
42. A. Maugeri, D. Puglisi, A new necessary and sufficient condition for the strong duality and the infinite dimensional Lagrange Multiplier rule. J. Math. Anal. Appl. **415**(2), 661–676 (2014)
43. A. Maugeri, D. Puglisi, Non-convex strong duality via subdifferential. Numer. Funct. Anal. Optim. **35**, 1095–1112 (2014)
44. A. Maugeri, D. Puglisi, On nonlinear strong duality and the infinite dimensional Lagrange multiplier rule. J. Nonlinear Convex Anal. **18**(3), 369–378 (2017)
45. A. Maugeri, F. Raciti, Remarks on infinite dimensional duality. J. Glob. Optim. **46**, 581–588 (2010)
46. A. Maugeri, L. Scrimali, New approach to solve convex infinite-dimensional bilevel problems: application to the pollution emission price problem. J. Optim. Theory Appl. **169**(2), 370–387 (2016)
47. R.T. Rockafellar, Conjugate duality and optimization, in *Conference Board of the Mathematical Science Regional Conference Series in Applied Mathematics*, vol. 16 (Society for Industrial and Applied Mathematics, Philadelphia, 1974)
48. J.F. Rodrigues, *Obstacle Problems in Mathematical Physics*. Mathematics Studies, vol. 134 (Elsevier, Amsterdam, 1987)
49. L. Scrimali, Infinite dimensional duality theory applied to investment strategies in environmental policy. J. Optim. Theory Appl. **154**, 258–277 (2012)
50. T.W. Ting, Elastic-plastic torsion of a square bar. Trans. Am. Math. Soc. **113**, 369–401 (1966)
51. T.W. Ting, Elastic-plastic torsion problem II. Arch. Ration. Mech. Anal. **25**, 342–366 (1967)
52. T.W. Ting, Elastic-plastic torsion problem III. Arch. Ration. Mech. Anal **34**, 228–244 (1969)
53. R. Von Mises, Three remarks on the theory of the ideal plastic body, in *Reissner Anniversary Volume* (Edwards, Ann Arbor, 1949)

Selective Priorities in Processing of Big Data

Nicholas J. Daras

1 Introduction

The aim of the present paper is to document a quantitative systemic modeling for the processing of big data flow. Since, according to official calculations, the total global flow of data exceeds 150 million petabytes annual rate, or nearly 500 exabytes per day, it is very clear that the ever-increasing volume of data will soon cause great difficulty in the efficient processing of information and will make extremely difficult task of processing the data flow.

In order to urgently overcome this obstacle, a good idea seems to be ***the appropriate choice of data amounts***. To this direction, this paper studies a reasonable question which arises and may be constitute a central subject of discussion in subsequent additional scientific studies. The question relates to the *preference of choices and priorities in the processing of big data*. Equivalently, *if each one of a group of data processors prefers to be limited to different sets of data amounts from a collection of big data, then how much the different priorities of processing could lead to equilibrium situations or contrasts*?

The paper is divided in two parts. The first part examines the case of a single data processor. Obviously, for each data entity in the domain of his competence, the processor can choose or use only an amount of data. Thus, in Sect. 2.1, we will describe how through its options, the data processor may prefer to focus only on some choices. A program of data selection for the processor specifies the data amount of each entity that the processor may take into account. Then, in Sect. 2.2, we will study the selectivity display of a processor in order to actually exploit a certain amount of data from another. A *data selection preference* is the relation

N. J. Daras (✉)
Department of Mathematics and Engineering Sciences, Hellenic Military Academy, Vari Attikis, Greece
e-mail: ndaras@sse.gr

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_6

that determines such any selectivity. In order to establish a well such preference, in Sect. 2.3, we will show how a processor should associate certain significance in each component of the system (vector) of times of data processing, while in Sect. 2.4 we will study the topology of the space of data selection preferences and we shall describe neighboring preferences of a given data selection preference. Having regard to all these, in the next Sect. 2.5 we will investigate the lower hemicontinuity of the relation defining the set of all *rational choices for the data amounts*, and in Sect. 2.6 we will deal with the concept of the *mean rational data amount choice for a set of data processors*. The second part of the paper is devoted to the case of several data processors. In this case, each of the processors has its own priorities and preferences, and, after a brief introduction, we will see that there are cores and equilibriums of contrasts, the study of which may provide useful information (Sects. 3.3 and 3.4).

2 Rational Choice of Data Sets

2.1 Programs of Data Selections

We begin by recalling some basic definitions.

Definition 2.1

(i) A ***data entity*** (index, concept, term, thing, etc.) in a given data system (or data complex) S is something that exists by itself, **although it need not be of material existence** (http://www.thefreedictionary.com/entity).

(ii) A ***measurable data entity*** in S is anything that can be measured in S [1].It is assumed that all entities over the given system S are distinguishable, measurable and indicated by an index i running from 1 to $\ell \in \mathbb{N} \cup \{\infty\}$.

(iii) The ***amount*** d_i ***of a measurable data entity*** i over S can be expressed by a natural number.

In what follows, without loss of generality, we will always assume that

1. *the amount of each data entity in a given system S takes values in the set* $\mathbb{R}$ *of the real numbers* and
2. *any unit vector*

$$1_{d_i} = \left(\underbrace{0, \ldots, 0, 1, 0, \ldots, 0}_{i\text{-}position} \right) \textit{ of } \mathbb{R}^{\ell}$$

is identified with one unit of the data amount d_i $(i = 1, 2, \ldots, \ell)$.

We point out that the first of these two assumptions does not contradict the Definition 2.1(iii), only facilitates, in a meaningful way, any technical documentation of the considerations that will follow in the paper, because the embedding in the larger space of real numbers gives greater capabilities and allows an easier processing.

Under these assumptions, we give the following two definitions.

Definition 2.2 The linear space $\mathbb{R}^{\ell}$, endowed with the corresponding product Euclidean topology, is called a ***continuous space of data amounts*** over the system S. Every bundle of data amounts $(d_1, d_2, \dots, d_{\ell})$ can be represented by a point in the measurable space $\mathbb{R}^{\ell}$ of data amounts over the system S.

Definition 2.3 For a *data processor* $\mathcal{M}$, a ***program of data selection*** over the system S specifies the data amount of ***each*** data entity in S that $\mathcal{M}$ takes into account, as well as the data amount of this entity which he will make available. We shall use the convention that the data amount of a data entity in S which has to be made available **by** the processor is represented by a negative number, while the amount of a data entity over S which has to be made available to the processor is represented by a positive number. Then, ***every program of data selection can be represented by an element***

$$x = (d_1, d_2, \dots, d_{\ell})$$

in the continuous measurable space $\mathbb{R}^{\ell}$ *of data amounts over the system* S.

Remark 2.1 It is obvious that every element in $\mathbb{R}^{\ell}$ can be interpreted meaningfully as a program of data selection.

It is assumed that *for every data processor* $\mathcal{M}$ *there is a nonempty closed subset* $\mathcal{X}_{\mathcal{M}}$ *in* $\mathbb{R}^{\ell}$, the ***focal data set of*** $\mathcal{M}$, or simply the ***focal set of*** $\mathcal{M}$, which describes the set of a priori possible programs of data selection over the system S. Here a priori possible means that, ignoring processing acts, the data processor can carry out the program of data selection over the system S. More specifically, we have the following.

Definition 2.4 A ***focal data set*** or simply ***focal set*** $\mathcal{X}$ over the system S is a nonempty subset of the data entity set over S which is closed, convex and bounded from below. Given a vector $b \in \mathbb{R}^{\ell}$ and a compact subset $E \subset \mathbb{R}^{\ell}$, we denote by $\mathcal{X}_{b;;E}$ the compact set of all focal sets $\mathcal{X}$ such that $b \in X$ and $X \cap E \neq \emptyset$.

Remark 2.2 A focal set over the system S will typically belong to a discrete (not necessarily finite) set in $\mathbb{R}^{\ell}$.

2.2 Data Selection Preferences

Definition 2.5 We say that ***a data processor*** $\mathcal{M}$ ***selects the program of data selection*** x ***instead of the program of data selection*** $x^{'}$ if he wants to select x whenever he is offered the alternatives x and $x^{'}$.

The binary relation "*selected*" becomes a powerful tool for modeling analysis if the behavior of the data processors *reveals* a certain 'consistency' of choices.

Definition 2.6 A ***data selection preference***, or simply ***data preference***, in the complex S is a pair $(\mathfrak{X}, \succ)$, where

1. $\mathfrak{X}$ is a focal set over S and
2. $\succ \mathfrak{X} \times \mathfrak{X}$ is a transitive and non-reflexive binary relation on $\mathfrak{X}$ such that $\succ$ is open in $\mathfrak{X} \times \mathfrak{X}$.

In what follows, instead of $(x, y) \in \succ$, we shall write

$$x \succ y.$$

Thus,

$$x \not\succ y \text{ means } (x, y) \in \not\succ .$$

Sometimes it is convenient to represent a data selection preference by an $\mathbb{R}$-valued function. Thus, we can give the following

Definition 2.7 Given a $(\mathfrak{X}, \succ)$, a ***data preference representation*** in S is a continuous function

$$u : \mathfrak{X} \to \mathbb{R}$$

such that
$x \succ y$ if and only if $u\,(x) > u\,(y)$.

Remark 2.3 It is well known that *if E is a compact subset of $\mathbb{R}^{\ell}$, then the set $\mathbb{P}\,(E)$ of all nonempty closed subsets of E together with the Hausdorff distance on E is a compact metric space*. We will assume that

1. $\mathfrak{K}\,(\subset \mathbb{P}\,(E))$ *is a compact subset of focal sets* $\mathfrak{X} \subset E$.

Notice that the particular choice of E is immaterial. To restrict in this way the "universe" of focal sets X is no restriction for our analysis; however, it simplifies the mathematical presentation, since the set $\mathfrak{K}$ will turn out to be compact.

Notation 2.1

1. *The set of all data selection preferences* $(\mathfrak{X}, \succ)$ *in* S *with* $\mathfrak{X} \in \mathfrak{K}$ *is denoted by*

$$\mathcal{P} = \mathcal{P}_{\mathfrak{K}}.$$

2. *The set of all data selection preferences* $(\mathfrak{X}, \succ)$ *in* S *with* $\mathfrak{X} \in \mathfrak{X}_{y;;E}$ *is denoted by*

$$\mathcal{P}_{y;;E}.$$

It is easy to verify the following result.

Proposition 2.1 *To every data selection preference* $(\mathcal{X}, \succ) \in \mathcal{P}$ *we associate the set*

$$\mathcal{F} := \{(x, y) \in E \times E : x \in \mathcal{X},\ y \in \mathcal{X} \text{ and } x \not\succ y\}.$$

The set $\mathcal{F}$ *is characterized by the system of the following four properties.*

1. $\mathcal{F}$ *is a closed subset in* $E \times E$.
2. *The set* $\{x \in E : \textit{there is a y with } (x, y) \in \mathcal{F}\}$ *belongs to the compact set* $\mathfrak{K}$.
3. $(x, y) \in \mathcal{F}$ *implies* $(x, x) \in \mathcal{F}$ *and* $(y, y) \in \mathcal{F}$.
4. $(x, y) \notin \mathcal{F}$ *and* $(y, z) \notin \mathcal{F}$ *implies* $(x, z) \notin \mathcal{F}$.

Conversely, given such a set $\mathcal{F}$, *we obtain the corresponding data selection preference* $(\mathcal{X}, \succ) \in \mathcal{P}$ *by setting*
$\mathcal{X} : \{x \in E : (x, x) \in \mathcal{F}\ \}$ *and* $\succ$ $(\mathcal{X} \times \mathcal{X})\ \{\mathcal{F}\}$.

In order to investigate the behavior of the modeling process, it is often required additional properties of the data selection preferences. For this purpose, we will now define some useful auxiliary subsets of $\mathcal{P}$.

Definition 2.8 Let $(\mathcal{X}, \succ) \in \mathcal{P}$ be a given data selection preference.

(i) $(\mathcal{X}, \succ)$ is said to be ***locally non-satiated*** in the complex S if for each $x \in \mathcal{X}$ and each neighborhood $U = U_x$ of x there exists a $x^{'} \in \mathcal{X} \cap U$ such that $x^{'} \succ x$. The set of all locally non-satiated data selection preferences in $\mathcal{P}$ is denoted by

$$\mathcal{P}_{lns}.$$

(ii) $(\mathcal{X}, \succ)$ is said to be ***monotonic*** in S if $0 \leq x \leq y$ and $x \neq y$ in $\mathcal{X}$ imply $y \succ x$. The set of all monotonic data selection preferences in $\mathcal{P}$ is denoted by

$$\mathcal{P}_{mo}.$$

(iii) $(\mathcal{X}, \succ)$ is said to be ***negatively transitive*** in S if for every $x, y, z \in \mathcal{X}$ with $x \not\succ y$ and $y \not\succ z$ we have $x \not\succ z$. The set of all negatively transitive data selection preferences in $\mathcal{P}$ is denoted by
$\mathcal{P}^{*}$.

For a data selection preference in $\mathcal{P}^{*}$ one defines the ***data selection indifference*** in S by
$x\ y$ if and only if $x \not\succ y$ and $y \not\succ x$.

The indifference relation on $\mathcal{X}$ is reflexive, transitive and symmetric. The relation $\not\succ$ is then written as $\precsim$. Obviously, ***the data selection indifference*** $\precsim$ ***is reflexive, transitive and complete***.

Definition 2.9 The data selection indifference $\big(\mathcal{X}, \precsim\big) \in \mathcal{P}^{*}$ is called:

1. ***convex*** in the complex S if for every $z \in \mathcal{X}$, the set $\big\{x \in \mathcal{X} : z \precsim x\big\}$ is convex and

2. ***strongly convex*** in S if for every $x\ x^{'}$, $x \neq x^{'}$, and every $0 < \lambda < 1$ it follows that $\lambda x + (1-\lambda)\,x^{'} \succ x$.
3. The set of all convex (strongly convex) data selection preferences in $\mathcal{P}^{*}$ is denoted by

$\mathcal{P}^{*}_{co}.\ (\mathcal{P}^{*}_{sco})$.

2.3 Weighted Data Systems and Data Amount Processing Capacities

A ***weighted data system*** w ***in*** S associates to every measurable data entity i in S two numerical values: its ***weight*** w_i and its ***balancing evaluation*** b_i. The concept of the weight for the data entity i depends upon the importance attributed to this entity by the manager of the data processor. Regarding the concept of balancing evaluation b_i, this means that b_i/b_{j} is the amount of available data d_j for the weighted entity j in order to obtain one unit of data amount for the weighted entity i.

Definition 2.10 Hereafter, for the weight or/and the balancing evaluation of a given weighted data entity in a system S, we will use, without any distinction and risk of confusion, the single term ***data significance*** of the entity.

Hence, a weighted data system in a complex S associates to every weighted data entity i in S a real number p_i, its data significance. Thus p can be considered as an element of $\mathbb{R}^{\ell}$.

If a data processor $\mathcal{M}$ in S decides to consider and use the weighted data system with data significance $p = (p_1, p_2, \ldots, p_\ell)$, then any $\mathcal{M}$'s choice of programs of data selection $x = (d_1, \ldots, d_\ell)$ in his data focal set $\mathcal{X}_{\mathcal{M}}$ is further restricted. Indeed,

Definition 2.11 The weighted data system's value px of x cannot exceed a certain number $\mathcal{C}_{\mathcal{M}}$ the ***data amount processing capacity*** of $\mathcal{M}$ in S.

The real number $\mathcal{C}_{\mathcal{M}}$ represents the maximum weighted value of a potential data processing by $\mathcal{M}$. Thus, a data amount processing capacity $\mathcal{C}_{\mathcal{M}}$ in S is typically a function of prevailing weights for the weighted data entities. However, it will be convenient to treat the data amount processing capacity as an independent argument.

Definition 2.12 Let $\mathcal{M}$ be a data processor, with data focal set $\mathcal{X}$ and data amount processing capacity $\mathcal{C}_{\mathcal{M}}$ in S. If $\mathcal{M}$ prefers a weighted data system with significance $p = (p_1, p_2, \ldots, p_\ell)$, we define the ***set of data amount processing options*** of $\mathcal{M}$ in S by
$\mathfrak{B}\,(\mathcal{X}, \mathcal{C}_{\mathcal{M}}, p) := \{x = (d_1, \ldots, d_\ell) \in \mathcal{X} : (p_1, p_2, \ldots, p_\ell)\,(d_1, d_2, \ldots, d_\ell) \leq \mathcal{C}_{\mathcal{M}}\}$.

The program of data selection which actually is chosen in the set of processing options $\mathfrak{B}\,(\mathcal{X}, \mathcal{C}_{\mathcal{M}}, p)$ depends directly on the data selection preferences.

Definition 2.13 Let $\mathcal{M}$ be a data processor, with data selection preference $(\mathfrak{X}, \succ)$ and data amount processing capacity $\mathcal{C}_{\mathcal{M}}$ in S. If $\mathcal{M}$ prefers a weighted data system with data significance p in S, we define the ***set*** $\mathfrak{A} = \mathfrak{A}\,(\mathfrak{X}, \succ, \mathcal{C}_{\mathcal{M}}, p)$ ***of all his rational choices for the data amount*** in S as the set of maximal elements in the set of data amount processing options, i.e.

$$\mathfrak{A}\,(\mathfrak{X}, \succ, \mathcal{C}_{\mathcal{M}}, p) = \Big\{x^* = (d_1, \ldots, d_\ell) \in \mathfrak{B}\,(\mathfrak{X}, \mathcal{C}_{\mathcal{M}}, p) :$$

$$there\ is\ no\ x = (d_1, \ldots, d_\ell) \in \mathfrak{B}\,(\mathfrak{X}, \mathcal{C}_{\mathcal{M}}, p)\ \ with\ x \succ x^*\Big\}.$$

Consequently, $x^* \in \mathfrak{A}\,(\mathfrak{X}, \succ, \mathcal{C}_{\mathcal{M}}, p)$ *if and only if* $x \succ x^*$ *implies* $x > \mathcal{C}_{\mathcal{M}}$.

2.4 Topology of the Space of Data Selection Preferences: Neighboring Selection Preferences

Our next purpose will be to investigate how $\mathcal{M}$'s set of rational choices for the data amount in S depends continuously on his data preference $(\mathfrak{X}, \succ)$, his data amount processing capacity $\mathcal{C}_{\mathcal{M}}$ and weighted data system with data significance p in S.

Surely, the discrete topology on $\mathcal{P}$ allows the correspondences on the set of data rational choices to have continuity properties. However, for clear reasons, we want a topology which is metrizable and separable or even compact.

Theorem 2.1

(i) The set $\mathcal{P}$ of all data selection preferences in the complex S endowed with the topology $\mathcal{T}_{\mathbf{c}losed}$ of closed convergence (http://www.math.kit.edu/iag4/lehre/stochgeom2010s/media/topology.pdf) is compact and metrizable.

(ii) A sequence $(\mathfrak{X}_n, \succ_n)_{n\in\mathbb{N}}$ of data selection preferences in S converges to $(\mathfrak{X}, \succ)$ in $(\mathcal{P}, \mathcal{T}_{\mathbf{c}losed})$ if and only if

$$liminf_{n\to\infty}\ \{(x, y) \in \mathfrak{X}_n \times \mathfrak{X}_n :\ x \not\succ y\} = limsup_{n\to\infty}\ \{(x, y) \in \mathfrak{X}_n \times \mathfrak{X}_n :$$

$$x \not\succ y\} = \{(x, y) \in \mathfrak{X} \times \mathfrak{X} :\ x \not\succ y\}.$$

(iii) The topology $\mathcal{T}_{\mathbf{c}losed}$ of closed convergence on the set $\mathcal{P}$ of data selection preferences in S is the coarsest topology on $\mathcal{P}$ which has the property that the set

$$\Big\{(\mathfrak{X}, \succ,\ x,\ y) \in \mathcal{P} \times \mathbb{R}^\ell \times \mathbb{R}^\ell : x, y \in \mathfrak{X}\ and\ x \not\succ y\Big\}$$

is closed.

Proof

(i) It is well known that the set $\mathcal{F}\left(\mathbb{R}^{\ell} \times \mathbb{R}^{\ell}\right)$ of all closed subsets of $\mathbb{R}^{\ell} \times \mathbb{R}^{\ell}$ endowed with the topology $\mathcal{T}_{\mathbf{c}losed}$ of closed convergence is compact and metrizable. In order to show that $(\mathcal{P}, \mathcal{T}_{\mathbf{c}losed})$ is compact and metrizable, it suffices to show that "*$\mathcal{P}$ is a closed subset of* $\left(\mathbb{R}^{\ell} \times \mathbb{R}^{\ell}, \mathcal{T}_{\mathbf{c}losed}\right)$". In this direction, let us assume that $(\mathfrak{X}_n, \succ_n)_{n\in\mathbb{N}}$ is a sequence in $\mathcal{P}$ and F is the closed limit of a sequence $(F_n)_{n\in\mathbb{N}}$ where $F_n := \left\{(x, y) \in \mathfrak{X}_n \times \mathfrak{X}_n : x \not\succ_n y\right\}$. We have to show that "*the data selection preference* $(\mathfrak{X}, \succ)$ *belongs to* F",where $\mathfrak{X} := \left\{x \in \mathbb{R}^{\ell} : (x, x) \in F\right\}$ and $\succ := (\mathfrak{X} \times \mathfrak{X}) \setminus F$. In other words, we have to show that

1. $\mathfrak{X}$ is a data focal set over a system S (i.e., a nonempty subset of the data space $\mathbb{R}^{\ell}$ which is closed, convex and bounded from below) and
2. $\succ \subset \mathfrak{X} \times \mathfrak{X}$ is a transitive and non-reflexive binary relation on $\mathfrak{X}$ such that $\succ$ is open in $\mathfrak{X} \times \mathfrak{X}$.

To do so, observe that, since $liminf_{n\to\infty} F_n = limsup_{n\to\infty} F_n = F$,

1. the set $\mathfrak{X}$ is the closed limit of the sequence $(\mathfrak{X}_n)_{n\in\mathbb{N}}$.

Further, the set $\mathfrak{X}$ is nonempty, since every set $\mathfrak{X}_n$ belongs to $\mathfrak{K}$ (Notation 2.1). It follows that

1. the set $\mathfrak{X}$ intersects a given compact set.

On the other hand, since every data focal set $\mathfrak{X}_n$ is convex,

1. the closed limit $\mathfrak{X}$ is a convex set.

Indeed, let $x, y \in \mathfrak{X}$ and $0 < \lambda < 1$. Since $\mathfrak{X} = liminf_{n\to\infty} \mathfrak{X}_n$, there are sequences $(x_n \in \mathfrak{X}_n)_{n\in\mathbb{N}}$ and $(y_n \in \mathfrak{X}_n)_{n\in\mathbb{N}}$ converging to x and y respectively. Since $\mathfrak{X}_n$ is convex, we have $\lambda x_n + (1-\lambda)\, y_n \in \mathfrak{X}_n$. Consequently, $\lambda x + (1-\lambda)\, y \in liminf_{n\to\infty} \mathfrak{X}_n = \mathfrak{X}$.

It is now easily seen that

1. $\mathfrak{X} \in \mathfrak{K}$.

We show now that the data selection preference $\succ$ on $\mathfrak{X}$ is non reflexive. Let $x \in \mathfrak{X}$. Then there is a sequence $(x_n \in \mathfrak{X}_n)_{n\in\mathbb{N}}$ converging to x. Since $\succ_n$ is non reflexive, we have $(x_n, x_n) \in F_n$. Hence $(x, x) \in F$, since $liminf_{n\to\infty} \mathfrak{X}_n = \mathfrak{X}$. Thus, we have $x \not\succ x$.

Next, we show that the data selection preference $\succ$ on $\mathfrak{X}$ is transitive. Let $x \succ y$ and $y \succ z$. To get a contradiction, let us assume that $x \not\succ z$, i.e., $(x, z) \in F$. Since $liminf_{n\to\infty} F_n = F$ there is a sequence $(x_n, z_n) \in F_n$ with $(x_n, z_n) \underset{n\to\infty}{\longrightarrow} (x, z)$. For n large enough, we have $\left(x_n, y_n\right) \notin F_n$ and $(y_n, z_n) \notin F_n$, where $(y_n \in \mathfrak{X}_n)_{n\in\mathbb{N}}$ converging to y. Indeed, if this were not true, it would follow that $(x, y) \in limsup_{n\to\infty} F_n = F$ or $(y, z) \in F$, which contradicts $x \succ y$ and $y \succ z$. Hence, by transitivity of $\succ_n$ we obtain $(x_n, z_n) \notin F_n$ which constitutes a contradiction.

(ii) It is well known that, in a compact metrizable space M endowed with the topology of closed convergence, a sequence $(F_n \subset M)_{n\in\mathbb{N}}$ of closed subsets of M converges to a closed set $F \subset M$ with respect to the topology of closed convergence in M if and only if $liminf_{n\to\infty} F_n = limsup_{n\to\infty} F_n = F$. Application for $M = (\mathcal{P}, \mathcal{T}_{\mathbf{c}losed})$ proves the desired assertion.

(iii) Since $(\mathcal{P}, \mathcal{T}_{\mathbf{c}losed})$ is a compact space, every separated coarser topology on $\mathcal{P}$ coincides with $\mathcal{T}_{\mathbf{c}losed}$. Thus, it remains to show that the set

$$\{(\mathfrak{X}, \succ, x, y) : x, y \in \mathfrak{X}\ and\ x \not\succ y\}$$

is closed in $(\mathcal{P}, \mathcal{T}_{\mathbf{c}losed}) \times \mathbb{R}^\ell \times \mathbb{R}^\ell$. Let $(\mathfrak{X}_n, \succ_n, x_n, y_n) \underset{n\to\infty}{\longrightarrow} (\mathfrak{X}, \succ, x, y)$, where $x_n, y_n \in \mathfrak{X}_n$ and $x_n \not\succ y_n$. Hence $(x_n, y_n) \in F_n$, which implies that $(x, y) \in liminf_{n\to\infty} F_n = F$, i.e. $x, y \in \mathfrak{X}$ and $x \not\succ y$.

For later easy reference we state three immediate consequences of Theorem 2.1.

Corollary 2.1 *The mapping* $(\mathfrak{X}, \succ) \mapsto \mathfrak{X}$ *of* $\mathcal{P}$ *into* $\mathbb{R}^\ell$ *is closed and lower hemicontinuous.*

Corollary 2.2 *The set* $\left\{(\mathfrak{X}, \succ, x, y) \in \mathcal{P} \times \mathbb{R}^\ell \times \mathbb{R}^\ell : x, y \in \mathfrak{X}\ and\ x \succ y\right\}$ *is a Borel subset of* $\mathcal{P} \times \mathbb{R}^\ell \times \mathbb{R}^\ell$.

Corollary 2.3 *Let* $(\mathfrak{X}, \succ) \in \mathcal{P}$, $x, y \in \mathfrak{X}$ *and* $x \succ y$. *Then there are neighborhoods* V, V_x *and* V_y *of* $(\mathfrak{X}, \succ)$ *in* $\mathcal{P}$, x *and* y *in* $\mathbb{R}^\ell$, *respectively, such that* $x' \succ' y'$, *for every* $\left(\mathfrak{X}', \succ\right) \in V$, $x' \in V_x \cap \mathfrak{X}'$ *and* $y' \in V_y \cap \mathfrak{X}'$.

In later sections, it will—for technical measure theoretical reasons—be important to know that the sets

$\mathcal{P}_{mo}$ *(the set of all monotonic data selection preferences in* $\mathcal{P}$*)*,
$\mathcal{P}^*$ *(the set of all negatively transitive data selection preferences in* $\mathcal{P}$*)*,
$\mathcal{P}^*_{co}$ *(the set of all convex (strongly convex) data selection preferences in* $\mathcal{P}^*$*)* and
$\mathcal{P}^*_{sco}$ *(the set of all convex (strongly convex) data selection preferences in* $\mathcal{P}^*$*)*

are Borel subsets of the compact metrizable space $\mathcal{P}$. In that regard, it is easy to show the following result.

Proposition 2.2 *The sets* $\mathcal{P}_{mo}$, $\mathcal{P}^*$, $\mathcal{P}^*_{co}$ *and* $\mathcal{P}^*_{sco}$ *are not closed* G_δ*-sets in* $\mathcal{P}$, *with closures different from* $\mathcal{P}$.

2.5 The Lower Hemicontinuity for the Rational Choice of Data Amount

Let $p = (p_1, p_2, \ldots, p_\ell)$ *be the data significance vector of the weighted data entity system* S.

Proposition 2.3 *The defining relation* $\mathfrak{B}$ *of the set*

$$\mathfrak{B}\,(\mathfrak{X}, \mathcal{C}_{\mathcal{M}}, p) :=\{x=(d_1, \ldots, d_\ell) \in \mathfrak{X} : (p_1, p_2, \ldots, p_\ell) \cdot (d_1, d_2, \ldots, d_\ell) \leq \mathcal{C}_{\mathcal{M}}\}$$

of the processing options of a data processor $\mathcal{M}$ *in the complex* S *is closed in* $\mathcal{P} \times \mathbb{R} \times \mathbb{R}^\ell$ *and lower hemicontinuous at every point* $(\mathfrak{X}, \mathcal{C}_{\mathcal{M}}, p) \in \mathcal{P} \times \mathbb{R} \times \mathbb{R}^\ell$.

Proof The defining relation $\mathfrak{B}$ is the intersection of the correspondence $(\mathfrak{X}, \succ) \mapsto \mathfrak{X}$ of $\mathcal{P}$ into $\mathbb{R}^\ell$ with the correspondence

$$(\mathfrak{X}, \mathcal{C}_{\mathcal{M}}, p) \longmapsto \left\{x \in \mathbb{R}^\ell : p \cdot x \leq \mathcal{C}_{\mathcal{M}}\right\}$$

of $\mathcal{P} \times \mathbb{R} \times \mathbb{R}^\ell$ into $\mathbb{R}^\ell$. Since, by Corollary 2.1, both mappings are closed, we infer that $\mathfrak{B}$ is closed. To show the lower hemicontinuity of $\mathfrak{B}$, let us consider the relation $\check{\mathfrak{B}}$ defined by

$$\check{\mathfrak{B}}\,(\mathfrak{X}, \mathcal{C}_{\mathcal{M}}, p) := \{x \in \mathfrak{X} : p \cdot x \leq \mathcal{C}_{\mathcal{M}}\}.$$

By assumption, there is a vector $x = (d_1, \ldots, d_\ell) \in \check{\mathfrak{B}}\,(\mathfrak{X}, \mathcal{C}_{\mathcal{M}}, p)$.
Let $\left(\mathfrak{X}_n, \succ_n, \mathcal{C}_{\mathcal{M}}^{(n)}, p_{\mathbf{n}}\right)_{n\in\mathbb{N}}$ be a sequence converging to $(\mathfrak{X}, \succ, \mathcal{C}_{\mathcal{M}}, w)$ in $\mathcal{P}$. By Corollary 2.1, the correspondence $(\mathfrak{X}, \succ) \mapsto \mathfrak{X}$ of $\mathcal{P}$ into $\mathbb{R}^\ell$ is low-hemicontinuous. Thus, there is a sequence $(x_n \in \mathfrak{X})_{n\in\mathbb{N}}$ converging to $x \in \mathfrak{X}$. Evidently, the strict inequality $w \cdot x < \mathcal{C}_{\mathcal{M}}$ implies $w_{\mathbf{n}} \cdot x_n < \mathcal{C}_{\mathcal{M}}^{(n)}$ for n large enough. Hence, $x_n \in \check{\mathfrak{B}}\left(\mathfrak{X}_n, \mathcal{C}_{\mathcal{M}}^{(n)}, p_{\mathbf{n}}\right)$ for enough large n, which proves that the relation $\check{\mathfrak{B}}$ is lower hemicontinuous $\underline{\text{at } (\mathfrak{X}, \mathcal{C}_{\mathcal{M}}, p)}$. The convexity of the data focal set $\mathfrak{X}$ implies that $\mathfrak{B}\,(\mathfrak{X}, \mathcal{C}_{\mathcal{M}}, p) = \check{\mathfrak{B}}\,(\mathfrak{X}, \mathcal{C}_{\mathcal{M}}, p)$. The desired assertion now follows, since the closure of a lower hemicontinuous mapping is also lower hemicontinuous.

Proposition 2.4 *The defining relation*

$$\mathfrak{A} of\ \ the\ \ set\ \ \mathfrak{A}\,(\mathfrak{X}, \succ, \mathcal{C}_{\mathcal{M}}, p) = \Big\{x^* = (d_1, \ldots, d_\ell) \in \mathfrak{B}\,(\mathfrak{X}, \mathcal{C}_{\mathcal{M}}, p) :$$

$$there\ is\ no\ x = (d_1, \ldots, d_\ell) \in \mathfrak{B}\,(\mathfrak{X}, \mathcal{C}_{\mathcal{M}}, p)\ \ with\ x \succ x^*\Big\}$$

of data amount's rational choices in the complex S *is nonempty and compact in* $\mathcal{P} \times \mathbb{R} \times \mathbb{R}^\ell$. *Further, it is lower hemicontinuous at every point* $(\mathfrak{X}, \succ, \mathcal{C}_{\mathcal{M}}, p) \in \mathcal{P} \times \mathbb{R} \times \mathbb{R}^\ell$ *where the set* $\mathfrak{B}\,(\mathfrak{X}, \mathcal{C}_{\mathcal{M}}, w)$ *of processing options of* $\mathcal{M}$ *is compact and* $\mathfrak{X}, \mathcal{C}_{\mathcal{M}}, p$ *satisfy the inequality*

$$\inf\{p \cdot \mathfrak{X}\} < \mathcal{C}_{\mathcal{M}}.$$

Note that the assumption $\inf\{p \cdot \mathfrak{X}\} < \mathcal{C}_{\mathcal{M}}$ *cannot be weakened to*

$$\inf\{p \cdot \mathfrak{X}\} \leq \mathcal{C}_{\mathcal{M}}.$$

Proof By Proposition 2.3, the defining relation $\mathfrak{B}$ of the set $\mathfrak{B}(\mathfrak{X}, \mathcal{C}_{\mathcal{M}}, p)$ is closed and lower hemicontinuous at every point $(\mathfrak{X}, \mathcal{C}_{\mathcal{M}}, w) \in \mathcal{P} \times \mathbb{R} \times \mathbb{R}^{\ell}$. Since the set $\mathfrak{B}(\mathfrak{X}, \mathcal{C}_{\mathcal{M}}, p)$ of processing options of $\mathcal{M}$ is compact and convex, the defining relation $\mathfrak{B}$ of $\mathfrak{B}(\mathfrak{X}, \mathcal{C}_{\mathcal{M}}, p)$ is continuous at the point $(\mathfrak{X}, \mathcal{C}_{\mathcal{M}}, p)$. Put $\mathfrak{S} := \mathcal{P} \times \mathbb{R} \times \mathbb{R}^{\ell}$. Since, by Theorem 2.1(iii) and Proposition 2.3, the set $\left\{(\mathfrak{s}, x, y) \in \mathfrak{S} \times \mathbb{R}^{\ell} \times \mathbb{R}^{\ell} : x, y \in \mathfrak{B}(\mathfrak{s}) \textit{ and } x \not\succ_{\mathfrak{s}} y \right\}$ is closed in $\mathfrak{S} \times \mathbb{R}^{\ell} \times \mathbb{R}^{\ell}$, the desired assertion follows.

2.6 Mean Rational Data Amount Choice

2.6.1 Data Sectors

We consider a finite set

$$\mathbb{M}$$

of data processors $\mathcal{M}$, each of whom is described by its data focal set $\mathfrak{X}_{\mathcal{M}}$ in the complex S, his data selection preference $\succ_{\mathcal{M}}$ *in* S and his data amount processing capacity $\mathcal{C}_{\mathcal{M}}$ *in* S. We introduce the map

$$\mathfrak{s} : \mathbb{M} \to \mathcal{P} \times \mathbb{R} : \mathcal{M} \longmapsto \mathfrak{s}(\mathcal{M}) = (\mathfrak{X}_{\mathcal{M}}, \succ_{\mathcal{M}}, \mathcal{C}_{\mathcal{M}}).$$

Notation 2.2 If $\mathcal{M}$ selects a weighted data system with data significance $p = (p_1, p_2, \ldots, p_{\ell})$ in S, then *the data amount's rational choice set of a data processor* $\mathcal{M}$ *with characteristics* $\mathfrak{s}(\mathcal{M}) \in \mathcal{P} \times \mathbb{R}$ will be denoted by
$\mathfrak{A}(\mathfrak{s}(\mathcal{M}), p)$.
Thus, we are leaded to the following.

Definition 2.14 If each data processor $\mathcal{M}$ selects the weighted data system with data significance $p = (p_1, p_2, \ldots, p_{\ell})$ in the complex S, the ***mean rational data amount choice*** of the set $\mathbb{M}$ in S is given by

$$\overline{\mathfrak{A}}(\mathfrak{s}, p) := \frac{1}{|\mathbb{M}|} \sum_{\mathcal{M} \in \mathbb{M}} \mathfrak{A}(\mathfrak{s}(\mathcal{M}), p)$$

Here the notation $|\cdot|$ means cardinality of set.

If χ denotes the ***normalized counting measure*** on $\mathbb{M}$, i.e.,

$$\chi(\varepsilon) := |\varepsilon| / |\mathbb{M}|$$

for every subset ε of $\mathbb{M}$, it is immediately verified that

$$\overline{\mathfrak{A}}(\mathfrak{s}, p) := \int_{\mathbb{M}} \mathfrak{A}(\mathfrak{s}(\cdot), p)\, d\chi.$$

Clearly, the integral is defined for more general mappings $\mathfrak{s}$ and measures χ. Indeed, we shall define later "mean rational data amount choice" by this formula in a more general situation. However, let us first prepare and motivate this step of abstraction.

Definition 2.15 The image measure ϱ of χ with respect to the mapping $\mathfrak{s}$ is called the ***data preference-capacity distribution*** of the set $\mathbb{M}$ of data processors in the complex S.

Thus,

$$\varrho\left(B\right)=\chi\left(\mathfrak{s}^{-1}\left(B\right)\right)$$

denotes the fraction of data processors in $\mathbb{M}$ whose characteristics belong to $B\subset\mathcal{P}\times\mathbb{R}$.

Definition 2.16 The marginal distributions
$\varrho^{\mathcal{P}}$ on $\mathcal{P}$ and $\varrho^{\mathbb{R}}$ on $\mathbb{R}$
are called the ***data preference distribution*** and ***data capacity distribution*** in the complex S, respectively.

Remark 2.4 The data preference-capacity distribution in the complex S may or may not to be the product of its marginal distributions, i.e., it *is not assumed that the data capacity distribution is independent of the data preference distribution.*

Notation 2.3 In some general cases, one is not primarily concerned about the total rational data choice of a small number of data processors. Typically, one is interested in the ***total rational data choice*** in S of all data processors in a large society. In this case, it seems natural and convenient (for analytical reasons) to view the data preference-capacity distribution ϱ as an ***atomless distribution***[1] [5] over the space of all characteristics $\mathcal{P}\times\mathbb{R}$, that is as a distribution satisfying $\varrho\left(\mathcal{X},\succ,\mathcal{C}_{\mathcal{M}}\right)=0$ for every $\left(\mathcal{X},\succ,\mathcal{C}_{\mathcal{M}}\right)\in\mathcal{P}\times\mathbb{R}$.

To view the distribution of the processor's characteristics of a finite set $\mathbb{M}$ of data processors as an atomless distribution means that t*he "actual" distribution is considered as a distribution of a sample of size* $|\mathbb{M}|$ *drawn from a "hypothetical" population of processors*. There is also another reason why one should consider atomless distributions of data processor's characteristics: *the very fact that data processors are not alike*—which means in our framework that *the support of the data preference-capacity distribution is "spread over" the set* $\mathcal{P}\times\mathbb{R}$ *can give rise to properties*, for example of the mean rational data amount choice in S, which would not hold without the diversification of data processor's characteristics. To be more specific, *if data selection preferences, say in* $\mathcal{P}^{*}$, *are not strongly convex in S, the mean rational data amount choice in S for a set of data processors is, in general, not unique*. However, given a weighted data system with data significance

[1] A distribution μ on $\mathbb{P}\times\mathbb{K}$ is atomless if $\mu\left(\mathcal{X},\succ,\mathcal{C}_{\mathcal{M}}\right)=0$ for every $\left(\mathcal{X},\succ,\mathcal{C}_{\mathcal{M}}\right)\in\mathcal{P}\times\mathbb{R}$.

vector $p \gg 0$, Proposition 2.4 guarantees that, for a topologically large subset of characteristics $\left(\precsim, \mathcal{C}_{\mathcal{M}}\right)$, the rational choice set of data amounts in S has a small diameter. Thus, ***for a "widely spread" distribution of data characteristics one can hope that, for "most" data processors, the rational choice set of data amounts in the complex** S **is small***.

The assumption of atomless distributions of data processors' characteristics, in particular, requires that "many" data processors be involved. Then the focal data decision in the complex S of a typical data processor will have only a small influence on the total rational data choice in S. In such a case, we do not need that the distribution of data processors' characteristics is atomless, but this distribution is induced from a very "large" set of data processors The above discussion motivates the following.

Definition 2.17 Let $\mathbb{M}$ be the set of all data processors in the complex S.

(i) A ***data sector*** in the complex S is a measurable mapping

$$\mathfrak{s} : (\mathbb{M}, \mathcal{A}, \upsilon\ ) \rightarrow \mathcal{P}\times\mathbb{R}$$

of a measure space $(\mathbb{M}, \mathcal{A}, \upsilon\ )$, consisting of the set $\mathbb{M}$, a σ-algebra $\mathcal{A}$ of subsets of $\mathbb{M}$ and a (probability) measure υ on $\mathcal{A}$, into the space $\mathcal{P}\times\mathbb{R}$ of data characteristics such that the mean data processing capacity

$$\int_{\mathbb{M}} \mathcal{C}_{\mathcal{M}} \circ \mathfrak{s}\ d\upsilon$$

is finite.

(ii) A data sector in S is called

1. ***simple***, if the measure space $(\mathbb{M}, \mathcal{A}, \upsilon\ )$ is simple, i.e., $\mathbb{M}$ is a finite set, $\mathcal{A}$ is the set of all subsets of $\mathbb{M}$, and $\upsilon\ (\mathcal{E}) = (|\mathcal{E}|/|\mathbb{M}|)$ whenever $\mathcal{E} \subset \mathbb{M}$;
2. ***partitionable***, if the measure space $(\mathbb{M}, \mathcal{A}, \upsilon\ )$ is atomless, i.e., for every $\mathcal{E} \in \mathcal{A}$ with $\upsilon\ (\mathcal{E}) > 0$ there is a set $\mathcal{K} \subset \mathcal{E}$ with $0 < \upsilon\ (\mathcal{K}) < \upsilon\ (\mathcal{E})$;
3. ***convex***, if almost all data processors of every atom of the measure space $(\mathbb{M}, \mathcal{A}, \upsilon\ )$ have convex data selection preferences.

Remark 2.5 According to this definition, *a partitionable data sector in* S *is always convex in* S.

Notation 2.4

(i) *The generic element in the set* $\mathbb{M}$ *of a data sector in* S *is a data processor* $\mathcal{M}$ *in* S.

(ii) *The data selection preference and data amount processing capacity of a data processor in* S *are denoted by*

$$\mathfrak{s}\ (\mathcal{M}) = \left(\mathfrak{X}_{\mathfrak{s}(\mathcal{M})}, \succ_{\mathfrak{s}(\mathcal{M})}, \mathcal{C}_{\mathfrak{s}(\mathcal{M})}\ \right).$$

(iii) *If it is clear which mapping* $\mathfrak{s}$ *is considered, we shall write, as usually, shorter*

$$(\mathfrak{X}_{\mathcal{M}}, \succ_{\mathcal{M}}, \mathcal{C}_{\mathcal{M}}) .$$

(iv) *The image measure*

$$v \circ \mathfrak{s}^{-1}$$

is called the ***data preference-capacity distribution of the data sector*** $\mathfrak{s} : (\mathbb{M}, \mathcal{A}, v) \to \mathcal{P} \times \mathbb{R}$ *in S and is denoted by* $\varrho_{\mathfrak{s}}$ *, or simply* ϱ*.*

(v) *Given a weighted data vector* $w \in \mathbb{R}^{\ell}$*, the integral*

$$\int_{\mathbb{M}} \mathfrak{A} (\mathfrak{s} (\cdot) , p) \, dv$$

is called the ***mean rational data amount choice of the data sector*** $\mathfrak{s} : (\mathbb{M}, \mathcal{A}, v) \to \mathcal{P} \times \mathbb{R}$ *in S. It is denoted by*

$$\overline{\mathfrak{A}} (\mathfrak{s}, p) .$$

A partitionable data sector in S is, in fact, a more abstract concept. Its interpretation relies on the analogy to the case of a simple data sector. It describes a data sector in S with a very large set of data processors—an uncountable infinite set—where every individual data processor has strictly no influence on the mean rational data amount choice.

The σ-algebra $\mathcal{A}$ has only been introduced for technical reasons. Conceptually $\mathcal{A}$ should be considered—as in the case of a simple data sector—as the set of all subsets of $\mathbb{M}$.

2.6.2 Data Preference-Capacity Distributions

One easily verifies (we shall prove a more general result in Theorem 2.2 below) that *the mean rational data amount choice* $\overline{\mathfrak{A}} (s, p)$ *in the complex S only depends on the data preference-capacity distribution* $\varrho = \chi \circ s^{-1}$ *in S, provided the data amount's rational choice sets* $\mathfrak{A} (s (\mathcal{M}) , p)$ *are convex in S*. More precisely, we have

$$\overline{\mathfrak{A}} (s, p) = \int_{\mathcal{P} \times \mathbb{R}} \mathfrak{A} (\cdot, p) \; d\varrho.$$

However, in general, the mean rational data amount choice in S depends on the data preference-capacity distribution in S and on the number $|\mathbb{M}|$ of data processors in $\mathbb{M}$. We shall now show in which situation the mean rational data amount choice in S is determined by the data preference-capacity distribution in S.

To prove the first result of this paragraph, we may quote some auxiliary material with necessary background.

Lemma 2.1 ([3]) *Let $(\Omega,\ A,\ m)$ be a measure space consisting of a set Ω, a σ-algebra A of subsets of Ω and a (probability) measure m on A.*

1. *Let φ be a mapping with a measurable graph of a measurable space T into $\mathbb{R}^n$.*
2. *If h is a measurable function of T into $\mathbb{R}^n$, then the mapping $\omega \longmapsto \varphi(\omega) + h(\omega)$ has a measurable graph.*
3. *If h is a measurable function of $(\Omega,\ A,\ v)$ into T, then the composition $\varphi \circ h : \omega \longmapsto \varphi(h(\omega))$ has a measurable graph.*
4. *Let φ be a mapping with a measurable graph of $(\Omega,\ A,\ m)$ into a complete separable metric space Υ and h a measurable mapping of Υ into a separable metric space M.*
5. *The mapping $h \circ \varphi : \omega \longmapsto h(\varphi(\omega))$ has an $(A \times \mathcal{B}(M))$-analytic graph.*[2]
6. *If φ is a mapping with a measurable graph of $(\Omega,\ A,\ v)$ into $\mathbb{R}^n$, then the mapping*

$$conv(\varphi) : \omega \longmapsto\ conv(\varphi)(\omega)$$

has an $(A \times \mathcal{B}^n)$-analytic graph.

1. *Suppose Ω is a topological complete space and Υ is another complete separable metric space. If φ is a mapping with a measurable graph of Ω into Υ and u is a measurable function of Υ into $\mathbb{R}$, then*
2. *the function*

$$\sup u(\varphi(\cdot))\ : \Omega \to\ \mathbb{R}\colon \omega \mapsto \sup u(\varphi(\omega)) := \sup\{u(x) : x \in \varphi(\omega)\}$$

is measurable and

1. *the mapping*

$$\varphi^u\ :\ \Omega \to\ \Upsilon\ :\ \omega \mapsto \{x \in \varphi(\omega) : u(x) = \sup u(\varphi(\omega))\ \}$$

has a measurable (analytic) graph.

1. *If, in particular, $\Upsilon = \mathbb{R}^n$, the graph of the mapping $conv(\varphi)$ is measurable.*
2. *If $(\Omega,\ A,\ m)$ is an atomless measure space [2] and φ is a mapping with a measurable graph of $(\Omega,\ A,\ m)$ into $\mathbb{R}^n$, then the following properties hold.*
3. *The integral*

$$\int_{\Omega} \varphi\, dm$$

is a convex set in $\mathbb{R}^n$.

1. *Let Υ be a set in $\mathbb{R}^n$. If $\varphi(\omega) := S$ for every $\omega \in\ \Omega$, then*

$$\int_{\Omega} \varphi\, dm = conv(\Upsilon)\,.$$

[2] $\mathcal{B}(M)$ denotes the Borel σ-algebra generated by the open subsets of M.

2. *If the mapping φ of $(\Omega,\ A,\ m)$ into $\mathbb{R}^n$ is closed-valued and integrably bounded, then the integral*

$$\int_{\Omega} \varphi \, dm$$

is a compact subset of $\mathbb{R}^n$.

1. *Let φ be a mapping with a measurable graph of the measurable space $(\Omega,\ A,\ m)$ into $\mathbb{R}^n$. If $\int \varphi \neq \varphi(\omega)$, $\omega \in A$, then*

$$sup\left\{p \cdot\ x : x \in \int \varphi\right\} = \int sup\left\{p\ \cdot x : x \in \varphi(\cdot)\right\}$$

for every vector $p \in \mathbb{R}^n$.

1. *Let φ be a mapping with a measurable graph of the measurable space $(\Omega,\ A,\ m)$ into $\mathbb{R}^n_+$. The following hold.*

$$conv\left(\int_{\Omega} \varphi \, dm\right) = \int_{\Omega} conv(\varphi)\ dm.$$

In particular, if the measure space is atomless, then $\int_{\Omega} \varphi \, dm = \int_{\Omega} conv(\varphi)\ dm$.

1. *Let φ be a mapping with a measurable graph of a measurable space $(T,\ \mathfrak{J})$ into $\mathbb{R}^n$ such that $\varphi(t)$is closed convex and contains no straight line whenever $t \in T$. If h is a measurable function of $(\Omega,\ A,\ v)$ into T, then*

$$\int_{\Omega} \varphi \circ h \, dm = \int_{T} \varphi \, d\left(m \circ h^{-1}\right).$$

Proof

(i)(a) The mapping $f : (\omega, x) \longmapsto (\omega, x - h(\omega))$ of $\Omega \times \mathbb{R}^n$ into $\Omega \times \mathbb{R}^n$ is $(A \otimes \mathcal{B}^n)$-measurable. Here $\mathcal{B}$ denotes the Borel σ-algebra on $\mathbb{R}$ generated by the open subsets of $\mathbb{R}$. Consequently, if $G_{\varphi+h}$ and G_{φ} are the graphs of $\varphi + h$ and φ, respectively then $G_{\varphi+h} = f^{-1}\left(G_{\varphi}\right) \in A \otimes \mathcal{B}^n$.

(b) Similarly, since the mapping $g : (\omega, x) \longmapsto (h(\omega), x)$ of $\Omega \times \mathbb{R}^n$ into $T \times \mathbb{R}^n$ is measurable, the graph $G_{\varphi \circ h} = g^{-1}\left(G_{\varphi}\right)$ is measurable.

(ii)(a) The set $G = \{(\omega, x, z) \in \Omega \times \Upsilon \times M{:}x \in \varphi(\omega)\ and\ z = h(x)\}$ belongs to $A \otimes \mathcal{B}(\Upsilon) \otimes \mathcal{B}(M)$. Since the graph $G_{h\circ\varphi}$ is obtained by projecting the set G on $\Omega \times M$, and since Υ is complete separable metric space, we infer that $G_{h\circ\varphi}$ is an $(A \times \mathcal{B}(M))$-analytic graph (see p. 34 in [4]).

(b) Let $\Delta = \left\{(\xi_1, \ldots, \xi_{n+1}) : \xi_i \geq 0\ and\ \sum_{i=1}^{n+1} \xi_i = 1\right\}$. The mapping $\psi : \omega \longmapsto \varphi(\omega) \times \cdots \times \varphi(\omega) \times \{(\xi_1, \ldots, \xi_{n+1})\}$ of $(\Omega,\ A,\ m)$ into $\mathbb{R}^{n(n+1)} \times \Delta$ has a measurable graph. The function

$h : (x_1, \ldots, x_{n+1}, \xi_1, \ldots, \xi_{n+1}) \longmapsto \sum_{i=1}^{n+1} \xi_i x_i$ of $\mathbb{K}^{n(n+1)} \times \Delta$ into $\mathbb{R}^n$ is continuous. Since $conv\,(\varphi) = h \circ \varphi$, part (ii)(a) implies that the graph $G_{conv(\varphi)}$ is analytic.

(iii)(a) We have to show that for every $c \in \mathbb{R}$ the set

$$\Omega_c^{(\varphi)} := \{\omega \in \Omega \colon \sup u\,(\varphi\,(\omega)) > c\,\}$$

belongs to A. Since $\Omega_c^{(\varphi)} = proj_{\Omega} \left\{(\omega, x) \in G_{\varphi} : u\,(x) > c\right\}$ (G_{φ} is the graph of φ) and since the assumptions on φ and u imply that

$$\left\{(\omega, x) \in G_{\varphi} : u\,(x) > c\right\} \in A \otimes\ \mathcal{B}\,(\Upsilon)$$

it follows from the Projection Theorem that $\Omega_c^{(\varphi)} \in A$.

(b) The second assertion now follows readily. The function $(\omega, x) \mapsto u\,(x) - \sup u\,(\varphi\,(\omega))$ is $A \int \mathcal{B}\,(\Upsilon)$ −measurable. Hence

$$V = \{(\omega, x) \in \Omega \times \Upsilon : u\,(x) = \sup u\,(\varphi\,(\omega))\,\} \in A \otimes\ \mathcal{B}\,(\Upsilon)$$

and consequently $G_{\varphi^u} = G_{\varphi} \bigcap V \in A \otimes\ \mathcal{B}\,(\Upsilon)$.

(c) Let $\varphi^k\,(\omega) = \{x \in \varphi\,(\omega) : |x| \leq k\}$ $(k = 1, 2, \ldots)$. One easily verifies that $conv\,(\varphi\,(\omega)) = \bigcup_{k=1}^{\infty} conv\left(\varphi^k\,(\omega)\right)$. For every $v \in \mathbb{R}^n$, consider the mapping H_v of Ω into $\mathbb{R}^n$:

$$H_v\,(\omega) := \left\{x \in \mathbb{R}^n : v \cdot x \leq \sup v \cdot \varphi^k\,(\omega)\,\right\}.$$

By part (iii)(b), the function $\omega \mapsto \sup v\varphi^k\,(\omega)$ is measurable, and hence the graph of H_v is measurable. Since $conv\left(\varphi^k\,(\omega)\right) = \bigcap_{v \in D} H_v\,(\omega)$, where D denotes a countable dense subset in $\mathbb{R}^n$, the graph of $conv\left(\varphi^k\right)$ $(k = 1, 2, \ldots)$ is measurable, and hence the graph of $conv\,(\varphi)$ is measurable.

(iv)(a) Let $x_1,\ x_2 \in \int \varphi dm$ and $0 < \lambda < 1$. We denote by $\mathfrak{L}_{\varphi}$ the set of m-integrable functions $f : \Omega \to \mathbb{R}^n$ such that $f\,(\omega) \in \varphi\,(\omega)$ almost everywhere in Ω. There are integrable functions $f_1,\ f_2 \in \mathfrak{L}_{\varphi}$ such that $x_1 = \int f_1 dm$ and $x_2 = \int f_2 dm$. From Liapunov's Theorem, it follows that the set

$$\left\{\left(\int_E f_1\ dm, \int_E f_2\ dm\ \right) \in \mathbb{R}^{2n} :\ \ E \in A\right\}$$

is convex. Since $(0, 0)$ and (x_1, x_2) belong to this set, there exists a set $E \in A$ such that

$$(\lambda x_1, \lambda\ x_2) = \left(\int_E f_1\ dm, \int_E f_2\ dm\ \right).$$

Define the function $f \in \mathfrak{L}_{\varphi}$ by

$$f(\omega) = \begin{cases} f_1(\omega), & if \ \omega \in E \\ f_2(\omega), & if \ \omega \notin E. \end{cases}$$

Then, one easily verifies that $\int f = \lambda x_1 + (1-\lambda)\ x_2$.This shows that *the integral $\int_{\Omega} \varphi\ dm$ is a convex set in $\mathbb{R}^n$.*

(b) From part (iv)a, it follows that $conv\,(\varUpsilon) \subset \int_{\Omega} \varphi\ dm$. On the other hand, it is easily proved, by induction on the dimension of $\mathbb{R}^n$, that $\int_{\Omega} f\ dm \in conv\,(\varUpsilon)$ for every f with $f(\omega) \in \varUpsilon$, almost everywhere on Ω. In particular, $\int_{\Omega} \varphi\ dm \subset conv\,(\varUpsilon)$.

(v) Observe that, by Fatou's lemma in n-dimension, *if $(\varphi_{\nu})_{\nu \in \mathbb{N}}$ is a sequence of mappings of $(\Omega,\ A,\ m)$ into $\mathbb{R}^n_+$ such that there exists a sequence $(g_{\nu})_{\nu \in \mathbb{N}}$ of functions of Ω into$\mathbb{R}^n_+$ with the properties:*

1. *$\varphi_{\nu}(\omega) = g_{\nu}(\omega)$, almost everywhere in Ω and*
2. *the sequence $(g_{\nu})_{\nu \in \mathbb{N}}$ is uniformly integrable and the set $\{g_{\nu}(\omega) : \nu \in \mathbb{N}\}$ is bounded almost everywhere in Ω,*

then $limsup_{\nu \in \mathbb{N}} \left(\int \varphi_{\nu}\right) \subset \int\ limsup_{\nu \in \mathbb{N}}(\varphi_{\nu})$. Letting $\varphi_{\nu} = \varphi\ (\nu \in \mathbb{N})$, we have $limsup_{\nu \in \mathbb{N}} \left(\int \varphi_{\nu}\right) = \varphi(\omega)$ since $\varphi(\omega)$ is closed. Thus, by the above remark, every adherent point of $\int \varphi$ belongs to $\int \varphi$. This proves assertion (v).

(vi) The left-hand side is clearly at most equal to the right-hand side. From part (iv), it follows that the function $\omega \mapsto sup\,\{p\,\varphi(\omega)\}$ is A-measurable. Since there is by assumption an integrable selection for φ, the right-hand side is well defined (it may be $+8$). Consider a real number $\alpha < \int s$, where $s(\omega) = sup\,\{p\,\varphi(\omega)\}$. We have to show that there is a function $f \in \mathfrak{L}_{\varphi}$ such that $\alpha < p \int f$. For this we choose an integrable selection $h \in \mathfrak{L}_{\varphi}$ and consider for every integer ν the truncated mapping $\varphi_{\nu}(\omega) = \{x \in \varphi(\omega) : |x - h(\omega)| = \nu\}$. Clearly, the graph of φ_{ν} is measurable. Hence, by part (c), the function $s_{\nu}(\omega) = sup\,\{p\,\varphi_{\nu}(\omega)\}$ is measurable. It is also integrable, since h is integrable. Since $(s_{\nu}(\omega))_{\nu \in \mathbb{N}} \nearrow s(\omega)$, we obtain, by the monotone convergence theorem, that $\int s_{\nu} \to \int s$. Consequently, for ν large enough, we have $\alpha < \int s_{\nu}$. Thus, there is an integrable function g of Ω into $\mathbb{R}$ such that $\alpha < \int g$ and $g(\omega) < s_{\nu}(\omega)$, $\omega \in \Omega$. Let $\psi(\omega) := \{x \in \varphi_{\nu}(\omega) : p\,x > g(\omega)\}$. Clearly $\psi(\omega) \neq \varphi(\omega)$, $\omega \in A$, and the graph of the mapping ψ is measurable. Consequently, by the Measurable Selection Theorem, there exists a measurable selection f of ψ, and hence of φ, which is even integrable. Since $g(\omega) < pf(\omega)$ we obtain $\int g < p \int f$ and consequently $< p \int f$.

(vii) We prove the assertion by induction on the dimension n of $\mathbb{K}^n$. Clearly the theorem holds for $n = 0$. First we show that

$$\int conv\,(\varphi) = \int \varphi\ \ if\ \ and\ \ only\ \ if\ \ conv\left(\int \varphi\right) = \int \varphi. \tag{1}$$

If $\int conv\left(\varphi\right) = \phi$, then $\int \varphi = \phi$ and $conv\left(\int \varphi\right) = \phi$. To show the converse, we assume that $\int conv\left(\varphi\right) \neq \phi$. Let f be an integrable selection of $conv\left(\varphi\right)$. Let $v \in \mathbb{R}^n$ and $v \gg 0$. Consider the set $\psi\left(\omega\right) = \{x \in \varphi\left(\omega\right) : vx = vf\left(\omega\right)\}$. Since $f\left(\omega\right) \in conv\left(\varphi\left(\omega\right)\right)$, we have $\psi\left(\omega\right) \neq \phi$ almost everywhere in Ω. The graph of the mapping ψ is measurable. Therefore, by the measurable selection theorem, there exists a measurable selection h of ψ. Since f is integrable, $v \gg 0$ and φ is positive, the selection h is integrable, and hence $\int \varphi \neq \phi$. In the remainder of the proof of (vii), we shall assume that $\int \varphi \neq \phi$. Next, we show that:

$$conv\left(\int \varphi\right) \ and \ \int conv\left(\varphi\right) \tag{2}$$

have the same closure.

For every $v \in \mathbb{R}^n$, one obtains $sup\left(v \int \varphi\right) \leq sup\left(v \int conv\left(\varphi\right)\right) \leq \int sup\left(v\varphi\right) = sup\left(v \int \varphi\right)$. Indeed, the two inequalities are trivial and the equality follows from (v). Hence, for every $v \in \mathbb{R}^n$, we have $sup\left(v \int conv\left(\varphi\right)\right) = sup\left(vconv\left(\int \varphi\right)\right)$, which proves property (2), since the two sets are convex. Now, for every subset X of $\mathbb{R}^n$ and every $v \in \mathbb{R}^n$ we define $X^v := \{x \in X : vx = \sup vX\}$. It remains to show that for every $v \in \mathbb{R}^n$ we have $\left(\int conv\left(\varphi\right)\right)^v = \left(conv \int \varphi\right)^v$. Using the measurable selection theorem, one easily shows that if ψ is a mapping of (Ω, A, m) into $\mathbb{R}^n$ whose graph is analytic and $\int \psi \neq \phi$, then for every $v \in \mathbb{R}^n$ we have $\left(\int \psi\right)^v \neq \phi$ if and only if $\psi^v\left(\omega\right) \neq \phi$ almost everywhere in Ω and $\int \psi^v \neq \left(\int \psi\right)^v$. We want to apply this to the mappings φ and $conv\left(\varphi\right)$. Since the graph of $conv\left(\varphi\right)$ is analytic, but may not be measurable, we have to use here the Measurable Selection Theorem for analytic sets. (if φ is closed-valued, then $conv\left(\varphi\right)$ has a measurable graph, by part (iii)(c)). Also one easily verifies that for every nonempty subset X of $\mathbb{R}^n$ one has $\left(convX\right)^v = conv\left(X^v\right)$. Consequently $\left(\int conv\left(\varphi\right)\right)^v = \int \left(conv\left(\varphi\right)\right)^v = \int conv\left(\varphi^v\right)$ and analogously $conv\left(\int \varphi\right)^v = conv\left(\int \varphi\right)^v = conv\left(\int \varphi^v\right)$. Thus, it remains to show that:

$$\int conv\left(\varphi^v\right) = conv\left(\int \varphi^v\right). \tag{3}$$

Since, by part (iii), the graph of the relation φ^v is measurable, it follows, from (1) that $\int conv\left(\varphi^v\right) = \phi$ if and only if $\int \varphi^v = \phi$. Thus we may assume in the remainder of the proof that $\int \varphi^v \neq \phi$. Consider the hyperplane $\mathcal{H} = \{x \in \mathbb{R}^n : vx = 0\}$. There exists a coordinate axis $\mathcal{L}$, say the first, not contained in $\mathcal{H}$. We consider the projection $\mathcal{Q}$ parallel to $\mathcal{L}$ into $\mathcal{H}$. Let h be the function of Ω into $\mathbb{R}^n$ defined by $\omega \longmapsto h\left(\omega\right) := x - \mathcal{Q}x$ for some $x \in \varphi^v\left(\omega\right)$. The function h is well-defined and measurable. Clearly $\varphi^v\left(\omega\right) = \mathcal{Q}\varphi^v\left(\omega\right) + h\left(\omega\right)$. One easily verifies that $conv\left(\int \varphi^{v}\right) = conv\left(\int(\mathcal{Q}\varphi^{v}+h)\right) = conv\left(\int \mathcal{Q}\varphi^{v}\right) + \int h$ and $\int conv\,(\varphi^{v}) = \int conv\,(\mathcal{Q}\varphi^{v}+h) = \int conv\,(\mathcal{Q}\varphi^{v}) + \int h$. Hence, in order to prove (3) it suffices to prove that

$$conv\left(\int \mathcal{Q}\varphi^{v}\right) = \int conv\left(\mathcal{Q}\varphi^{v}\right). \tag{4}$$

This follows from the induction hypothesis. Indeed, the vectors $\mathcal{Q}e_2,\ldots,\ \mathcal{Q}e_n$ form a basis for the hyperplane $\mathcal{H}$ (e_j denotes the jth unit vector in $\mathbb{R}^n$). With respect to this basis, the mapping $\mathcal{Q}\varphi^{v}$ becomes a mapping φ^{v} of Ω into $\mathbb{K}^{n-1}$. The mapping φ^{v} is positive, since φ is positive. Moreover, φ^{v} has a measurable graph, since $\mathcal{Q}\varphi^{v}$ has a measurable graph. Therefore, by induction hypothesis, we obtain $conv\left(\int \varphi^{v}\right) = \int conv\,(\varphi^{v})$. Let $\mathcal{T}$ denote the linear and injective mapping of $\mathbb{R}^{n-1}$ into $\mathbb{R}^n$, defined by $\mathcal{T}(\zeta_2,\ldots,\zeta_n) = \sum_j \zeta_j \mathcal{Q}e_j$. Clearly $\mathcal{Q}\varphi^{v}(\omega) = \mathcal{T}\varphi^{v}(\omega)$. One easily verifies that $\int conv\,(\mathcal{T}\circ\varphi^{v}) = \mathcal{T}\int conv\,(\varphi^{v}) = \mathcal{T}conv\left(\int\varphi^{v}\right) = conv\left(\int \mathcal{T}\circ\varphi^{v}\right)$.

Thus, we obtain (4).

(viii) It is well known that if f is a measurable selection for φ, then $f\circ h$ is a measurable selection for $\varphi\circ h$. Therefore, the "change-of-variable formula"[3] implies that $\int \varphi\, d\left(m\circ h^{-1}\right) \subset \int \varphi\circ h\ \ dm$. In order to prove the converse inclusion we have to show that for every $x\in\int\varphi\circ h\ \ dm$ we can find an integrable selection $g\in\mathcal{L}_{\varphi\circ h}$ with $x\in\int g$ which is of the form: $g = f\circ h$, where f is a measurable function of T into $\mathbb{R}^n$. There exists a measurable function $f: T\to\mathbb{R}^n$ such that $g = f\circ h$ if and only if g is $h^{-1}(\mathfrak{J})$-measurable. Hence it remains to show that for every $g\in\mathcal{L}_{\varphi\circ h}$ there exists a $h^{-1}(\mathfrak{J})$-measurable selection of $\varphi\circ h$ with the same integral. But such a selection is easily found. Let $\mathcal{K} = h^{-1}(\mathfrak{J})$. Consider the conditional expectation $\mathbb{E}^{\mathcal{K}}g$ of g given the σ-algebra $\mathcal{K}$ (the conditional expectation is taken coordinatewise). By definition, $\mathbb{E}^{\mathcal{K}}g$ is a $\mathcal{K}$-measurable function of Ω into $\mathbb{R}^n$ and one has $\int \mathbb{E}^{\mathcal{K}}g\, dm = \int g\, dm$. Thus we have only to show that the function $\mathbb{E}^{\mathcal{K}}g$ is a selection for $\varphi\circ h$. Let $\psi = \varphi\circ h$. Since $g\in\mathcal{L}_{\psi}$, we obtain for every $v\in\mathbb{R}^n$ that $\inf v\psi(\omega) = vg(\omega)$ almost everywhere in Ω. By part (i)b, the graph of ψ belongs to $\mathcal{K}\otimes\mathcal{B}^n$. Recall that $\mathcal{B}$ denotes the Borel σ-algebra on $\mathbb{R}$ generated by the open subsets of $\mathbb{R}$. Hence, part (ii) implies that $\inf v\psi\,()$ is $\mathcal{K}_v$-measurable, and thus is almost everywhere equal to a $\mathcal{K}$-measurable function. Consequently, almost everywhere in Ω, depending on v, one obtains that

$$\inf v\psi(\omega) = \left(\mathbb{E}^{\mathcal{K}}\inf v\psi\right)(\omega) = \left(\mathbb{E}^{\mathcal{K}}v\psi\right)(\omega) = v\left(\mathbb{E}^{\mathcal{K}}g\right)(\omega).$$

[3] If M is a metric space, f is a measurable mapping of Ω into M and h is a measurable mapping of M into $\mathbb{R}$, then h is $m\circ f^{-1}$-integrable if and only if $h\circ f$ is m-integrable and $\int_M h\, dm = \int_\Omega h\ \circ f\, dm$.

Thus, if Q denotes a countable dense subset of $\mathbb{R}^n$, we have shown that, almost everywhere in Ω, one has
$\inf v\psi\,(\omega)\ = v\left(\mathbb{E}^{\mathcal{K}}g\right)(\omega)$, foe every $v \in Q$.
Since $\psi\,(\omega)$ is closed, convex and contains no straight line, this implies that almost everywhere in Ω, it holds
$\left(\mathbb{E}^{\mathcal{K}}g\right)(\omega) \in\ \psi\,(\omega)$.

We are now in position to prove the main result of this section.

Theorem 2.2 *For every data sector $\mathfrak{s} : (\mathbb{M}, \mathcal{A}, v\ ) \to \mathcal{P}{\times}\mathbb{R}$ in the complex S and every weighted data system with data significance vector $p \gg 0$ in S, one has*

(i) conv $\left(\overline{\mathfrak{A}}\,(\mathfrak{s}, p)\right) = conv\left(\int_{\mathcal{P}\times\mathbb{R}} \mathfrak{A}\,(\cdot, p)\ d\varrho\right)$, where $\varrho = v \circ \mathfrak{s}^{-1}$.

(ii) If the data sector $\mathfrak{s}$ is convex in S, then the mean rational data amount choice set $\overline{\mathfrak{A}}\,(\mathfrak{s}, p)$ in S is convex.

(iii) If $\underline{inf}\ \{p \cdot \mathfrak{X}_{\mathcal{M}}\} \le \mathcal{C}_{\mathcal{M}}$, a.e. in $\mathbb{M}$, then the mean rational data amount choice set $\overline{\mathfrak{A}}\,(\mathfrak{s}, p)$ in S is nonempty and compact.

Proof

(i) Since, by Lemma 2.1(ii)(a) and Lemma 2.1(i)(a), the mappings $\mathfrak{A}\,(\cdot, p)$ and $\mathfrak{A}\,(\mathfrak{s}\,(\cdot)\,, p)$ have both measurable graph and since they are bounded from below, we have

$$conv\left(\int \mathfrak{A}\,(\mathfrak{s}\,(\cdot)\,, p)\ dv\right) = \int conv\,(\mathfrak{A}\,(\mathfrak{s}\,(\cdot)\,, p))dv =$$

$$\int conv\,(\mathfrak{A}\,(\cdot, p))d\varrho = conv \int \mathfrak{A}\,(\cdot, p)\ d\varrho.$$

Indeed, the first and third equality follows from Lemma 2.1(vii). The second equality follows from the transformation formula of Lemma 2.1(viii), since, by Lemma 2.1(iii)(c), the graph of the mapping

$$conv\,(\mathfrak{A}\,(\cdot, p))$$

belongs to $\mathcal{B}_{\varrho}\,(\mathcal{P}{\times}\mathbb{R}) \times \mathcal{B}^{\ell}$.

(ii) It is easily seen that the measure space $(\mathbb{M}, \mathcal{A}, v\ )$ can be decomposed into a countable union of atoms and an atomless part. Since on atoms the data selection preferences are convex, the ***rational choice set of data amounts*** is also convex. Therefore, by Lemma 2.1(iv)(a), the mean rational data amount choice set $\overline{\mathfrak{A}}\,(\mathfrak{s}, p)$ is convex.

(iii) It remains to show that $\int \mathfrak{A}\,(\cdot, p)\ d\varrho$ is nonempty and compact. Since
$p \gg 0$ and $\mathfrak{X}_{\mathcal{M}} \le \mathcal{C}_{\mathcal{M}}$, ϱ-almost everywhere on $\mathcal{P}{\times}\mathbb{R}$,
The rational choice set $\mathfrak{A}\,(\mathfrak{X}, \succ\ , \mathcal{C}_{\mathcal{M}}, p\ )$ of data amounts is nonempty almost everywhere. Thus, by the measurable selection theorem and Lemma 2.1(v), the integral

$$\int \mathfrak{A}\,(\cdot, p)\ d\varrho$$

is nonempty and compact if the mapping $\mathfrak{A}(\cdot, p)$ is integrably bounded. To show this, one can assume without loss of generality that $\mathfrak{A}$ takes values only in $\mathbb{R}^{\ell}_{+}$. Then consider the function

$$\mathbb{V} : \mathcal{P} \times \mathbb{R} \to \mathbb{R}^{\ell} : (\mathfrak{X}, \succ, \mathcal{C}_{\mathcal{M}}) \mapsto \mathbb{V}(\mathfrak{X}, \succ, \mathcal{C}_{\mathcal{M}}) := \left(\frac{\mathcal{C}_{\mathcal{M}}}{p_1}, \ldots, \frac{\mathcal{C}_{\mathcal{M}}}{p_{\ell}}\right).$$

Clearly, $\mathfrak{A}(\mathfrak{X}, \succ, \mathcal{C}_{\mathcal{M}}, p) \leq \mathbb{V}(\mathfrak{X}, \succ, \mathcal{C}_{\mathcal{M}})$. Since, by assumption, $\int \mathcal{C}_{\mathcal{M}} d\varrho < \infty$, we conclude that the function $\mathbb{V}$ is ϱ-integrable.

3 Contrasting Selective Priorities

3.1 Introduction

In this chapter we will study the evaluation of selective priorities for the data amounts by several processors. To this end, let us consider a set $\mathbb{M}$ of data processors $\mathcal{M}$, each of whom is described by its focal data set $\mathfrak{X}_{\mathcal{M}}$ over a complex S, his corresponding data selection preference $\succ_{\mathcal{M}}$ over S and his available data amount over S. Hence, each data processor is characterized by an element in the space $\mathcal{P} \times \mathbb{R}^{\ell}$. *A* ***contrast of selective priorities for the data amounts*** *is defined to be a mapping of a finite set* $\mathbb{M}$ *of data processors into the space* $\mathcal{P} \times \mathbb{R}^{\ell}$ *of processors' characteristics.* For reasons which will become clear later, we shall also consider sets $\mathbb{M}$ of data processors which are *infinite*. Of course, in this case, the "total available data amount over *S*" is infinitely large. To overcome the problems that this creates, we shall replace the concept of "*total available data amount in S*" by that of "*mean available data amount in S*".

Obviously, the outcome of any contrast of selective priorities for the data amounts can be viewed as a ***redistribution*** of the initially available data amount over *S*. The analysis of a contrast, as presented here, consists of specifying a certain class of redistributions as possible outcomes and investigates two equilibrium concepts: *the collaborative concept* and *the non-collaborative concept.*

Let us consider first the *collaborative concept. The* ***contrast core of selective priorities for the data amounts*** *consists of those redistributions of the available data amount over S which no other group of data processors can "improve upon"*. A group of data processors can ***improve upon redistribution*** if the group, by using the data amount available to it, can make each member of that group better off, regardless of the actions of the data processors outside that group. Let us now turn to the *non- collaborative* concept. A ***contrast equilibrium of selective priorities for the data amounts*** *consists of a redistribution of the available data amount over S and a vector of weighted data such that no individual data processor acting independently can improve upon his situation when these weighted data prevail.* To say that ***certain weighted data prevail*** means that every data processor takes these weighted data as given (beyond his influence) and that there is a "program" where the data processors can use any amount of every weighted data by using these weighted data.

In Proposition 3.1, we show that *any contrast equilibrium of selective priorities belongs to the contrast core of selective priorities*. To study the converse, we need a meaning for the data concept of "***pure contrasting***", that is a set of data processors each of whom cannot influence the outcome of their collective activity but certain interplays of whom can influence that outcome. This leads logically to the concept of the ***partitionable data processing***, also called ***data processing with a "continuum of data processors."*** The essential result is *the identity of the contrast core of selective priorities and the set of contrast equilibriums of selective priorities for such a partitionable data processing* (see Theorem 3.1).

3.2 Main Definitions

In the context of pure contrasting, a data processor is described by a point in the space $\mathcal{P} \times \mathbb{R}^{\ell}$ the space of processors' characteristics. In order to simplify the presentation we shall often assume that

1. *the data focal set* $\mathfrak{X}$ *over a complex* S *is equal to the positive orthant* $\mathbb{R}^{\ell}_{+}$ and
2. *the vector* δ *of available data amount is* ≥ 0.

With this formalism we are in position to give a more rigorous definition for the concept of the contrasting of selective priorities.

Definition 3.1

(i) A ***contrast*** $\mathfrak{I}$ ***of selective priorities for the data amounts*** over the complex S is a measurable mapping $\mathfrak{I} : (\mathbb{M}, \mathcal{A}, v\,) \to \mathcal{P} \times \mathbb{R}^{\ell}$ of a measure space $(\mathbb{M}, \mathcal{A}, v\,)$, consisting of the set $\mathbb{M}$, a σ-algebra $\mathcal{A}$ of subsets of $\mathbb{M}$ and a (probability) measure v on $\mathcal{A}$, into the space $\mathcal{P} \times \mathbb{R}^{\ell}$ of data processors' characteristics such that the mean available data amount over S

$$\int_{\mathbb{M}} \delta \circ \mathfrak{I}\, dv$$

is finite.

(ii) An ***allocation for the*** **contrast** $\mathfrak{I}$ over S is an integrable function $f : (\mathbb{M}, \mathcal{A}, v\,) \to \mathbb{R}^{\ell}$ such that almost everywhere in $\mathbb{M}$, the focus vector $f(\mathcal{M})$ belongs to the data focal set of the data processor $\mathcal{M}$.

(iii) An allocation f for the contrast $\mathfrak{I}$ over S is called ***attainable*** or a ***state*** of the $\mathfrak{I}$ if

$$\int_{f(\mathbb{M})} f\, dv = \int_{\mathbb{M}} \delta \circ \mathfrak{I}\, dv.$$

(iv) A contrast $\mathfrak{I}$ of selective priorities over S is called

1. ***simple*** if the measure space $(\mathbb{M}, \mathcal{A}, v\,)$ is simple, i.e.

$\mathbb{M}$ is a finite set, $\mathcal{A}$ is the set of all subsets of $\mathbb{M}$ and $v(\mathcal{E}) = (|\mathcal{E}|/|\mathbb{M}|)$ whenever $\mathcal{E} \subset \mathbb{M}$;

1. ***partitionable*** if the measure space $(\mathbb{M}, \mathcal{A}, v)$ is atomless, i.e., for every $\mathcal{E} \in \mathcal{A}$ with $v(\mathcal{E})$ there is a $\mathcal{K} \subset \mathcal{E}$ such that $\mathcal{K} \in \mathcal{A}$ and $0 < v(\mathcal{K}) < v(\mathcal{E})$;
2. ***convex*** if almost all data processors of every atom of the measure space $(\mathbb{M}, \mathcal{A}, v)$ have convex data selection preferences.

The focal data set, data selection preference and totally available data amount S of a data processor $\mathcal{M}$ in $\mathbb{M}$ are denoted by

$$\mathfrak{I}(\mathcal{M}) = \left(\mathfrak{X}(\mathfrak{I}(\mathcal{M})),\ \succ_{\mathfrak{I}(\mathcal{M})}, \delta(\mathfrak{I}(\mathcal{M}))\right).$$

If it is clear which contrast $\mathfrak{I}$ is considered, we will shorten this
$(\mathfrak{X}(\mathcal{M}),\ \succ_{\mathcal{M}}, \delta(\mathcal{M}))$ or $(\mathfrak{X}_{\mathcal{M}},\ \succ_{\mathcal{M}}, \delta_{\mathcal{M}})$.

Definition 3.2

(i) Subsets of $\mathbb{M}$ belonging to $\mathcal{A}$ are called ***data processors' synergies***.
(ii) The distribution of $\mathfrak{I}$, i.e. the measure $v \circ \mathfrak{I}^{-1}$ on $\mathcal{P} \times \mathbb{R}^{\ell}$ is called the ***preference-availability distribution*** of the contrast $\mathfrak{I}$ and is denoted by $\mu_{\mathfrak{I}}$ or simply by μ.

If f is an allocation for the simple contrast $\mathfrak{I}$, then $f(\mathcal{M})$ denotes the vector of data significances allocated to the data processor $\mathcal{M}$ and $\int \delta\, dv = (1/|\mathbb{M}|) \sum_{\mathcal{M} \in \mathbb{M}} \delta_{\mathcal{M}}$ is the mean available data amount δ of the contrast $\mathfrak{I}$ *over S*. We emphasize that $\int_{\mathcal{E}} f\, dv$ does not mean the vector of data significances allocated to a synergy $\mathcal{E}$, indeed, if $\mathfrak{I}$ is a *simple* contrast of selective priorities, then

$$\int_{\mathcal{E}} f\, dv = (1/|\mathbb{M}|) \sum_{\mathcal{E} \in \mathcal{A}} f.$$

A partitionable contrast of selective priorities for the data amounts is, in fact, a quite abstract concept. The interpretation relies on analogy to the case of a simple contrast. As in the case of a simple contrast, $f(\mathcal{M})$ denotes the vector of data significances allocated to the data processor $\mathcal{M}$. The number $v(\mathcal{E})$ is interpreted as the fraction of the totality of data processors belonging to $\mathcal{E}$ and the integral $\int \delta\, dv$ is the mean available data amount δ of the contrast $\mathfrak{I}$ *over S*. Further, the σ-algebra $\mathcal{A}$ of synergies is introduced for technical measure theoretic reasons. As in the case of a simple contrast, there is no a priori restriction on possible synergies. Since for a partitionable measure space $(\mathbb{M}, \mathcal{A}, v)$, the set $\mathbb{M}$ must be uncountably infinite, we shall speak of a "*continuum of data processors*" as the set of participants. The results of Sects. 3.3 and 3.4 provide a strong justification for considering the partitionable contrast of selective priorities as the proper mathematical formulation of the concept of "pure contrasting", that is to say, a set of data processors, each of whom cannot influence the outcome of their collective activity but certain synergies can influence that outcome. The later concept is, in fact, as abstract as the former, which has the decisive advantage of being mathematically well defined.

A *convex* contrast of selective priorities has been defined in order to have a concise way of referring to a contrast that is either partitionable or simple with convex data selection preferences. From a formal point of view one might also consider a measure space $(\mathbb{M}, \mathcal{A}, v)$ with atoms and a non-atomic (atomless) part. In terms of the interpretation given earlier, one could consider a ***data processing atom*** or simply atom as a group of data processors which cannot split up; either all of them join a synergy or none does so. Note, however, that the mapping $\mathfrak{I}$, and also every allocation f, must be constant on an atom. This means that all data processors in the atom must have identical characteristics and must receive the same bundle in an allocation. This is so special a case that it makes the interpretation of atoms as synergies of little contrast significance. An alternative approach is to consider an atom as a "*big*" data processor. In the framework of the model under consideration "big" can only mean "big" in terms of the available data amount. Thus, *a data processing atom would be a data processor that has infinitely more available data amount than any data processor in the partitionable part.* Now, the measure $v(\mathcal{K})$ has a different interpretation. Formerly it expressed the relative number of data processors in the synergy $\mathcal{K}$, here it expresses something like the relative size of the available data amount of $\mathcal{K}$. Moreover the allocation $f(\mathcal{M})$ and the preferences $\succ_{\mathcal{M}}$ must also be reinterpreted. The net result is far from clear.

3.3 Contrast Core and Contrast Equilibriums

A state of contrast is clearly not in equilibrium if one data processor or a group of data processors could carry out decisions under the current circumstances and arrive at a position which is more advantageous to all members of the group than the current state. The underlying notion of equilibrium is based on the behavioral assumption that data processors want to improve their position, and that to achieve a preferred situation they are willing to cooperate. This notion of equilibrium leads to the basic concept of the contrast core of selective priorities for the data amounts. In such a case, the synergy of data processors can improve upon a redistribution of the available data amount over S if the synergy, by using the data amount available to it, can make each member better off. The contrast core of selective priorities for the data amounts is defined as the set of all redistributions that no synergy can improve upon. Formally:

Definition 3.3 Let

$$\mathfrak{I} : (\mathbb{M}, \mathcal{A}, v) \to \mathcal{P} \times \mathbb{R}^{\ell}$$

be a contrast of selective priorities for the data amounts over the complex S. Let also f be an allocation for $\mathfrak{I}$. The synergy $\mathcal{C} \in \mathcal{A}$ can ***improve upon*** the allocation f if there exists another allocation g for $\mathfrak{I}$ such that

1. $g(\mathcal{M}) \succ_{\mathcal{M}} f(\mathcal{M})$, almost everywhere in the synergy $\mathcal{C}$,
2. $v(\mathcal{C}) > 0$ and $\int_{\mathcal{C}} g\, dv = \int_{\mathcal{C}} \delta\, dv$.

The set of all attainable allocations for the contrast $\mathfrak{I}$ that no synergy in $\mathcal{A}$ can improve upon is called the ***core of selective priorities*** for the contrast $\mathfrak{I}$**,** or simply ***the contrast core*** for $\mathfrak{I}$**,** and is denoted by
$\mathfrak{C}(\mathfrak{I})$.

The meaning of the definitions of "improve" and core of selective priorities for a contrast is clear in the case of simple contrasts. In the framework presented here, all externalities of focal data sets are excluded (i.e., data selection preferences do not depend on the data amounts available to other data processors); the utility level of the members in a synergy does not depend on actions taken by data processors outside the synergies. The contrast core expresses what synergies can or cannot do for them, not what they can or cannot do to their opponents. Therefore, we used the term "to improve upon" and not "to block."

We now introduce a different concept of equilibrium. Suppose the following hold.

1. *Every weighted data system over S has a data significance p.*
2. *Every data processor in a contrast considers this data significance as given.*

Then, the data processor $\mathcal{M}$ with characteristics $(\mathfrak{X}_{\mathcal{M}}, \succ_{\mathcal{M}}, \delta_{\mathcal{M}})$ considers only vectors of data amounts in his set of data options

$$\{x \in \mathfrak{X}_{\mathcal{M}} :\ p \cdot x \leq p \cdot \delta_{\mathcal{M}}\}$$

and chooses a most desired vector in that set. If all these individually taken decisions—decentralized through the data significance system p—yield a situation where the rational choice set of data amounts equals the total supply we call that ***state of contrast equilibrium for the data amounts***. This concept of equilibrium is based on the behavioral assumption that data processors consider the data significance system as given and make their decisions independently of each other. The only link between these individual decisions is the data significance system. Formally: *An allocation f for the contrast $\mathfrak{I}$ over S is called*

Definition 3.4 An allocation f for the contrast $\mathfrak{I} : (\mathbb{M}, \mathcal{A}, v\ ) \to \mathcal{P} \times \mathbb{R}^{\ell}$ together with a data significance system $p \in \mathbb{R}^{\ell}$ is said to be an ***equilibrium of contrasts of selective priorities*** for $\mathfrak{I}$, or simply a ***contrast equilibrium*** for $\mathfrak{I}$**,** if the following two conditions are satisfied.

1. $f(\mathcal{M}) \in \mathfrak{A}\left(\mathfrak{X}_{\mathcal{M}},\ \succ_{\mathcal{M}}, \underbrace{p \cdot \delta_{\mathcal{M}}}_{\mathcal{C}_{\mathcal{M}}}, p\right)$ almost everywhere in $\mathbb{M}$,

i.e., $f(\mathcal{M})$ is a maximal element for $\succ_{\mathcal{M}}$ in the set

$$\mathfrak{B} = \mathfrak{B}(\mathfrak{X}, \mathcal{C}_{\mathcal{M}}, p) := \{x = (d_1, \ldots, d_\ell) \in \mathfrak{X} :$$

$$(p_1, p_2, \ldots, p_\ell) \cdot (d_1, d_2, \ldots, d_\ell) \leq \mathcal{C}_{\mathcal{M}}\}$$

of data options of $\mathcal{M}$.

1. $\int_{f(\mathbb{M})} f\, dv = \int_{\mathbb{M}} \delta\, dv$,

i.e., mean rational data amount choice equals mean data availability.
The allocation f for the contrast $\mathfrak{I}$ is called a ***contrast allocation*** if there exists a weight vector $p \in \mathbb{R}^{\ell}$ such that (f, p) is an equilibrium of contrasts of selective priorities for $\mathfrak{I}$. The set of all equilibriums of contrasts of selective priorities for $\mathfrak{I}$ is denoted by

$$\mathfrak{W}(\mathfrak{I}).$$

A data significance system $p \in \mathbb{R}^{\ell}$ is said to be a ***data equilibrium of significance levels*** or simply a ***data equilibrium vector*** for the contrast $\mathfrak{I}$ if there exists an allocation f for $\mathfrak{I}$ such that (f, p) is an equilibrium of contrasts of selective priorities for $\mathfrak{I}$. The set of all data equilibrium vectors for $\mathfrak{I}$ which are normalized, i.e.$|p| = 1$, is denoted by
$\mathfrak{e}(\mathfrak{I})$.

Under what conditions is justified the behavioral assumption that the data processors can adapt to the prevailing system of data significance system? The obvious answer is the following: *data processors take significances as given if they have no influence on them*. This leads to a partitionable contrast of selective priorities. In that case, one could ask what is special about the contrast allocations among the allocations that cannot be improved upon. We shall show that there is nothing special; *every deviation from a contrast allocation can be improved upon*. That is to say, *contrast allocations, and only they, belong to the contrast core of selective priorities*, i.e.

$$\mathfrak{W}(\mathfrak{I}) = \mathfrak{C}(\mathfrak{I}).$$

One part of the identity is trivial:

Proposition 3.1 *For every contrast of selective priorities for the data amounts $\mathfrak{I}$ in S, we have*

$$\mathfrak{W}(\mathfrak{I}) \subset \mathfrak{C}(\mathfrak{I}).$$

Proof Let $f \in \mathfrak{W}(\mathfrak{I})$ but $f \notin \mathfrak{C}(\mathfrak{I})$. Thus, there is a synergy $\mathcal{C} \in \mathcal{A}$, $v(\mathcal{C}) > 0$, and there is an allocation g such that

1. $g(\mathcal{M}) \succ_{\mathcal{M}} f(\mathcal{M})$, almost everywhere in the synergy $\mathcal{C}$
2. $v(\mathcal{C}) > 0$ and $\int_{g(\mathcal{C})} g\, dv = \int_{\mathcal{C}} \delta\, dv$.

By (i) and the definition of a contrast allocation, we obtain
$p \cdot \delta(\mathcal{M}) < p \cdot g(\mathcal{M})$ almost everywhere in $\mathcal{C}$,
where p denotes an equilibrium significance associated with f. Hence

$$p \cdot \int_{\mathcal{C}} \delta\, dv < p \cdot \int_{\mathcal{C}} g\, dv,$$

which contradicts (ii).

The central result of this section is proved in the following.

Theorem 3.1 *Let* $\mathfrak{I} : (\mathbb{M}, \mathcal{A}, v) \to \mathcal{P}_{mo} \times \mathbb{R}_+^{\ell}$ *be a partitionable contrast of selective priorities for the data amounts in S with*

$$\int \delta dv \gg 0.$$

Then

$$\mathfrak{W}(\mathfrak{I}) = \mathfrak{C}(\mathfrak{I}).$$

Proof By Proposition 3.1, it is enough to show that $f \in \mathfrak{C}(\mathfrak{I})$ implies $f \in \mathfrak{W}(\mathfrak{I})$. Consider for every data processor $\mathcal{M} \in \mathbb{M}$, the sets
$\prec_{\mathcal{M}}(f) := \{x \in \mathfrak{X}_{\mathcal{M}} : x \succ_{\mathcal{M}} f(\mathcal{M})\}$ and $\mathfrak{h}(\mathcal{M}) := \{\prec_{\mathcal{M}}(f) - \delta(\mathcal{M})\} \bigcup \{0\}$.
Since the measure space $(\mathbb{M}, \mathcal{A}, v)$ is atomless, the integral $\int \mathfrak{h} dv$ is a convex subset in $\mathbb{R}^{\ell}$. Since $0 \in \int \mathfrak{h} dv$, it is clear that

$$\int \mathfrak{h} dv \neq \mathfrak{A}.$$

We now claim that $\int \mathfrak{h} dv \bigcap \mathbb{R}_-^{\ell} = \{0\}$. Assume to the contrary that there is an integrable function h in the set $\mathcal{L}_{\mathfrak{h}}$ of all integrable selections of $\mathfrak{h}$, (that is of the set of all v-integrable $h : \mathbb{M} \to \mathbb{R}^{\ell}$ which have the property that $h(\mathcal{M}) \in \mathfrak{h}(\mathcal{M})$ almost everywhere in $\mathbb{M}$), with $\int \mathfrak{h} dv < 0$. Then the interplay $\mathcal{C} = \{\mathcal{M} \in \mathbb{M} : \mathfrak{h}(\mathcal{M}) \neq 0\}$ can improve the allocation f with the allocation

$$g(\mathcal{M}) = \mathfrak{h}(\mathcal{M}) + \delta(\mathcal{M}) - \frac{\int \mathfrak{h} dv}{v(\mathcal{C})}.$$

Indeed, $v(\mathcal{C}) > 0$, $g(\mathcal{M}) \succ_{\mathcal{M}} f(\mathcal{M})$ for every $\mathcal{M} \in \mathcal{C}$ and $\int_{\mathcal{C}} g\, dv = \int_{\mathcal{C}} \delta\, dv$. Consequently, there exists a hyperplane separating the two convex sets $\int \mathfrak{h} dv$ and $\mathbb{R}_-^{\ell}$, i.e. there is a vector $\mathfrak{p} \in \mathbb{R}^{\ell}$, $\mathfrak{p} \geq 0$, $\mathfrak{p} \neq 0$, such that

$$0 \leq \mathfrak{p} \cdot z \quad for \quad every \quad z \in \int \mathfrak{h} dv. \tag{5}$$

The graph of the mapping $\mathfrak{h}$ is measurable. Indeed the set

$$G := \left\{(\mathfrak{X}, \succ, x, y) \in \mathcal{P}_{mo} \times \mathbb{R}_+^{\ell} \times \mathbb{R}_+^{\ell} : x \succ y\right\}$$

is a Borel set in $\mathcal{P}_{mo} \times \mathbb{R}_+^{\ell} \times \mathbb{R}_+^{\ell}$. Now the graph of the mapping $\mathcal{M} \mapsto \mathfrak{h}(\mathcal{M}) \setminus \{0\}$, i.e., the set

$$\left\{(\mathcal{M}, x) \in \mathbb{M} \times \mathbb{R}^{\ell+1} : x + \delta(\mathcal{M}) \succ_{\mathcal{M}} f(\mathcal{M})\right\}$$

is equal to $\mathfrak{h}^{-1}(G)$ where $\mathfrak{h}$ is a mapping of $\mathbb{M} \times \mathbb{R}^{\ell}$ into $\mathbb{M} \times \mathbb{R}^{\ell} \times \mathbb{R}^{\ell}$ defined by

$$\mathfrak{h}(\mathcal{M}, x)?\left(\mathfrak{X}_{\mathcal{M}}, \succ_{\mathcal{M}}, x + \delta(\mathcal{M}), f(\mathcal{M})\right).$$

Clearly the mapping $\mathfrak{h}$ is measurable, and hence the graph of $\mathfrak{h}$ is measurable. Therefore it follows

$$inf_{z \in \int \mathfrak{h} dv}\, \mathfrak{p} \cdot z = \int inf_{x \in \mathfrak{h}(\cdot)}\, \mathfrak{p} \cdot x\, dv.$$

Consequently, we obtain from (5) that $0 \leq \int \inf \mathfrak{p} \cdot \mathfrak{h}\ \ dv$. Since by definition the set $\mathfrak{h}(\mathcal{M})$ contains 0, we clearly have $\inf p \cdot \mathfrak{h}(\mathcal{M})\ \leq 0$. Hence, it follows that, almost everywhere in $\mathbb{M}$, $\inf \mathfrak{p} \cdot \mathfrak{h}(\mathcal{M})\ = 0$. Thus, we have shown that

$$almost\ \ everywhere\ \ in\ \ \mathbb{M}, \mathfrak{p} \cdot \delta(\mathcal{M}) \leq \mathfrak{p} \cdot x\ \ for\ \ every\ \ x \succ_{\mathcal{M}} f(\mathcal{M}). \tag{6}$$

It follows from (6) that almost everywhere in $\mathbb{M}$,

$$\mathfrak{p} \cdot \delta(\mathcal{M}) = \mathfrak{p} \cdot f(\mathcal{M}).$$

Indeed, first we obtain from (6) that $\mathfrak{p} \cdot \delta(\mathcal{M}) \leq \mathfrak{p} \cdot f(\mathcal{M})$ almost everywhere in $\mathbb{M}$. Now, if $\mathfrak{p} \cdot \delta(\mathcal{M}) < \mathfrak{p} \cdot f(\mathcal{M})$for a set of data processors with positive measure, then we obtain

$$\mathfrak{p} \cdot \int \delta\, dv < \mathfrak{p} \cdot \int f\, dv,$$

which contradicts $\int \delta\, dv = \int f\, dv$. Since by assumption $\int \delta\, dv \gg 0$ and since $\mathfrak{p} \geq 0$, $\mathfrak{p} \neq 0$, we surely have

$$v\{\mathcal{M} \in \mathbb{M} :\ \mathfrak{p} \cdot \delta(\mathcal{M}) > 0\} > 0.$$

But for a data processor $\mathcal{M}$ with positive income, i.e., $\mathfrak{p} \cdot \delta(\mathcal{M}) > 0$, property (6) implies that

$$f(\mathcal{M}) \in \mathfrak{A}(\mathfrak{X}, \succ_{\mathcal{M}}, \mathfrak{p} \cdot \delta_{\mathcal{M}}(\mathcal{M}), \mathfrak{p})\ \ .$$

Indeed, for $x \in \mathbb{R}_{+}^{\ell}$ with $\mathfrak{p} \cdot x < \mathfrak{p} \cdot \delta(\mathcal{M})$, it follows from (6) that

$$x \not\succ_{\mathcal{M}} f(\mathcal{M}).$$

Since in the case $\mathfrak{p} \cdot \delta(\mathcal{M}) > 0$ for every $x \in \mathbb{R}_{+}^{\ell}$ with $\mathfrak{p} \cdot x = \mathfrak{p} \cdot \delta(\mathcal{M})$ is limit of a sequence (x_n) with $\mathfrak{p} \cdot x_n < \mathfrak{p} \cdot \delta(\mathcal{M})$, the continuity of the selection preference relation $\succ_{\mathcal{M}}$ implies $x \not\succ_{\mathcal{M}} f(\mathcal{M})$. Thus $f(\mathcal{M})$ is a maximal element for $\not\succ_{\mathcal{M}}$ in the set of data amount processing options $\left\{x \in \mathbb{R}_{+}^{\ell} : \mathfrak{p} \cdot x \leq \mathfrak{p} \cdot \delta(\mathcal{M})\right\}$. This, together with the monotony of the selection preferences, implies that $\mathfrak{p} \gg 0$. Hence $f(\mathcal{M})$ belongs to the rational choice set of data amounts $\mathfrak{A}(\mathfrak{X}, \succ_{\mathcal{M}}, \mathfrak{p} \cdot \delta_{\mathcal{M}}(\mathcal{M}), \mathfrak{p})$ even in the case $\mathfrak{p} \cdot \delta(\mathcal{M}) = 0$ since, by (6), the vector $\delta(\mathcal{M})$ belongs to the set of data amount processing options which in this case is equal to $\{0\}$. This proves that $(f, \mathfrak{p})$ is an equilibrium of contrasts of selective priorities for $\mathfrak{I}$.

3.4 Determinateness of Data Equilibrium Vectors

In this section we will investigate the existence of data equilibrium vectors for a contrast $\mathfrak{I}$ of selective priorities for the data amounts over the complex S with particular emphasis on the case where the data selection preferences are not assumed to be convex. Clearly, the classical assumption of convex preferences cannot simply be dropped. Indeed, in a contrast of selective priorities, where the influence of a certain individual data processor cannot be neglected, the convexity of his preferences is essential in proving the existence of data equilibrium vectors. The extreme case, where the contrast of selective priorities for the data amounts is partitionable, is particularly simple (see Theorem 3.2 below).

The results of this section will show *the important role that plays the number of participants into a contrast (of selective priorities for the data amounts) to the issue of existence of data equilibrium vectors, in the case where the selection preferences are not convex*. Proposition 3.3 and its consequences will show that the *data equilibrium vectors depend in a continuous way on the data defining the contrast of selective priorities*.

Let us introduce some notation. We shall write

$$\mathfrak{A}(t, p) \equiv \mathfrak{A}\left(\underbrace{\mathfrak{X}, \succ, p \cdot \delta}_{t}, p\right)$$

(instead of $= (\mathfrak{X}, \succ, p \cdot \delta) \in \mathcal{P} \times \mathbb{R}^{\ell+1}$ and $p \in \mathbb{R}^{\ell}$).

Consequently, given the data equilibrium vector p, the mean rational choice set of data amounts of a contrast $\mathfrak{I} : (\mathbb{M}, \mathcal{A}, v) \to \mathcal{P} \times \mathbb{R}^{\ell}$, of selective priorities for the data amounts is denoted by

$$\mathfrak{u}(\mathfrak{I}, p) := \int \mathfrak{A}(\mathfrak{I}(\cdot), p)\, dv.$$

Given $\mathfrak{I}$ and p, the mapping

$$\mathfrak{I}^{p} : (\mathbb{M}, \mathcal{A}, v) \to \mathcal{P} \times \mathbb{R} : \mathfrak{I}^{p}(\mathcal{M}) := \left(\mathfrak{X}_{\mathfrak{I}(\mathcal{M})}, \succ_{\mathcal{M}}, p \cdot \delta_{\mathfrak{I}(\mathcal{M})}\right)$$

defines a data sector in the complex S. With the Notation 2.4(v) of Sect. 2.6, we clearly have

$$\mathfrak{u}(\mathfrak{I}, p) = \overline{\mathfrak{A}}\left(\mathfrak{I}^{p}, p\right).$$

However, the sets
$\int_{\mathcal{P}\times\mathbb{R}^{\ell}} \mathfrak{u}\,(t,\, p)d\varrho_{\mathfrak{I}}$ and $\int_{\mathcal{P}\times\mathbb{R}} \mathfrak{u}\,(\cdot,\ \cdot,\, p)d\varrho_{\mathfrak{I}^{p}}$
may well be defined; only their convex hulls are identical.

It is easy to prove the following.

Proposition 3.2 *If the sequence* $(\mathfrak{I}_n)_{n\in\mathbb{N}}$ *of contrasts of selective priorities converges in distribution to the contrast* $\mathfrak{I}$ *of selective priorities for the data amounts and if* $lim_{n\to\infty}p_{\mathbf{n}}{=}p$*, then the sequence* $\left(\mathfrak{I}_n^{p_{\mathbf{n}}}\right)_{n\in\mathbb{N}}$ *of the data sectors in the complex* S *converges in distribution to the data sector* $\mathfrak{I}^{p}$ *in the complex* S*.*

Before giving the main result on the existence of data equilibrium vectors for a contrast $\mathfrak{I}$ of selective priorities for the data amounts over the complex S, we need some preparatory material.

Definition 3.5 Let $\mathfrak{I}$ be a contrast of selective priorities for the data amounts over the complex S. For every weighted data system with significance vector $p{=}\,(p_1, p_2, \ldots, p_\ell)$, we define the ***mean excess rational data choice*** $\mathcal{Z}\,(\,p)$ by
$\mathcal{Z}\,(p) :=\ \mathfrak{u}\,(\mathfrak{I}, p) - \int \delta\, dv.$

As it is readily seen, if $0\ \in\ \mathcal{Z}\,(p^{*})$, then *a significance vector* $p^{*}\ =\ \left(p_1^{*},\ p_2^{*}, \ldots, p_\ell^{*}\right)$ *is a data equilibrium vector for a contrast* $\mathfrak{I}$ *of selective priorities for the data amounts over the complex* S. The existence of data significances for the contrast $\mathfrak{I}$ therefore depends on properties of the mean excess rational data choice relation $\mathcal{Z}$. The relevant properties of $\mathcal{Z}$ are summarized in the following.

Proposition 3.3 *Let* $\mathfrak{I}\ :\ (\mathbb{M}, \mathcal{A}, v\,)\ \to\ \mathcal{P}_{mo}\ \times\ \mathbb{R}_{+}^{\ell}$ *be a contrast of selective priorities for the data amounts over the complex* S *with* $\int \delta\ dv\ \gg\ 0$*. Then the mean excess rational data choice mapping* $\mathcal{Z}$ *has the following properties.*

1. $\mathcal{Z}$ *is homogeneous of degree zero (i.e., for every* $p \gg 0$ *and* $\lambda > 0$ *one has*

$$\mathcal{Z}\,(p) = \mathcal{Z}\,(\lambda p)).$$

2. *For every weighted data system with data significance vector*

$$p{=}\,(p_1, p_2, \ldots, p_\ell) \gg 0$$

and

$$z \in \mathcal{Z}\,(p)$$

one has

$$p \cdot z = 0.$$

3. *The mapping $\mathcal{Z}$ is compact-valued, bounded from below and upper hemi-continuous*[4].
4. *If the sequence $\left(p^{(n)} = \left(p_1^{(n)} p_2^{(n)}, \ldots, p_\ell^{(n)}\right)\right)_{n\in\mathbb{N}}$ of strictly positive significance vectors converges to p which is not strictly positive, then*

$inf_{n\in\mathbb{N}}\left\{\sum_{i=1}^{\ell} z_i : z \in \mathcal{Z}\left(p^{(n)}\right)\right\} > 0$ *for n large enough.*

Proof Property (i) follows immediately from the definition of the set $\mathcal{Z}(p)$. Since data selection preferences are monotonic, we have $p \cdot x = p \cdot \delta(\mathcal{M})$ for every $x \in \mathfrak{A}(\mathfrak{I}(\mathcal{M}), p)$. This clearly implies property (ii). Let now $\overline{p} \gg 0$. Then there is a neighborhood $U_{\overline{p}}$ of $\overline{p}$ consisting of strictly positive vectors. For any fixed $\mathcal{M} \in \mathbb{M}$, the mapping $p \longmapsto \mathfrak{u}(\mathfrak{I}(\mathcal{M}), p)$ is closed at $\overline{p}$. Further, there is an integrable real function h of $\mathbb{M}$ such that
$|\mathfrak{u}(\mathfrak{I}(\mathcal{M}), p)| \leq h(\mathcal{M})$ whenever $\mathcal{M} \in \mathbb{M}$ and $p \in U_{\overline{p}}$,
e.g.

$$h(\mathcal{M}) = \frac{1}{min\left\{p_i : p \in U_{\overline{p}},\ i = 1, 2, \ldots, \ell\right\}}\ |\delta(\mathcal{M})|.$$

Thus, the mapping $p \longmapsto \mathfrak{u}(\mathfrak{I}, p)$ is closed at $\overline{p}$. Since the correspondence $\mathfrak{u}(\mathfrak{I}, \cdot)$ is bounded on the neighborhood $U_{\overline{p}}$ of $\overline{p}$, it follows that $\mathfrak{u}(\mathfrak{I}, \cdot)$ is compact-valued and upper hemi-continuous at $\overline{p}$. This clearly implies property (iii). Finally, property (iv) follows from the fact that the data selection preferences are assumed to be monotone and $\int \delta\, dv \gg 0$.

The following Proposition is the fundamental mathematical result in contrast equilibrium analysis.

Proposition 3.4 *Let $\mathcal{Z}$ be a mapping of*

$$int\Delta = \left\{p = (p_1, p_2, \ldots, p_\ell) \in \mathbb{R}_+^\ell : \sum_{i=1}^{\ell} p_i = 1\right\}$$

into $\mathbb{R}_+^\ell$ which has the properties (ii), (iii) and (iv) of Proposition 3.3. Then there exists a vector $p^* \gg 0$ such that

$0 \in conv\mathcal{Z}(p^*)$ (=the convex hull of $\mathcal{Z}(p^*)$).

Proof For any $n \geq \ell$, we set

$$\Delta_{\mathbf{n}} := \left\{p = (p_1, p_2, \ldots, p_\ell) \in \mathbb{R}_+^\ell : \sum_{i=1}^{\ell} p_i = 1 \ and\ p_i \geq \frac{1}{n}\ \forall i = 1, 2, \ldots, \ell\right\}.$$

[4] A relation φ of the metric space M into the metric space N is said to be upper hemi-continuous at $x \in$ M if $\varphi(x) \not\equiv \varphi$ and if for every neighborhood $U_{\varphi(x)}$ of $\varphi(x)$ there exists a neighborhood U_x of x such that $\varphi(U_x) \subset U_{\varphi(x)}$. A relation ψ of the metric space M into the metric space N is said to be lower hemi-continuous at $x \in$ M if $\psi(x) \not\equiv \psi$ and if for every open set G in N with $\psi(x) \cap G \neq \psi$ there exists a neighborhood U_x of x such that $\psi(U_x) \cap G \neq \psi$.

Applying the fixed point theorem to the map $\triangle_{\mathbf{n}} \to \mathbb{R}^{\ell}: p \mapsto conv\mathcal{Z}\,(p)$, we infer the existence of vectors
$p^{(n)} = \left(p_1^{(n)} p_2^{(n)}, \ \ldots, p_{\ell}^{(n)}\right) \in \triangle_{\mathbf{n}}$ and $\mathbf{z}^{(n)} = \left(\mathbf{z}_1^{(n)} \mathbf{z}_2^{(n)}, \ \ldots, \mathbf{z}_{\ell}^{(n)}\right) \in \mathbb{R}^{\ell}$
such that

$$\mathbf{z}^{(n)} \in conv\mathcal{Z}\left(p^{(n)}\right) \tag{7}$$

and

$$p \cdot \mathbf{z}^{(n)} \leq 0 \ \ for \ \ every \ \ p \in \triangle_{\mathbf{n}}(n \geq \ell). \tag{8}$$

It suffices to show that $\mathbf{z}^{(n)} = 0$ for some n. Without of generality, we can assume that the sequence $\left(p^{(n)} = \left(p_1^{(n)} p_2^{(n)}, \ \ldots, p_{\ell}^{(n)}\right)\right)_{n\in\mathbb{N}}$ is convergent, say $lim_{n\to\infty} p^{(n)} = p \in \Delta$. One may claim that $p \gg 0$. Otherwise, it would follow that $\sum_{i=1}^{\ell} \mathbf{z}_i^{(n)} > 0$ for n large enough, which contradicts (8). Since $p \gg 0$, it follows that
$\mathbf{z}^{(n)} = 0$ for n large enough.
Indeed, let $\overline{n}$ be such that $int\triangle_{\overline{n}}$ contains p. Clearly, we have $p^{(n)} \cdot \mathbf{z}^{(n)} = 0$, and, since for n large enough, $p^{(n)} \in int\triangle_{\overline{n}}$, it follows, from (8), that $\mathbf{z}^{(n)} = 0$.

As an immediate consequence of Propositions 3.3 and 3.4, we have the following result.

Theorem 3.2 *Let $\mathfrak{I} : (\mathbb{M}, \mathcal{A}, v\ ) \to \mathcal{P}_{mo} \times \mathbb{R}_{+}^{\ell}$ be a contrast of selective priorities for the data amounts over the complex S with $\int \delta\, dv \gg 0$. Then there exists an equilibrium (f, p^{*}) of contrasts of selective priorities for $\mathfrak{I}$, with $p^{*} \gg 0$.*

Proof It suffices only to note that the mean rational data amount choice $\overline{\mathfrak{A}}\,(\mathfrak{I}, p^{*})$ of a convex data processing is convex (Theorem 2.2).

Corollary 3.1 *The core is nonempty for every convex contrast of selective priorities for the data amounts $\mathfrak{I} : (\mathbb{M}, \mathcal{A}, v\ ) \to \mathcal{P}_{mo} \times \mathbb{R}_{+}^{\ell}$ over the complex S with $\int \delta\, dv \gg 0$.*

References

1. M.P. Barcellos et al., A well-founded software measurement ontology, in *Formal Ontology in Information Systems, Proceedings of the Sixth International Conference, FOIS 2010*, Toronto, ed. by A. Galton, R. Mizoguchi (2010)
2. V.I. Bogachev, *Measure Theory*, vols. 1 and 2 (Springer, New York, 2007), 1105 pp.
3. P.R. Halmos, *Measure Theory*. Graduate Texts in Mathematics, vol. 18 (Springer, Berlin, 1974), 304 pp. ISBN 03-540-90088-8
4. P.A. Meyer, *Probability and Potentials* (Blaisdell Publishing Company, London, 1966)
5. T. Suzuki, *General Equilibrium Analysis of Production and Increasing Returns* (World Scientific, Business & Economics, Singapore, 2009), 285 pp.

General Inertial Mann Algorithms and Their Convergence Analysis for Nonexpansive Mappings

Qiao-Li Dong, Yeol Je Cho, and Themistocles M. Rassias

1 Introduction

Let $\mathscr{H}$ be a Hilbert space and C be a nonempty closed convex subset of $\mathscr{H}$. A mapping $T : C \to C$ is said to be *nonexpansive* if

$$\|Tx - Ty\| \leq \|x - y\|$$

for all $x, y \in C$ and $Fix(T) := \{x \in C : Tx = x\}$ denotes the set of fixed points of T.

In this paper, we consider the following fixed point problem:

Problem 1 Suppose that $T : C \to C$ is a nonexpansive mapping with $Fix(T) \neq \emptyset$. Find a point $x^* \in C$ such that

$$T(x^*) = x^*.$$

Approximating fixed point problems for nonexpansive mappings has a variety of specific applications since many problems can be seen as a fixed point problem of

Q.-L. Dong
College of Science, Civil Aviation University of China, Tianjin, China
e-mail: dongql@lsec.cc.ac.cn

Y. J. Cho (✉)
Department of Mathematics Education and RINS, Gyeongsang National University, Jinju, South Korea
Center for General Education, China Medical University, Taichung, Taiwan
e-mail: yjcho@gnu.ac.kr

Th. M. Rassias
Department of Mathematics, National Technical University of Athens, Athens, Greece
e-mail: trassias@math.ntua.gr

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_7

nonexpansive mappings such as convex feasibility problems, monotone variational inequalities (see [3, 4] and references therein). In 2011, Micchelli et al. [27] proposed fixed-point framework in the study of the total-variation model for image denoising and finding a fixed point of a nonexpansive mapping was embedded in their algorithms. Recently, in 2013 and 2016, Chen et al. [9, 10] showed the convergence of the primal-dual fixed point algorithms with aid of the fixed point theories of the nonexpansive mappings. A great deal of literature on the iteration methods for fixed points problems of nonexpansive mappings have been published (for example, see [11, 13, 15, 16, 19, 30–32, 35, 38]).

One of the most used algorithms is the *Mann algorithm* [20, 22] as follows:

$$x_{n+1} = \alpha_n x_n + (1 - \alpha_n) T x_n \tag{1}$$

for each $n \geq 0$. The iterative sequence $\{x_n\}$ converges weakly to a fixed point of T provided that $\{\alpha_n\} \subset [0, 1]$ satisfies $\sum_{n=1}^{\infty} \alpha_n (1 - \alpha_n) = +\infty$.

In generally, the convergence rate of the Mann algorithm is very slow, especially, for large scale problems. In 2014, Sakurai and Liduka [36] pointed out that, to guarantee practical systems and networks (see, for example, [17, 18]) stable and reliable, the fixed point has to be quickly found. So, there are increasing interests in study of fast algorithms for approximating fixed points of nonexpansive mappings.

To the best of our knowledge, there are two main ways to speed up the Mann algorithm. One way is to combine conjugate gradient methods [29] and the Mann algorithm to construct the accelerated Mann algorithm (see [12]). We will make further analysis of the accelerated Mann algorithm in Sect. 3. Another way is to combine the inertial extrapolation with Mann algorithm.

Consider the following *minimization problem*:

$$\min \varphi(x) \tag{2}$$

for all $x \in \mathscr{H}$, where $\varphi(x)$ is differentiable. There are many methods to solve the problem (2), the most popular two methods among which are the steepest descent method and the conjugate gradient method. The later is a popular acceleration method of the former.

To accelerate speed of convergence of the algorithms, multi-step methods have been proposed in the literature, which can usually be viewed as certain discretizations of the second-order dynamical system with friction:

$$\ddot{x}(t) + \gamma \dot{x}(t) + \nabla \varphi(x(t)) = 0,$$

where $\gamma > 0$ represents a friction parameter. One of the simplest method is the two-step heavy ball method, in which, given x_n and x_{n-1}, the next point x_{n+1} is determined via

$$\frac{x_{n+1} - 2x_n + x_{n-1}}{h^2} + \gamma \frac{x_n - x_{n-1}}{h} + \nabla \varphi(x_n) = 0,$$

which results in an iterative algorithm of the form

$$x_{n+1} = x_n + \beta(x_n - x_{n-1}) - \alpha\nabla\varphi(x_n) \tag{3}$$

for each $n \geq 0$, where $\beta = 1 - \gamma h$ and $\alpha = h^2$. In 1964, Polyak [33] firstly used (3) to solve the minimization problem (2) and called it an *inertial type extrapolation algorithm*. In 1987, Polyak [33, 34] also considered the relation between the heavy ball method and the following conjugate gradient method:

$$x_{n+1} = x_n + \beta_k(x_n - x_{n-1}) - \alpha_k\nabla\varphi(x_n) \tag{4}$$

for each $n \geq 0$, where α_k and β_k can be chosen through different ways. It is obvious that the only difference between the heavy ball method (3) is the choice of the parameters.

From Polyak's work, as an acceleration process, the inertial extrapolation algorithms were widely studied. Especially recently, researchers constructed many iterative algorithms by using inertial extrapolation, such as inertial forward-backward algorithm [2, 7, 21], inertial extragradient methods [14] and fast iterative shrinkage thresholding algorithms (FISTA) (see [5, 8]). The inertial extrapolation algorithm is a two-step iterative method and its main feature is that the next iterate is defined by making use of the previous two iterates.

By using the technique of the inertial extrapolation, in 2008, Mainge [23] introduced the classical inertial Mann algorithm:

$$\begin{cases} y_n = x_n + \alpha_n(x_n - x_{n-1}), \\ x_{n+1} = (1 - \lambda_n)y_n + \lambda_n T(y_n) \end{cases} \tag{5}$$

for each $n \geq 1$. He showed that $\{x_n\}$ converges weakly to a fixed point of T under the following conditions:

(B1) $\alpha_n \in [0, \alpha)$ for each $n \geq 1$, where $\alpha \in [0, 1)$;
(B2) $\sum_{n=1}^{\infty} \alpha_n \|x_n - x_{n-1}\|^2 < +\infty$;
(B3) $\inf_{n\geq 1} \lambda_n > 0$ and $\sup_{n\geq 1} \lambda_n < 1$.

For satisfying the summability condition (B2) of the sequence $\{x_n\}$, one need to calculate α_n at each step (see [28]). In 2015, Bot and Csetnek [7] got rid of the condition (B2) and substituted (B1) and (B3) with the following conditions, respectively:

(C1) for each $n \geq 1$, $\{\alpha_n\} \subset [0, \alpha]$ is nondecreasing with $\alpha_1 = 0$ and $0 \leq \alpha < 1$;
(C2) for each $n \geq 1$,

$$\delta > \frac{\alpha^2(1+\alpha) + \alpha\sigma}{1 - \alpha^2}, \quad 0 < \lambda \leq \lambda_n \leq \frac{\delta - \alpha[\alpha(1+\alpha) + \alpha\delta + \sigma]}{\delta[1 + \alpha(1+\alpha) + \alpha\delta + \sigma]},$$

where $\lambda, \sigma, \delta > 0$.

In this paper, we introduce a general inertial Mann algorithm which includes the classical inertial Mann algorithm and the accelerated Mann algorithm as special cases. The numerical experiments show that the accelerated Mann behaves better than other algorithms.

The structure of the paper is as follows. In Sect. 2, we present some lemmas which will be used in the main result. In Sect. 3, we revisit first the accelerated Mann algorithm and show that it is an inertial type algorithm. Then we analyze the relationship between the general inertial Mann algorithm with some other ones. The weak convergence of the general inertial Mann algorithm is discussed in Sect. 4. We apply the general inertial Mann algorithm to the minimization problems and propose a general inertial type gradient-projection algorithm in Sect. 5. In the final section, Sect. 6, some numerical results are provided, which give the best choice of the parameters in the general inertial Mann algorithm.

2 Preliminaries

We use the notation:

1. $\rightharpoonup$ for weak convergence and $\to$ for strong convergence;
2. $\omega_w(x^k) = \{x : \exists x^{k_j} \rightharpoonup x\}$ denotes the weak ω-limit set of $\{x^k\}$.

The following identity will be used several times in the paper (see Corollary 2.14 of [4]):

$$\|\alpha x + (1-\alpha)y\|^2 = \alpha\|x\|^2 + (1-\alpha)\|y\|^2 - \alpha(1-\alpha)\|x-y\|^2 \tag{6}$$

for all $\alpha \in \mathbb{R}$ and $(x, y) \in \mathscr{H} \times \mathscr{H}$.

Definition 1 A mapping $T : \mathscr{H} \to \mathscr{H}$ is called an *averaged mapping* if it can be written as the average of the identity I and a nonexpansive mapping, that is,

$$T = (1-\alpha)I + \alpha S, \tag{7}$$

where α is a number in $]0, 1[$ and $S : \mathscr{H} \to \mathscr{H}$ is a nonexpansive mapping. More precisely, when (7) holds, we say that T is *α-averaged.*

It is obvious that a averaged mapping is nonexpansive.

Lemma 1 ([1]) *Let* $\{\psi_n\}$, $\{\delta_n\}$ *and* $\{\alpha_n\}$ *be the sequences in* $[0, +\infty)$ *such that* $\psi_{n+1} \le \psi_n + \alpha_n(\psi_n - \psi_{n-1}) + \delta_n$ *for each* $n \ge 1$, $\sum_{n=1}^{\infty} \delta_n < +\infty$ *and there exists a real number* α *with* $0 \le \alpha_n \le \alpha < 1$ *for all* $n \in \mathbb{N}$. *Then the following hold:*

(1) $\sum_{n\ge1}[\psi_n - \psi_{n-1}]_+ < +\infty$, *where* $[t]_+ = \max\{t, 0\}$;
(2) *there exists* $\psi^* \in [0, +\infty)$ *such that* $\lim_{n\to+\infty} \psi_n = \psi^*$.

Lemma 2 ([4]) *Let D be a nonempty closed convex subset of $\mathscr{H}$ and $T : D \to \mathscr{H}$ be a nonexpansive mapping. Let $\{x_n\}$ be a sequence in D and $x \in \mathscr{H}$ such that $x_n \rightharpoonup x$ and $Tx_n - x_n \to 0$ as $n \to +\infty$. Then $x \in Fix(T)$.*

Lemma 3 ([4]) *Let C be a nonempty subset of $\mathscr{H}$ and $\{x_n\}$ be a sequence in $\mathscr{H}$ such that the following two conditions hold:*

(i) *for all $x \in C$, $\lim_{n\to\infty} \|x_n - x\|$ exists;*
(ii) *every sequential weak cluster point of $\{x_n\}$ is in C.*

Then the sequence $\{x_n\}$ converges weakly to a point in C.

3 The General Inertial Mann Algorithms

In this section, first, we revisit the accelerated Mann algorithm. Then we propose the general inertial Mann algorithm and show that it includes some other algorithms as special cases.

3.1 Revisit the Accelerated Mann Algorithm

In 2014, Sakurai and Liduka [36] first proposed an acceleration of the Halpern algorithm to search for a fixed point of a nonexpansive mapping. Inspired by their work, by combining the Mann algorithm (1) and conjugate gradient methods [29], the authors [12] proposed the following accelerated Mann algorithm:

$$d_{n+1} := \frac{1}{\gamma}(T(x_n) - x_n) + \beta_n d_n, \tag{8}$$

$$y_n := x_n + \gamma d_{n+1}, \tag{9}$$

$$x_{n+1} := \lambda_n x_n + (1 - \lambda_n) y_n \tag{10}$$

for each $n \geq 1$, where $\gamma > 0$. The sequence $\{x_n\}$ converges weakly to a fixed point of T provided that the sequences $\{\lambda_n\}$ and $\{\beta_n\}$ satisfy the following conditions:

(A1) $\sum_{n=0}^{\infty} \lambda_n(1 - \lambda_n) = \infty$;
(A2) $\sum_{n=0}^{\infty} \beta_n < \infty$.

Moreover, the sequence $\{x_n\}$ satisfies the following condition:

(A3) $\{T(x_n) - x_n\}$ is bounded.

Remark 1 The condition (A3) is very strict. Sakurai and Liduka [36] discussed it on two cases:

(1) Suppose that $Fix(T)$ is bounded. Let C be a bounded closed convex set such that $Fix(T) \subset C$ and P_C can be easily computed (for example, C is a closed ball with a large enough radius). Then compute

$$x_{n+1} := P_C(\lambda_n x_n + (1 - \lambda_n) y_n)$$

for each $n \geq 1$ instead of the x_{n+1} in (10). The boundedness of C and the nonexpansivity of T mean that $\{x_n\}$ and $\{T(x_n)\}$ are bounded. Therefore, the condition (A3) holds.

(2) Suppose that $Fix(T)$ is unbounded. One cannot choose a bounded C satisfying that $Fix(T) \subset C$ and verify the boundedness of $\{T(x_n) - x_n\}$.

Next, we rewrite the accelerated Mann algorithm (8)–(10). Based on the new formula, its convergence will be reanalyzed in Sect. 4.

Substitute (9) into (10), we have

$$\begin{aligned} x_{n+1} &= \lambda_n x_n + (1 - \lambda_n)(x_n + \gamma d_{n+1}) \\ &= x_n + (1 - \lambda_n)\gamma d_{n+1} \end{aligned} \tag{11}$$

for each $n \geq 1$, which implies that

$$d_{n+1} = \frac{1}{(1 - \lambda_n)\gamma}(x_{n+1} - x_n) \tag{12}$$

for each $n \geq 1$. Combining (8) and (9), we have

$$\begin{aligned} y_n &= T(x_n) + \gamma \beta_n d_n \\ &= T(x_n) + \frac{\beta_n}{1 - \lambda_{n-1}}(x_n - x_{n-1}) \end{aligned} \tag{13}$$

for each $n \geq 1$, where the second equality comes from (12). Substitute (13) into (10), we obtain

$$\begin{aligned} x_{n+1} &= \lambda_n x_n + (1 - \lambda_n)\left[T(x_n) + \frac{\beta_n}{1 - \lambda_{n-1}}(x_n - x_{n-1})\right] \\ &= \lambda_n x_n + (1 - \gamma_n)T(x_n) + \frac{\beta_n(1 - \lambda_n)}{1 - \lambda_{n-1}}(x_n - x_{n-1}) \\ &= \lambda_n\left[x_n + \frac{\beta_n(1 - \lambda_n)}{\lambda_n(1 - \lambda_{n-1})}(x_n - x_{n-1})\right] + (1 - \lambda_n)T(x_n) \end{aligned} \tag{14}$$

for each $n \geq 1$. Set

$$\alpha_n = \frac{\beta_n(1 - \lambda_n)}{\lambda_n(1 - \lambda_{n-1})} \tag{15}$$

and

$$y_n = x_n + \alpha_n (x_n - x_{n-1}) \tag{16}$$

for each $n \geq 1$. Then the formula (8)–(10) can be rewrite as:

$$\begin{cases} y_n = x_n + \alpha_n (x_n - x_{n-1}), \\ x_{n+1} = \lambda_n y_n + (1 - \lambda_n) T x_n \end{cases} \tag{17}$$

for each $n \geq 1$.

3.2 *Algorithms*

Now we present the general inertial Mann algorithm as follows:

$$\begin{cases} y_n = x_n + \alpha_n (x_n - x_{n-1}), \\ z_n = x_n + \beta_n (x_n - x_{n-1}), \\ x_{n+1} = (1 - \lambda_n) y_n + \lambda_n T(z_n) \end{cases} \tag{18}$$

for each $n \geq 1$, where $\{\alpha_n\}$, $\{\beta_n\}$ and $\{\lambda_n\}$ satisfy the following conditions:

(D1) $\{\alpha_n\} \subset [0, \alpha]$ and $\{\beta_n\} \subset [0, \beta]$ are nondecreasing with $\alpha_1 = \beta_1 = 0$ and $\alpha, \beta \in [0, 1)$;

(D2) for any $\lambda, \sigma, \delta > 0$,

$$\delta > \frac{\alpha\xi(1 + \xi) + \alpha\sigma}{1 - \alpha^2}, \quad 0 < \lambda \leq \lambda_n \leq \frac{\delta - \alpha[\xi(1 + \xi) + \alpha\delta + \sigma]}{\delta[1 + \xi(1 + \xi) + \alpha\delta + \sigma]}, \tag{19}$$

where $\xi = \max\{\alpha, \beta\}$.

Remark 2 By form, the general inertial Mann algorithm is the most general Mann algorithm with inertial effects we are aware of. It is easy to show that the general inertial Mann algorithm includes other algorithms as special cases. The relations between the algorithm (18) with other work are as follows:

(1) $\alpha_n = \beta_n$, i.e., $y_n = z_n$: this is the classical inertial Mann algorithm [23];
(2) $\beta_n = 0$: this becomes the accelerated Mann algorithm [12];
(3) $\alpha_n = 0$: it becomes the following algorithm

$$\begin{cases} z_n = x_n + \beta_n (x_n - x_{n-1}), \\ x_{n+1} = (1 - \lambda_n) x_n + \lambda_n T(z_n) \end{cases} \tag{20}$$

for each $n \geq 1$, which has not been studied before. Inspired by Malitsky [26] and Mainge [24, 25], we call the algorithm (20) the *reflected Mann algorithm*.

4 Convergence Analysis

In this section, we prove the convergence of the general inertial Mann algorithm and then deduce the convergence of other methods.

Theorem 1 *Suppose that* $T : \mathscr{H} \to \mathscr{H}$ *is nonexpansive with* $Fix(T) \neq \emptyset$*. Assume the conditions* (D1) *and* (D2) *hold. Then the sequence* $\{x_n\}$ *generated by the general inertial Mann algorithm* (18) *converges weakly to a point of* $Fix(T)$*.*

Proof Take arbitrarily $p \in Fix(T)$. From (6), it follows that

$$\begin{aligned}\|x_{n+1}-p\|^2 &= (1-\lambda_n)\|y_n-p\|^2+\lambda_n\|Tz_n-p\|^2-\lambda_n(1-\lambda_n)\|Tz_n-y_n\|^2\\ &\le (1-\lambda_n)\|y_n-p\|^2+\lambda_n\|z_n-p\|^2-\lambda_n(1-\lambda_n)\|Tz_n-y_n\|^2.\end{aligned} \tag{21}$$

Using (6) again, we have

$$\begin{aligned}\|y_n-p\|^2 &= \|(1+\alpha_n)(x_n-p)-\alpha_n(x_{n-1}-p)\|^2\\ &= (1+\alpha_n)\|x_n-p\|^2-\alpha_n\|x_{n-1}-p\|^2+\alpha_n(1+\alpha_n)\|x_n-x_{n-1}\|^2\end{aligned} \tag{22}$$

Similarly, we have

$$\|z_n-p\|^2 = (1+\beta_n)\|x_n-p\|^2-\beta_n\|x_{n-1}-p\|^2+\beta_n(1+\beta_n)\|x_n-x_{n-1}\|^2. \tag{23}$$

Combining (21), (22) and (23), we have

$$\begin{aligned}&\|x_{n+1}-p\|^2-(1+\theta_n)\|x_n-p\|^2+\theta_n\|x_{n-1}-p\|^2\\ &\quad\le -\lambda_n(1-\lambda_n)\|Tz_n-y_n\|^2\\ &\qquad+[(1-\lambda_n)\alpha_n(1+\alpha_n)+\lambda_n\beta_n(1+\beta_n)]\|x_n-x_{n-1}\|^2,\end{aligned} \tag{24}$$

where

$$\theta_n = \alpha_n(1-\lambda_n)+\beta_n\lambda_n.$$

From (D1), (D2) and $\lambda_n \in (0,1)$, it follows that the $\theta_n \subset [0,\xi]$ is nondecreasing with $\theta_1 = 0$. Using (18), we have

$$\begin{aligned}\|Tz_n - y_n\| &= \left\| \frac{1}{\lambda_n}(x_{n+1} - x_n) + \frac{\alpha_n}{\lambda_n}(x_{n-1} - x_n) \right\|^2 \\ &= \frac{1}{{\lambda_n}^2}\|x_{n+1} - x_n\|^2 + \frac{\alpha_n^2}{{\lambda_n}^2}\|x_{n-1} - x_n\|^2 \\ &\quad + 2\frac{\alpha_n}{{\lambda_n}^2}\langle x_{n+1} - x_n, x_{n-1} - x_n\rangle \\ &\geq \frac{1}{{\lambda_n}^2}\|x_{n+1} - x_n\|^2 + \frac{\alpha_n^2}{{\lambda_n}^2}\|x_{n-1} - x_n\|^2 \\ &\quad + \frac{\alpha_n}{{\lambda_n}^2}\Big(- \rho_n\|x_{n+1} - x_n\|^2 - \frac{1}{\rho_n}\|x_{n-1} - x_n\|^2\Big),\end{aligned} \tag{25}$$

where we denote $\rho_n := \frac{1}{\alpha_n + \delta\lambda_n}$. From (24) and (25), we can derive the inequality

$$\begin{aligned}&\|x_{n+1} - p\|^2 - (1 + \theta_n)\|x_n - p\|^2 + \theta_n\|x_{n-1} - p\|^2 \\ &\leq \frac{(1 - \lambda_n)(\alpha_n\rho_n - 1)}{\lambda_n}\|x_{n+1} - x_n\|^2 + \mu_n\|x_n - x_{n-1}\|^2,\end{aligned} \tag{26}$$

where

$$\mu_n = (1 - \lambda_n)\alpha_n(1 + \alpha_n) + \lambda_n\beta_n(1 + \beta_n) + \alpha_n(1 - \lambda_n)\frac{1 - \rho_n\alpha_n}{\rho_n\lambda_n} \geq 0 \tag{27}$$

since $\rho_n\alpha_n \leq 1$ and $\lambda_n \in (0, 1)$. Again, taking into account the choice of ρ_n, we have

$$\delta = \frac{1 - \rho_n\alpha_n}{\rho_n\lambda_n},$$

and, from (27),

$$\mu_n = (1 - \lambda_n)\alpha_n(1 + \alpha_n) + \lambda_n\beta_n(1 + \beta_n) + \alpha_n(1 - \lambda_n)\delta \leq \xi(1 + \xi) + \alpha\delta \tag{28}$$

for each $n \geq 1$. In the following, we apply some techniques from [2, 7] adapted to our setting. Define the sequences $\phi_n := \|x_n - p\|^2$ for all $n \in \mathbb{N}$ and $\Psi_n := \phi_n - \theta_n\phi_{n-1} + \mu_n\|x_n - x_{n-1}\|^2$ for all $n \geq 1$. Using the monotonicity of $\{\theta_n\}$ and the fact that $\phi_n \geq 0$ for all $n \in \mathbb{N}$, we have

$$\Psi_{n+1} - \Psi_n \leq \phi_{n+1} - (1 + \theta_n)\phi_n + \theta_n\phi_{n-1} + \mu_{n+1}\|x_{n+1} - x_n\|^2 - \mu_n\|x_n - x_{n-1}\|^2.$$

By (26), we know

$$\Psi_{n+1} - \Psi_n \leq \Big(\frac{(1 - \lambda_n)(\alpha_n\rho_n - 1)}{\lambda_n} + \mu_{n+1}\Big)\|x_{n+1} - x_n\|^2. \tag{29}$$

Now, we claim that

$$\frac{(1-\lambda_n)(\alpha_n\rho_n-1)}{\lambda_n}+\mu_{n+1}\leq-\sigma \tag{30}$$

for each $n\geq 1$. Indeed, by (27) and the monotonicity of $\{\lambda_n\}$, we have

$$\begin{aligned}
&\frac{(1-\lambda_n)(\alpha_n\rho_n-1)}{\lambda_n}+\mu_{n+1}\leq-\sigma\\
\Longleftrightarrow\ &\lambda_n(\mu_{n+1}+\sigma)+(1-\lambda_n)(\alpha_n\rho_n-1)\leq 0\\
\Longleftrightarrow\ &\lambda_n(\mu_{n+1}+\sigma)-\frac{\delta\lambda_n(1-\lambda_n)}{\alpha_n+\delta\lambda_n}\leq 0\\
\Longleftrightarrow\ &(\alpha_n+\delta\lambda_n)(\mu_{n+1}+\sigma)+\delta\lambda_n\leq\delta.
\end{aligned}$$

Employing (28), we have

$$(\alpha_n+\delta\lambda_n)(\mu_{n+1}+\sigma)+\delta\lambda_n\leq(\alpha+\delta\lambda_n)[\xi(1+\xi)+\alpha\delta+\sigma]+\delta\lambda_n\leq\delta,$$

where the last inequality follows by using the upper bound for (λ_n) in (19). Hence the claim in (30) is true. It follows from (29) and (30) that

$$\Psi_{n+1}-\Psi_n\leq-\sigma\|x_{n+1}-x_n\|^2 \tag{31}$$

for each $n\geq 1$. The sequence $(\Psi_n)_{n\geq 1}$ is non-increasing and the boundness for $(\theta_n)_{n\geq 1}$ delivers

$$-\xi\phi_{n-1}\leq\phi_n-\xi\phi_{n-1}\leq\Psi_n\leq\Psi_1 \tag{32}$$

for each $n\geq 1$. Thus we obtain

$$\phi_n\leq\xi^n\phi_0+\Psi_1\sum_{k=1}^{n-1}\xi^k\leq\xi^n\phi_0+\frac{\Psi_1}{1-\xi} \tag{33}$$

for each $n\geq 1$, where we notice that $\Psi_1=\phi_1\geq 0$ (due to the relation $\theta_1=\alpha_1=\beta_1=0$). Using (31)–(33), for all $n\geq 1$, we have

$$\sigma\sum_{k=1}^{n}\|x_{k+1}-x_k\|^2\leq\Psi_1-\Psi_{n+1}\leq\Psi_1+\xi\phi_n\leq\xi^{n+1}\phi_0+\frac{\Psi_1}{1-\xi},$$

which means that

$$\sum_{n=1}^{\infty}\|x_{n+1}-x_n\|^2<+\infty. \tag{34}$$

Thus we have

$$\lim_{n\to\infty} \|x_{n+1} - x_n\| = 0. \tag{35}$$

From (20), we have

$$\begin{aligned}\|y_n - x_{n+1}\| &\le \|x_n - x_{n+1}\| + \alpha_n \|x_n - x_{n-1}\| \\ &\le \|x_n - x_{n+1}\| + \alpha \|x_n - x_{n-1}\|,\end{aligned}$$

which with (35) implies that

$$\lim_{n\to\infty} \|y_n - x_{n+1}\| = 0. \tag{36}$$

Similarly, we obtain

$$\lim_{n\to\infty} \|z_n - x_{n+1}\| = 0. \tag{37}$$

For an arbitrary $p \in Fix(T)$, by (26), (28), (34) and Lemma 1, we derive that $\lim_{n\to\infty} \|x_n - p\|$ exists (we take into consideration also $\lambda_n \in (0, 1)$ in (26)). On the other hand, let x be a sequential weak cluster point of $\{x_n\}$, that is, there exists a subsequence $\{x_{n_k}\}$ which converge weakly to x. By (37), it follows that $z_{n_k} \rightharpoonup x$ as $k \to \infty$. Furthermore, from (18), we have

$$\begin{aligned}\|Tz_n - z_n\| &\le \|Tz_n - y_n\| + \|y_n - z_n\| \\ &\le \frac{1}{\lambda_n}\|x_{n+1} - y_n\| + \|y_n - x_{n+1}\| + \|z_n - x_{n+1}\| \\ &\le \left(1 + \frac{1}{\lambda}\right)\|x_{n+1} - y_n\| + \|z_n - x_{n+1}\|.\end{aligned}$$

Thus, by (36) and (37), we obtain $\|Tz_{n_k} - z_{n_k}\| \to 0$ as $k \to \infty$. Applying now Lemma 2 for the sequence $\{z_{n_k}\}$, we conclude that $x \in Fix(T)$. From Lemma 3, it follows that $\{x_n\}$ converges weakly to a point in $Fix(T)$. This completes the proof. □

Let $\alpha_n = \beta_n$ and then Theorem 1 becomes Theorem 5 in [7].

Theorem 2 *Suppose that* $T : \mathscr{H} \to \mathscr{H}$ *is a nonexpansive mapping with* $Fix(T) \neq \emptyset$. *Assume the conditions* (C1) *and* (C2) *hold. Then the sequence* $\{x_n\}$ *generated by the classical Mann algorithm* (5) *converges weakly to a point of* $Fix(T)$.

Let $\beta_n = 0$ and then we obtain another convergence condition of the accelerated Mann algorithm.

Theorem 3 *Suppose that $T : \mathscr{H} \to \mathscr{H}$ is a nonexpansive mapping with $Fix(T) \neq \emptyset$. Assume that $\{\alpha_n\} \subset [0, \alpha]$ is nondecreasing with $\alpha_1 = 0$ and $0 \leq \alpha < 1$ and $\{\lambda_n\}$ satisfies*

$$\delta > \frac{\alpha^2(1+\alpha)+\alpha\sigma}{1-\alpha^2}, \quad 0 < \lambda \leq \lambda_n \leq \frac{\delta - \alpha[\alpha(1+\alpha)+\alpha\delta+\sigma]}{\delta[1+\alpha(1+\alpha)+\alpha\delta+\sigma]},$$

where $\lambda, \sigma, \delta > 0$. Then the sequence $\{x_n\}$ generated by the accelerated Mann algorithm (17) converges weakly to a point of $Fix(T)$.

Remark 3 It is obvious that Theorem 3 does not need the strict condition (A3).

Let $\alpha_n = 0$ and then we obtain the convergence theorem of the reflected Mann algorithm.

Theorem 4 *Suppose that $T : \mathscr{H} \to \mathscr{H}$ is a nonexpansive mapping with $Fix(T) \neq \emptyset$. Assume that $\{\beta_n\} \subset [0, \beta]$ is nondecreasing with $\beta_1 = 0$ and $0 \leq \beta < 1$ and $\{\lambda_n\}$ satisfies*

$$0 < \lambda \leq \lambda_n \leq \frac{1}{1+\beta(1+\beta)+\sigma},$$

where $\lambda, \sigma > 0$. Then the sequence $\{x_n\}$ generated by the reflected Mann algorithm (20) converges weakly to a point of $Fix(T)$.

5 Applications

Consider the following *constrained convex minimization problem*:

$$\min_{x \in C} \varphi(x), \tag{38}$$

where C is a closed convex subset of a Hilbert space $\mathscr{H}$ and $\varphi : C \to \mathbb{R}$ is a real-valued convex function. If $\varphi(x)$ is differentiable, then the problem (38) is equivalent to the following fixed point problem:

$$x = P_C(x - \gamma \nabla \varphi(x)), \tag{39}$$

where $\gamma > 0$. Then the gradient-projection algorithm generates a iterative sequence via

$$x_{n+1} = P_C(x_n - \gamma \nabla \varphi(x_n)) \tag{40}$$

for each $n \geq 1$, where the initial guess x_0 is taken from C arbitrarily, the parameter γ is a positive real number and P_C is the metric projection from $\mathscr{H}$ onto C.

Remark 4 There are some inertial type algorithms for solving the minimization problems (38). We first review them as follows:

(1) In 2015, Bot and Csetnek [6] proposed the so-called *inertial hybrid proximal extragradient algorithm*, which includes the following algorithm as a special case:

$$\begin{cases} y^k = x^k + \alpha_k(x^k - x^{k-1}), \\ x^{k+1} = P_C(y^k - \lambda_k \nabla\varphi(x^k)) \end{cases}$$

for each $k \geq 1$. They showed the convergence of the algorithm provided that $\nabla\varphi$ is γ-cocoercive and $\{\alpha_k\}$ is nondecreasing with $\alpha_1 = 0$, $0 \leq \alpha_k \leq \alpha$ and $0 < \underline{\lambda} \leq \lambda_k \leq 2\gamma\sigma^2$ for any $\alpha, \sigma \geq 0$ such that $\alpha(5 + 4\sigma^2) + \sigma^2 < 1$.

(2) In 2015, Malitski [26] proposed the *projected reflected method*:

$$x^{k+1} = P_C\left(x^k - \lambda\nabla\varphi(2x^k - x_{k-1})\right)$$

for each $k \geq 1$. In 2016, Mainge [24, 25] extended the above method to more general cases as follows:

$$\begin{cases} y^k = x^k + \alpha_k(x^k - x^{k-1}), \\ x^{k+1} = P_C\left(x^k - \lambda_n \nabla\varphi(y^k)\right) \end{cases}$$

for each $k \geq 1$, where $\alpha_k \geq 0$ and $\{\lambda_n\} \subset [0, 1]$ satisfies some conditions. They proved the convergence of the method when $\nabla\varphi$ is Lipshitz continuous and monotone.

(3) In 2016, Dong et al. [14] introduced the *extragradient method* with inertial effects:

$$\begin{cases} w_k = x_k + \alpha_k(x_k - x_{k-1}), \\ y_k = P_C(w_k - \tau\nabla\varphi(w_k)), \\ x_{k+1} = (1 - \lambda_k)w_k + \lambda_k P_C(w_k - \tau\nabla\varphi(y_k)) \end{cases} \tag{41}$$

for each $k \geq 1$. The numerical experiments show that the inertial algorithm (41) speeds up the extragradient method.

Assume that $\nabla\varphi$ is L-Lipschitz continuous, namely, there is a constant $L > 0$ such that

$$\|\nabla\varphi(x) - \nabla\varphi(y)\| \leq L\|x - y\| \tag{42}$$

for all $x, y \in C$. In 2011, Xu [37] showed that the composite $P_C(I - \gamma\nabla\varphi)$ is $((2 + \gamma L)/4)$-averaged for $0 < \gamma < 2/L$. So the composite $P_C(I - \gamma\nabla\varphi)$ is nonexpansive and we use the general inertial Mann methods (18) to construct the *general inertial gradient-projection algorithm* for (38) as follows:

$$\begin{cases} y_n = x_n + \alpha_n(x_n - x_{n-1}), \\ z_n = x_n + \beta_n(x_n - x_{n-1}), \\ x_{n+1} = (1-\lambda_n)y_n + \lambda_n P_C(z_n - \gamma\nabla\varphi(z_n)) \end{cases} \tag{43}$$

for each $n \geq 1$, where $\{\alpha_n\}$, $\{\beta_n\}$ and $\{\lambda_n\}$ satisfy the conditions (D1) and (D2).

To generalize Theorem 1, we have the following convergent result:

Theorem 5 *Assume that the minimization problem* (38) *is consistent and the gradient* $\nabla\varphi$ *satisfies the Lipschitz condition* (42). *Let* γ *be a number such that* $0 < \gamma < 2/L$. *Then the sequence* $\{x_n\}$ *generated by the general inertial gradient-projection algorithm* (43) *converges weakly to a minimizer of the problem* (38).

6 Numerical Examples and Conclusions

In this section, we present a numerical example to illustrate the choice of the parameters $\{\alpha_n\}$ and $\{\beta_n\}$ in the general inertial algorithm (18). All the programs are written in Matlab version 7.0. and performed on a PC Desktop Intel(R) Core(TM) i5-4200U CPU @ 1.60 GHz 2.30 GHz, RAM 4.00 GB.

Problem 2 (see [36]) For any nonempty closed convex set $C_i \subset \mathbb{R}^N$ for each $i = 0, 1, \cdots, m$,

$$\text{Find } x^* \in C := \bigcap_{i=0}^{m} C_i,$$

where one assumes that $C \neq \emptyset$.

Define a mapping $T : \mathbb{R}^N \to \mathbb{R}^N$ by

$$T := P_0\Big(\frac{1}{m}\sum_{i=1}^{m} P_i\Big), \tag{44}$$

where $P_i = P_{C_i}$ $(i = 0, 1, \cdots, m)$ stands for the metric projection onto C_i. Since P_i $(i = 0, 1, \cdots, m)$ is nonexpansive, the mapping T defined by (44) is also nonexpansive. Moreover, we find that

$$Fix(T) = Fix(P_0)\bigcap_{i=1}^{m} Fix(P_i) = C_0 \bigcap_{i=1}^{m} C_i = C.$$

Table 1 The general inertial Mann algorithm with $\alpha_n = 0.4$, $\lambda_n = 0.5$, $N = 500$, $m = 300$

The initial value	β_n	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$100 \times rand(N, 1)$	Iter.	8	18	20	1367	1291	1195	796	32	36	41	47
$(1, 1, \cdots, 1)$	Iter.	4	13	15	17	1111	1031	928	27	31	36	42
$(1, -1, \cdots, 1, -1)$	Iter.	4	13	15	17	1429	1333	1217	27	31	36	42

Table 2 The general inertial Mann algorithm with $\beta_n = 0.0$, $\lambda_n = 0.5$, $N = 500$, $m = 700$

The initial value	α_n	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$100 \times rand(N, 1)$	Iter.	2181	4093	3786	4	8	2588	12	11	15	15	19
$(1, 1, \cdots, 1)$	Iter.	4798	4557	4315	3585	4	3	3	8	8	7	7
$(10, -10, \cdots, 10, -10)$	Iter.	3608	3427	3249	5	3213	8	7	12	11	11	15

In the experiment, we set C_i $(i = 0, 1, \cdots, m)$ as a closed ball with center $c_i \in \mathbb{R}^N$ and radius $r_i > 0$. Thus P_i $(i = 0, 1, \cdots, m)$ can be computed with

$$P_i(x) := \begin{cases} c_i + \dfrac{r_i}{\|c_i - x\|}(x - c_i) & \text{if} \quad \|c_i - x\| > r_i, \\ x & \text{if} \quad \|c_i - x\| \le r_i. \end{cases}$$

Choose $r_i := 1\,(i = 0, 1, \cdots, m)$, $c_0 := 0$, $c_i \in (-1/\sqrt{N}, 1/\sqrt{N})^N$ $(i = 1, \cdots, m)$ are randomly chosen.

In the numerical results listed in the tables, "Iter." denotes the number of iterations. We take $E(x) = \|x_n - x_{n-1}\| < 10^{-6}$ as the stopping criterion and test three initial values x_0.

In the general inertial Mann algorithm, there are three parameters α_n, β_n, λ_n. To compare the different algorithms, we choose $\lambda_n = 0.5$ and test different choices of α_n and β_n.

Table 1 illustrates that the number of iterations for the general inertial Mann algorithm with $\beta_n = 0$ is minimal, that is, the accelerated Mann algorithm is best.

From Table 2, we conclude that the number of the iteration is small for the accelerated algorithm with $\alpha_n \in [0.6, 1.0]$.

Acknowledgements The first author would like to thank Professor Chunlin Wu for the helpful discussion and valuable suggestions. We wish to express our thanks to Professor Mihai Turinici for reading the paper and providing very helpful comments.

References

1. F. Alvarez, Weak convergence of a relaxed and inertial hybrid projection-proximal point algorithm for maximal monotone operators in Hilbert space. SIAM J. Optim. **14**, 773–782 (2004)

2. F. Alvarez, H. Attouch, An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping. Set-Valued Anal. **9**, 3–11 (2001)
3. H.H. Bauschke, J.M. Borwein, On projection algorithms for solving convex feasibility problems. SIAM Rev. **38**, 367–426 (1996)
4. H.H. Bauschke, P.L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces* (Springer, Berlin, 2011)
5. A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
6. R.I. Bot, E.R. Csetnek, A hybrid proximal-extragradient algorithm with inertial effects. Numer. Funct. Anal. Optim. **36**, 951–963 (2015)
7. R.I. Bot, E.R. Csetnek, C. Hendrich, Inertial Douglas-Rachford splitting for monotone inclusion problems. Appl. Math. Comput. **256**, 472–487 (2015)
8. A. Chambolle, C. Dossal, On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". J. Optim. Theory Appl. **166**, 968–982 (2015)
9. P. Chen, J. Huang, X. Zhang, A primal-dual fixed point algorithm for convex separable minimization with applications to image restoration. Inverse Prob. **29**, 025011 (33 pp.) (2013)
10. P. Chen, J. Huang, X. Zhang, A primal-dual fixed point algorithm for minimization of the sum of three convex separable functions. Fixed Point Theory Appl. **2016**, 54 (2016)
11. Q.L. Dong, Y.Y. Lu, A new hybrid algorithm for a nonexpansive mapping. Fixed Point Theory Appl. **2015**, 37 (2015)
12. Q.L. Dong, H.Y. Yuan, Accelerated Mann and CQ algorithms for finding a fixed point of a nonexpansive mapping. Fixed Point Theory Appl. **2015**, 125 (2015)
13. Q.L. Dong, S. He, Y.J. Cho, A new hybrid algorithm and its numerical realization for two nonexpansive mappings. Fixed Point Theory Appl. **2015**, 150 (2015)
14. Q.L. Dong, Y.Y. Lu, J. Yang, The extragradient algorithm with inertial effects for solving the variational inequality. Optimization **65**(12), 2217–2226 (2016)
15. B. Halpern, Fixed points of nonexpanding maps. Bull. Am. Math. Soc. **73**, 957–961 (1967)
16. S. He, C. Yang, Boundary point algorithms for minimum norm fixed points of nonexpansive mappings. Fixed Point Theory Appl. **2014**, 56 (2014)
17. H. Iiduka, Iterative algorithm for triple-hierarchical constrained nonconvex optimization problem and its application to network bandwidth allocation. SIAM J. Optim. **22**, 862–878 (2012)
18. H. Iiduka, Fixed point optimization algorithms for distributed optimization in networked systems. SIAM J. Optim. **23**, 1–26 (2013)
19. S. Ishikawa, Fixed points by a new iteration method. Proc. Am. Math. Soc. **44**, 147–150 (1974)
20. M.A. Krasnoselskii, Two remarks on the method of successive approximations. Usp. Mat. Nauk **10**, 123–127 (1955)
21. D.A. Lorenz, T. Pock, An inertial forward-backward algorithm for monotone inclusions. J. Math. Imaging Vis. **51**, 311–325 (2015)
22. W.R. Mann, Mean value methods in iteration. Proc. Am. Math. Soc. **4**, 506–510 (1953)
23. P.E. Mainge, Convergence theorems for inertial KM-type algorithms. J. Comput. Appl. Math. **219**, 223–236 (2008)
24. P.E. Mainge, Numerical approach to monotone variational inequalities by a one-step projected reflected gradient method with line-search procedure. Comput. Math. Appl. **72**, 720–728 (2016)
25. P.E. Mainge, M.L. Gobinddass, Convergence of one-step projected gradient methods for variational inequalities. J. Optim. Theory Appl. **171**, 146–168 (2016)
26. Y. Malitski, Projected reflected gradient method for variational inequalities. SIAM J. Optim. **25**, 502–520 (2015)
27. C.A. Micchelli, L. Shen, Y. Xu, Proximity algorithms for image models: denoising. Inverse Prob. **27**, 45009–45038 (2011)
28. A. Moudafi, M. Oliny, Convergence of a splitting inertial proximal method formonotone operators. J. Comput. Appl. Math. **155**, 447–454 (2003)

29. J. Nocedal, S.J. Wright, *Numerical Optimization*, 2nd edn. Springer Series in Operations Research and Financial Engineering (Springer, Berlin, 2006)
30. P.M. Pardalos, T.M. Rassias, *Contributions in Mathematics and Engineering; In Honor of Constantin Caratheodory* (Springer, Berlin, 2016)
31. P.M. Pardalos, P.G. Georgiev, H.M. Srivastava, *Nonlinear Analysis, Stability, Approximation, and Inequalities; In Honor of Themistocles M. Rassias on the Occasion of his 60th Birthday* (Springer, Berlin, 2012)
32. E. Picard, Memoire sur la theorie des equations aux derivees partielles et la methode des approximations successives. J. Math. Pures et Appl. **6**, 145–210 (1890)
33. B.T. Polyak, Some methods of speeding up the convergence of iteration methods. U.S.S.R. Comput. Math. Math. Phys. **4**, 1–17 (1964)
34. B.T. Polyak, *Introduction to Optimization* (Optimization Software Inc., Publications Division, New York, 1987)
35. T.M. Rassias, L. Toth, *Topics in Mathematical Analysis and Applications* (Springer, Berlin, 2014)
36. K. Sakurai, H. Liduka, Acceleration of the Halpern algorithm to search for a fixed point of a nonexpansive mapping. Fixed Point Theory Appl. **2014**, 202 (2014)
37. H.K. Xu, Averaged mappings and the gradient-projection algorithm. J. Optim. Theory Appl. **150**, 360–378 (2011)
38. C. Yang, S. He, General alternative regularization methods for nonexpansive mappings in Hilbert spaces. Fixed Point Theroy Appl. **2014**, 203 (2014)

Reverses of Jensen's Integral Inequality and Applications: A Survey of Recent Results

Silvestru Sever Dragomir

1 Introduction

Let $(\Omega, \mathcal{A}, \mu)$ be a measurable space consisting of a set Ω, a σ-algebra $\mathcal{A}$ of parts of Ω and a countably additive and positive measure μ on $\mathcal{A}$ with values in $\mathbb{R} \cup \{\infty\}$.

For a μ-measurable function $w : \Omega \to \mathbb{R}$, with $w(x) \geq 0$ for μ-a.e. (almost every) $x \in \Omega$, consider the Lebesgue space $L_w(\Omega, \mu) := \{f : \Omega \to \mathbb{R},\ f$ is μ-measurable and $\int_\Omega w(x) |f(x)| \, d\mu(x) < \infty\}$. For simplicity of notation we write everywhere in the sequel $\int_\Omega w d\mu$ instead of $\int_\Omega w(x) \, d\mu(x)$. We also assume that $\int_\Omega w d\mu = 1$.

An useful result that is used to provide simpler upper bounds for the difference in Jensen's inequality is the Gruss' inequality. We recall now some facts related to this famous result.

If $f, g : \Omega \to \mathbb{R}$ are μ-measurable functions and $f, g, fg \in L_w(\Omega, \mu)$, then we may consider the *Čebyšev functional*

$$T_w(f, g) := \int_\Omega wfg d\mu - \int_\Omega wf d\mu \int_\Omega wg d\mu. \tag{1.1}$$

The following result is known in the literature as the *Grüss inequality*

$$|T_w(f, g)| \leq \frac{1}{4} (\Gamma - \gamma)(\Delta - \delta), \tag{1.2}$$

S. S. Dragomir (✉)
Mathematics, College of Engineering & Science, Victoria University, Melbourne City, VIC, Australia

DST-NRF Centre of Excellence in the Mathematical and Statistical Sciences, School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa
e-mail: sever.dragomir@vu.edu.au

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_8

provided

$$-\infty < \gamma \leq f(x) \leq \Gamma < \infty, \quad -\infty < \delta \leq g(x) \leq \Delta < \infty \tag{1.3}$$

for μ-a.e. a. $x \in \Omega$. The constant $\frac{1}{4}$ is sharp in the sense that it cannot be replaced by a smaller quantity.

Note that if $\Omega = \{1, \ldots, n\}$ and μ is the discrete measure on Ω, then we obtain the discrete Grüss inequality

$$\left| \sum_{i=1}^{n} w_i x_i y_i - \sum_{i=1}^{n} w_i x_i \cdot \sum_{i=1}^{n} w_i y_i \right| \leq \frac{1}{4} (\Gamma - \gamma)(\Delta - \delta), \tag{1.4}$$

provided $\gamma \leq x_i \leq \Gamma$, $\delta \leq y_i \leq \Delta$ for each $i \in \{1, \ldots, n\}$ and $w_i \geq 0$ with $W_n := \sum_{i=1}^{n} w_i = 1$.

With the above assumptions, if $f \in L_w(\Omega, \mu)$ then we may define

$$D_w(f) := D_{w,1}(f) := \int_\Omega w \left| f - \int_\Omega w f d\mu \right| d\mu. \tag{1.5}$$

In 2002, Cerone and Dragomir [5] obtained the following refinement of the Grüss inequality (1.2):

Theorem 1 (Cerone and Dragomir [5]) *Let w, f, $g : \Omega \to \mathbb{R}$ be μ-measurable functions with $w \geq 0$ μ-a.e. (almost everywhere) on Ω and $\int_\Omega w d\mu = 1$. If f, g, $fg \in L_w(\Omega, \mu)$ and there exists the constants δ, Δ such that*

$$-\infty < \delta \leq g(x) \leq \Delta < \infty \ \text{ for } \ \mu\text{-a.e. } x \in \Omega, \tag{1.6}$$

then we have the inequality

$$|T_w(f, g)| \leq \frac{1}{2} (\Delta - \delta) D_w(f). \tag{1.7}$$

The constant $\frac{1}{2}$ is sharp in the sense that it cannot be replaced by a smaller quantity.

Remark 1 The inequality (1.7) was obtained for the particular case $\Omega = [a, b]$ and the uniform weight $w(t) = 1$, $t \in [a, b]$ by Cheng and Sun in [7]. However, in that paper the authors did not prove the sharpness of the constant $\frac{1}{2}$.

For $f \in L_{p,w}(\Omega, \mathcal{A}, \mu) := \left\{ f : \Omega \to \mathbb{R}, \int_\Omega w |f|^p d\mu < \infty \right\}$, $p \geq 1$ we may also define

$$D_{w,p}(f) := \left[\int_\Omega w \left| f - \int_\Omega w f d\mu \right|^p d\mu \right]^{\frac{1}{p}} = \left\| f - \int_\Omega w f d\mu \right\|_{\Omega,p} \tag{1.8}$$

where $\|\cdot\|_{\Omega,p}$ is the usual p-norm on $L_{p,w}(\Omega, \mathcal{A}, \mu)$, namely,

$$\|h\|_{\Omega,p} := \left(\int_{\Omega} w\,|h|^p\,d\mu\right)^{\frac{1}{p}}, \quad p \geq 1.$$

Using Hölder's inequality we get

$$D_{w,1}(f) \leq D_{w,p}(f) \quad \text{for } p \geq 1,\ f \in L_{p,w}(\Omega, \mathcal{A}, \mu); \tag{1.9}$$

and, in particular for $p = 2$

$$D_{w,1}(f) \leq D_{w,2}(f) := \left[\int_{\Omega} wf^2 d\mu - \left(\int_{\Omega} wf d\mu\right)^2\right]^{\frac{1}{2}}, \tag{1.10}$$

if $f \in L_{2,w}(\Omega, \mathcal{A}, \mu)$.

For $f \in L_{\infty}(\Omega, \mathcal{A}, \mu) := \left\{f : \Omega \to \mathbb{R},\ \|f\|_{\Omega,\infty} := \operatorname{essup}_{x\in\Omega}|f(x)| < \infty\right\}$ we also have

$$D_{w,p}(f) \leq D_{w,\infty}(f) := \left\|f - \int_{\Omega} wf d\mu\right\|_{\Omega,\infty}. \tag{1.11}$$

The following corollary may be useful in practice.

Corollary 1 *With the assumptions of Theorem 1, we have*

$$\begin{aligned} |T_w(f,g)| &\leq \frac{1}{2}(\Delta - \delta) D_w(f) \qquad (1.12)\\ &\leq \frac{1}{2}(\Delta - \delta) D_{w,p}(f) \quad \textit{if } f \in L_p(\Omega, \mathcal{A}, \mu),\ 1 < p < \infty;\\ &\leq \frac{1}{2}(\Delta - \delta) D_{w,\infty}(f) \quad \textit{if } f \in L_{\infty}(\Omega, \mathcal{A}, \mu). \end{aligned}$$

Remark 2 The inequalities in (1.12) are in order of increasing coarseness. If we assume that $-\infty < \gamma \leq f(x) \leq \Gamma < \infty$ for μ-a.e. $x \in \Omega$, then by the Grüss inequality for $g = f$ we have for $p = 2$

$$\left[\int_{\Omega} wf^2 d\mu - \left(\int_{\Omega} wf d\mu\right)^2\right]^{\frac{1}{2}} \leq \frac{1}{2}(\Gamma - \gamma). \tag{1.13}$$

By (1.12), we deduce the following sequence of inequalities

$$|T_w(f,g)| \leq \frac{1}{2}(\Delta - \delta)\int_{\Omega} w\left|f - \int_{\Omega} wf d\mu\right| d\mu \tag{1.14}$$

$$\leq \frac{1}{2}(\Delta - \delta)\left[\int_{\Omega} wf^2 d\mu - \left(\int_{\Omega} wf d\mu\right)^2\right]^{\frac{1}{2}}$$
$$\leq \frac{1}{4}(\Delta - \delta)(\Gamma - \gamma)$$

for $f, g : \Omega \to \mathbb{R}$, μ-measurable functions and so that $-\infty < \gamma \leq f(x) < \Gamma < \infty$, $-\infty < \delta \leq g(x) \leq \Delta < \infty$ for μ-a.e. $x \in \Omega$. Thus, the inequality (1.14) is a refinement of Grüss' inequality (1.2).

In order to provide a reverse of the celebrated Jensen's integral inequality for convex functions, Dragomir obtained in 2002 [14] the following result:

Theorem 2 (Dragomir [14]) *Let $\Phi : [m, M] \subset \mathbb{R} \to \mathbb{R}$ be a differentiable convex function on (m, M) and $f : \Omega \to [m, M]$ so that $\Phi \circ f$, f, $\Phi' \circ f$, $\left(\Phi' \circ f\right) f \in L_w(\Omega, \mu)$, where $w \geq 0$ μ-a.e. on Ω with $\int_{\Omega} w d\mu = 1$. Then we have the inequality:*

$$0 \leq \int_{\Omega} w(\Phi \circ f) d\mu - \Phi\left(\int_{\Omega} wf d\mu\right) \tag{1.15}$$
$$\leq \int_{\Omega} w\left(\Phi' \circ f\right) f d\mu - \int_{\Omega} w\left(\Phi' \circ f\right) d\mu \int_{\Omega} wf d\mu$$
$$\leq \frac{1}{2}\left[\Phi'(M) - \Phi'(m)\right] \int_{\Omega} w \left| f - \int_{\Omega} wf d\mu \right| d\mu.$$

For a generalization of the first inequality when differentiability is not assumed and the derivative Φ' is replaced with a selection φ from the subdifferential $\partial\Phi$, see the paper [41] by Niculescu.

Remark 3 If $\mu(\Omega) < \infty$ and $\Phi \circ f$, f, $\Phi' \circ f$, $\left(\Phi' \circ f\right) f \in L(\Omega, \mu)$, then we have the inequality:

$$0 \leq \frac{1}{\mu(\Omega)} \int_{\Omega} (\Phi \circ f) d\mu - \Phi\left(\frac{1}{\mu(\Omega)} \int_{\Omega} f d\mu\right) \tag{1.16}$$
$$\leq \frac{1}{\mu(\Omega)} \int_{\Omega} \left(\Phi' \circ f\right) f d\mu - \frac{1}{\mu(\Omega)} \int_{\Omega} \left(\Phi' \circ f\right) d\mu \frac{1}{\mu(\Omega)} \int_{\Omega} f d\mu$$
$$\leq \frac{1}{2}\left[\Phi'(M) - \Phi'(m)\right] \frac{1}{\mu(\Omega)} \int_{\Omega} \left| f - \frac{1}{\mu(\Omega)} \int_{\Omega} f d\mu \right| d\mu.$$

Remark 4 On making use of (1.15) and (1.14), one can state the following string of reverse inequalities for the Jensen's difference

$$0 \leq \int_{\Omega} w(\Phi \circ f) d\mu - \Phi\left(\int_{\Omega} wf d\mu\right) \tag{1.17}$$

$$
\begin{aligned}
&\le \int_\Omega w\left(\Phi'\circ f\right) f d\mu - \int_\Omega w\left(\Phi'\circ f\right) d\mu \int_\Omega wf d\mu \\
&\le \frac{1}{2}\left[\Phi'(M) - \Phi'(m)\right] \int_\Omega w \left| f - \int_\Omega wf d\mu \right| d\mu \\
&\le \frac{1}{2}\left[\Phi'(M) - \Phi'(m)\right] \left[\int_\Omega wf^2 d\mu - \left(\int_\Omega wf d\mu\right)^2\right]^{\frac{1}{2}} \\
&\le \frac{1}{4}\left[\Phi'(M) - \Phi'(m)\right](M-m).
\end{aligned}
$$

We notice that the inequality between the first, second and last term from (1.17) was proved in the general case of positive linear functionals in 2001 by Dragomir in [13].

The discrete case is as follows. Let $\bar{\mathbf{a}} = (a_1, \ldots, a_n)$, $\bar{\mathbf{b}} = (b_1, \ldots, b_n)$, $\bar{\mathbf{p}} = (p_1, \ldots, p_n)$ be n-tuples of real numbers with $p_i \ge 0$ $(i \in \{1, \ldots, n\})$ and $\sum_{i=1}^n p_i = 1$. If $b \le b_i \le B,\ i \in \{1, \ldots, n\}$, then one has the inequality

$$
\begin{aligned}
\left|\sum_{i=1}^n p_i a_i b_i - \sum_{i=1}^n p_i a_i \sum_{i=1}^n p_i b_i\right| &\le \frac{1}{2}(B-b)\sum_{i=1}^n p_i \left|a_i - \sum_{j=1}^n p_j a_j\right| \qquad (1.18) \\
&\le \frac{1}{2}(B-b)\left[\sum_{i=1}^n p_i \left|a_i - \sum_{j=1}^n p_j a_j\right|^p\right]^{\frac{1}{p}} \\
&\le \frac{1}{2}(B-b)\max_{i=\overline{1,n}}\left|a_i - \sum_{j=1}^n p_j a_j\right|,
\end{aligned}
$$

where $1 < p < \infty$. The constant $\frac{1}{2}$ is sharp in the first inequality.

If more information about the vector $\bar{\mathbf{a}} = (a_1, \ldots, a_n)$ is available, namely, if there exists the constants a and A such that $a \le a_i \le A,\ i \in \{1, \ldots, n\}$, then

$$
\begin{aligned}
\left|\sum_{i=1}^n p_i a_i b_i - \sum_{i=1}^n p_i a_i \sum_{i=1}^n p_i b_i\right| &\le \frac{1}{2}(B-b)\sum_{i=1}^n p_i \left|a_i - \sum_{j=1}^n p_j a_j\right| \qquad (1.19) \\
&\le \frac{1}{2}(B-b)\left[\sum_{i=1}^n p_i \left|a_i - \sum_{j=1}^n p_j a_j\right|^2\right]^{\frac{1}{2}} \\
&\le \frac{1}{4}(B-b)(A-a),
\end{aligned}
$$

with the constants $\frac{1}{2}$ and $\frac{1}{4}$ best possible.

Corollary 2 *Let $\Phi : [m, M] \to \mathbb{R}$ be a differentiable convex function on (m, M). If $x_i \in [m, M]$ and $w_i \geq 0$ $(i = 1, \ldots, n)$ with $W_n := \sum_{i=1}^{n} w_i = 1$, then one has the reverse of Jensen's weighted discrete inequality:*

$$\begin{aligned} 0 &\leq \sum_{i=1}^{n} w_i \Phi(x_i) - \Phi\left(\sum_{i=1}^{n} w_i x_i\right) \\ &\leq \sum_{i=1}^{n} w_i \Phi'(x_i) x_i - \sum_{i=1}^{n} w_i \Phi'(x_i) \sum_{i=1}^{n} w_i x_i \\ &\leq \frac{1}{2}\left[\Phi'(M) - \Phi'(m)\right] \sum_{i=1}^{n} w_i \left| x_i - \sum_{j=1}^{n} w_j x_j \right|. \end{aligned} \tag{1.20}$$

Remark 5 We notice that the inequality between the first and second term in (1.20) was proved in 1994 by Dragomir and Ionescu, see [25].

On utilizing (1.20) and (1.19) we can state the string of inequalities

$$\begin{aligned} 0 &\leq \sum_{i=1}^{n} w_i \Phi(x_i) - \Phi\left(\sum_{i=1}^{n} w_i x_i\right) \\ &\leq \sum_{i=1}^{n} w_i \Phi'(x_i) x_i - \sum_{i=1}^{n} w_i \Phi'(x_i) \sum_{i=1}^{n} w_i x_i \\ &\leq \frac{1}{2}\left[\Phi'(M) - \Phi'(m)\right] \sum_{i=1}^{n} w_i \left| x_i - \sum_{j=1}^{n} w_j x_j \right| \\ &\leq \frac{1}{2}\left[\Phi'(M) - \Phi'(m)\right] \left[\sum_{i=1}^{n} w_i x_i^2 - \left(\sum_{i=1}^{n} w_i x_i\right)^2\right]^{1/2} \\ &\leq \frac{1}{4}\left[\Phi'(M) - \Phi'(m)\right](M - m). \end{aligned} \tag{1.21}$$

We notice that the inequality between the first, second and last term in (1.21) was proved in 1999 by Dragomir in [12].

In this paper we survey several new reverses of the celebrated Jensen's inequality for convex functions and Lebesgue integral on measurable spaces. Applications for weighted discrete means, to Hölder inequality, Cauchy-Bunyakovsky-Schwarz inequality and for f-divergence measures in information theory are also given. Finally, applications for functions of selfadjoint operators in Hilbert spaces with some examples of interest are also provided.

2 A Refinement and a Divided-Difference Reverse

2.1 General Results

Following Roberts and Varberg [45, p. 5], we recall that if $f : I \to \mathbb{R}$ is a convex function, then for any $x_0 \in \overset{\circ}{I}$ (the interior of the interval I) the limits

$$f'_{-}(x_0) := \lim_{x \to x_0 -} \frac{f(x) - f(x_0)}{x - x_0} \text{ and } f'_{+}(x_0) := \lim_{x \to x_0 +} \frac{f(x) - f(x_0)}{x - x_0}$$

exists and $f'_{-}(x_0) \le f'_{+}(x_0)$. The functions f'_{-} and f'_{+} are monotonic nondecreasing on $\overset{\circ}{I}$ and this property can be extended to the whole interval I (see [45, p. 7]).

From the monotonicity of the lateral derivatives f'_{-} and f'_{+} we also have *the gradient inequality*

$$f'_{-}(x)(x - y) \ge f(x) - f(y) \ge f'_{+}(y)(x - y)$$

for any $x, y \in \overset{\circ}{I}$.

If $I = [a, b]$, then at the end points we also have the inequalities

$$f(x) - f(a) \ge f'_{+}(a)(x - a)$$

for any $x \in (a, b]$ and

$$f(y) - f(b) \ge f'_{-}(b)(y - b)$$

for any $y \in [a, b)$.

For a real function $g : [m, M] \to \mathbb{R}$ and two distinct points $\alpha, \beta \in [m, M]$ we recall that the *divided difference* of g in these points is defined by

$$[\alpha, \beta; g] := \frac{g(\beta) - g(\alpha)}{\beta - \alpha}.$$

In what follows, we assume that $w : \Omega \to \mathbb{R}$, with $w(x) \ge 0$ for μ-a.e. $x \in \Omega$, is a μ-measurable function with $\int_\Omega w d\mu = 1$.

Theorem 3 (Dragomir [23]) *Let $\Phi : I \to \mathbb{R}$ be a continuous convex function on the interval of real numbers I and $m, M \in \mathbb{R}$, $m < M$ with $[m, M] \subset \overset{\circ}{I}$, $\overset{\circ}{I}$ the interior of I. If $f : \Omega \to \mathbb{R}$, is μ-measurable, satisfying the bounds*

$$-\infty < m \le f(x) \le M < \infty \text{ for } \mu\text{-a.e. } x \in \Omega \tag{2.1}$$

and such that $f, \Phi \circ f \in L_w(\Omega, \mu)$, then by denoting

$$\overline{f}_{\Omega,w} := \int_\Omega w f d\mu \in [m, M]$$

and assuming that $\overline{f}_{\Omega,w} \neq m, M$, *we have*

$$\begin{aligned}
&\left|\int_\Omega \left|\Phi(f) - \Phi\left(\overline{f}_{\Omega,w}\right)\right| \operatorname{sgn}\left[f - \overline{f}_{\Omega,w}\right] w d\mu\right| \qquad (2.2)\\
&\leq \int_\Omega (\Phi \circ f)\, w d\mu - \Phi\left(\overline{f}_{\Omega,w}\right)\\
&\leq \frac{1}{2}\left(\left[\overline{f}_{\Omega,w}, M; \Phi\right] - \left[m, \overline{f}_{\Omega,w}; \Phi\right]\right) D_w(f)\\
&\leq \frac{1}{2}\left(\left[\overline{f}_{\Omega,w}, M; \Phi\right] - \left[m, \overline{f}_{\Omega,w}; \Phi\right]\right) D_{w,2}(f)\\
&\leq \frac{1}{4}\left(\left[\overline{f}_{\Omega,w}, M; \Phi\right] - \left[m, \overline{f}_{\Omega,w}; \Phi\right]\right)(M - m),
\end{aligned}$$

where sgn *is the sign function, i.e.* $\operatorname{sgn}(x) = \frac{x}{|x|}$ *for* $x \neq 0$ *and* $\operatorname{sgn}(0) = 0$. *The constant* $\frac{1}{2}$ *in the second inequality from (2.2) is best possible.*

Proof We recall that if $\Phi : I \to \mathbb{R}$ is a continuous convex function on the interval of real numbers I and $\alpha \in I$ then the divided difference function $\Phi_\alpha : I \setminus \{\alpha\} \to \mathbb{R}$,

$$\Phi_\alpha(t) := [\alpha, t; \Phi] := \frac{\Phi(t) - \Phi(\alpha)}{t - \alpha}$$

is monotonic nondecreasing on $I \setminus \{\alpha\}$.

For f as considered in the statement of the theorem we can assume that that it is not constant μ-almost every where, since for that case the inequality (2.2) is trivially satisfied.

For $\overline{f}_{\Omega,w} \in (m, M)$, we consider now the function defined μ-almost everywhere on Ω by

$$\Phi_{\overline{f}_{\Omega,w}}(x) := \frac{\Phi(f(x)) - \Phi\left(\overline{f}_{\Omega,w}\right)}{f(x) - \overline{f}_{\Omega,w}}.$$

We will show that $\Phi_{\overline{f}_{\Omega,w}}$ and $h := f - \overline{f}_{\Omega,w}$ are synchronous μ-a.e. on Ω.

Let $x, y \in \Omega$ with $f(x), f(y) \neq \overline{f}_{\Omega,w}$. Assume that $f(x) \geq f(y)$, then

$$\Phi_{\overline{f}_{\Omega,w}}(x) = \frac{\Phi(f(x)) - \Phi\left(\overline{f}_{\Omega,w}\right)}{f(x) - \overline{f}_{\Omega,w}} \geq \frac{\Phi(f(y)) - \Phi\left(\overline{f}_{\Omega,w}\right)}{f(y) - \overline{f}_{\Omega,w}} = \Phi_{\overline{f}_{\Omega,w}}(y) \tag{2.3}$$

and

$$h(x) \geq h(y), \tag{2.4}$$

which shows that

$$\left[\Phi_{\overline{f}_{\Omega,w}}(x)-\Phi_{\overline{f}_{\Omega,w}}(y)\right][h(x)-h(y)]\geq 0. \tag{2.5}$$

If $f(x)<f(y)$, then the inequalities (2.3) and (2.4) reverse but the inequality (2.5) still holds true.

This show that for μ-a.e. $x, y\in\Omega$ we have (2.5) and the claim is proven as stated.

Utilising the continuity property of the modulus we have

$$\begin{aligned}&\left|\left[\left|\Phi_{\overline{f}_{\Omega,w}}(x)\right|-\left|\Phi_{\overline{f}_{\Omega,w}}(y)\right|\right][h(x)-h(y)]\right|\\&\leq\left|\left[\Phi_{\overline{f}_{\Omega,w}}(x)-\Phi_{\overline{f}_{\Omega,w}}(y)\right][h(x)-h(y)]\right|\\&=\left[\Phi_{\overline{f}_{\Omega,w}}(x)-\Phi_{\overline{f}_{\Omega,w}}(y)\right][h(x)-h(y)]\end{aligned}$$

for μ-a.e. $x, y\in\Omega$.

Multiplying with $w(x), w(y)\geq 0$ and integrating over $\mu(x)$ and $\mu(y)$ we have

$$\begin{aligned}&\left|\int_\Omega\int_\Omega\left[\left|\Phi_{\overline{f}_{\Omega,w}}(x)\right|-\left|\Phi_{\overline{f}_{\Omega,w}}(y)\right|\right]\right. \\&\times[h(x)-h(y)]\,w(x)\,w(y)\,d\mu(x)\,d\mu(y)|\\&\leq\int_\Omega\int_\Omega\left[\Phi_{\overline{f}_{\Omega,w}}(x)-\Phi_{\overline{f}_{\Omega,w}}(y)\right]\\&\times[h(x)-h(y)]\,w(x)\,w(y)\,d\mu(x)\,d\mu(y).\end{aligned}\tag{2.6}$$

A simple calculation shows that

$$\begin{aligned}&\frac{1}{2}\int_\Omega\int_\Omega\left[\left|\Phi_{\overline{f}_{\Omega,w}}(x)\right|-\left|\Phi_{\overline{f}_{\Omega,w}}(y)\right|\right]\\&\times[h(x)-h(y)]\,w(x)\,w(y)\,d\mu(x)\,d\mu(y)\\&=\int_\Omega\left|\Phi_{\overline{f}_{\Omega,w}}(x)\right|h(x)\,w(x)\,d\mu(x)\\&-\int_\Omega\left|\Phi_{\overline{f}_{\Omega,w}}(x)\right|w(x)\,d\mu(x)\int_\Omega w(x)\,h(x)\,d\mu(x)\\&=\int_\Omega\left|\frac{\Phi(f(x))-\Phi\left(\overline{f}_{\Omega,w}\right)}{f(x)-\overline{f}_{\Omega,w}}\right|\left[f(x)-\overline{f}_{\Omega,w}\right]w(x)\,d\mu(x)\\&=\int_\Omega\left|\Phi(f(x))-\Phi\left(\overline{f}_{\Omega,w}\right)\right|\operatorname{sgn}\left[f(x)-\overline{f}_{\Omega,w}\right]w(x)\,d\mu(x)\end{aligned}\tag{2.7}$$

and

$$\begin{aligned}
&\frac{1}{2}\int_{\Omega}\int_{\Omega}\left[\Phi_{\overline{f}_{\Omega,w}}(x)-\Phi_{\overline{f}_{\Omega,w}}(y)\right] \\
&\times[h(x)-h(y)]\,w(x)\,w(y)\,d\mu(x)\,d\mu(y) \\
&=\int_{\Omega}\Phi_{\overline{f}_{\Omega,w}}(x)\,h(x)\,w(x)\,d\mu(x) \\
&-\int_{\Omega}\Phi_{\overline{f}_{\Omega,w}}(x)\,w(x)\,d\mu(x)\int_{\Omega}h(x)\,w(x)\,d\mu(x) \\
&=\int_{\Omega}\frac{\Phi(f(x))-\Phi\left(\overline{f}_{\Omega,w}\right)}{f(x)-\overline{f}_{\Omega,w}}\left[f(x)-\overline{f}_{\Omega,w}\right]w(x)\,d\mu(x) \\
&=\int_{\Omega}\left[\Phi(f(x))-\Phi\left(\overline{f}_{\Omega,w}\right)\right]w(x)\,d\mu(x) \\
&=\int_{\Omega}w\,(\Phi\circ f)\,d\mu-\Phi\left(\overline{f}_{\Omega,w}\right).
\end{aligned} \tag{2.8}$$

On making use of the identities (2.7) and (2.8) we obtain from (2.6) the first inequality in (2.2).

Now, since f satisfies the condition (2.1) then we have that

$$\begin{aligned}
\left[m,\overline{f}_{\Omega,w};\Phi\right]=\frac{\Phi\left(\overline{f}_{\Omega,w}\right)-\Phi(m)}{\overline{f}_{\Omega,w}-m}&\le\Phi_{\overline{f}_{\Omega,w}}(x) \\
&\le\frac{\Phi(M)-\Phi\left(\overline{f}_{\Omega,w}\right)}{M-\overline{f}_{\Omega,w}}=\left[\overline{f}_{\Omega,w},M;\Phi\right]
\end{aligned} \tag{2.9}$$

for μ-a.e. $x\in\Omega$.

Applying now the Grüss' type inequality (1.7) and taking into account the second part of the equality in (2.7) we have that

$$\begin{aligned}
&\int_{\Omega}w\,(\Phi\circ f)\,d\mu-\Phi\left(\overline{f}_{\Omega,w}\right) \\
&\le\frac{1}{2}\left(\left[\overline{f}_{\Omega,w},M;\Phi\right]-\left[m,\overline{f}_{\Omega,w};\Phi\right]\right)\int_{\Omega}w\left|f-\overline{f}_{\Omega,w}\right|d\mu
\end{aligned}$$

which proves the second inequality in (2.2).

The other two bounds are obvious from the comments in the introduction.

It is obvious that from (2.2) we get the following reverse of the first Hermite-Hadamard inequality for the convex function $\Phi:[a,b]\to\mathbb{R}$

$$\begin{aligned}&\frac{1}{b-a}\int_a^b \Phi(t)\,dt-\Phi\left(\frac{a+b}{2}\right) \\ &\le\frac{1}{2}\left(\left[\frac{a+b}{2},b;\Phi\right]-\left[a,\frac{a+b}{2};\Phi\right]\right)D_w(e)\end{aligned} \tag{2.10}$$

where $e(t)=t,t\in[a,b]$.

Since a simple calculation shows that

$$\begin{aligned}&\frac{1}{2}\left(\left[\frac{a+b}{2},b;\Phi\right]-\left[a,\frac{a+b}{2};\Phi\right]\right) \\ &=\frac{2}{b-a}\left[\frac{\Phi(a)+\Phi(b)}{2}-\Phi\left(\frac{a+b}{2}\right)\right]\end{aligned}$$

and

$$D_w(e)=\frac{1}{b-a}\int_a^b\left|t-\frac{a+b}{2}\right|dt=\frac{1}{4}(b-a),$$

and we get from (2.10) that

$$\begin{aligned}0&\le\frac{1}{b-a}\int_a^b \Phi(t)\,dt-\Phi\left(\frac{a+b}{2}\right) \\ &\le\frac{1}{2}\left[\frac{\Phi(a)+\Phi(b)}{2}-\Phi\left(\frac{a+b}{2}\right)\right].\end{aligned} \tag{2.11}$$

To prove the sharpness of the constant $\frac{1}{2}$ in the second inequality from (2.2) we need now only to show that the equality case in (2.11) is realized.

If we take, for instance $\Phi(t)=\left|t-\frac{a+b}{2}\right|$, $t\in[a,b]$, then we observe that Φ is convex and we get in both sides of (2.11) the same quantity $\frac{1}{4}(b-a)$. □

Corollary 3 *With the assumptions in Theorem 3 and if the lateral derivatives $\Phi'_+(m)$ and $\Phi'_-(M)$ are finite, then we have the inequalities*

$$\begin{aligned}0&\le\int_\Omega(\Phi\circ f)\,w\,d\mu-\Phi\left(\overline{f}_{\Omega,w}\right) \\ &\le\frac{1}{2}\left(\left[\overline{f}_{\Omega,w},M;\Phi\right]-\left[m,\overline{f}_{\Omega,w};\Phi\right]\right)D_w(f) \\ &\le\frac{1}{2}\left(\Phi'_-(M)-\Phi'_+(m)\right)D_w(f) \\ &\le\frac{1}{2}\left(\Phi'_-(M)-\Phi'_+(m)\right)D_{w,2}(f) \\ &\le\frac{1}{4}\left(\Phi'_-(M)-\Phi'_+(m)\right)(M-m).\end{aligned} \tag{2.12}$$

The constant $\frac{1}{2}$ in the second and third inequality from (2.12) is best possible.

Proof We need to prove only the third inequality.

By the convexity of Φ we have the gradient inequalities

$$\frac{\Phi(M)-\Phi\left(\overline{f}_{\Omega,w}\right)}{M-\overline{f}_{\Omega,w}}\leq\Phi'_{-}(M)$$

and

$$\frac{\Phi\left(\overline{f}_{\Omega,w}\right)-\Phi(m)}{\overline{f}_{\Omega,w}-m}\geq\Phi'_{+}(m).$$

These imply that

$$\left[\overline{f}_{\Omega,w},M;\Phi\right]-\left[m,\overline{f}_{\Omega,w};\Phi\right]\leq\Phi'_{-}(M)-\Phi'_{+}(m)$$

and the proof is concluded.

We observe that from (2.12) we get the following reverse of the Hermite-Hadamard inequality for the convex function $\Phi:[a,b]\to\mathbb{R}$ having finite lateral derivative $\Phi'_{+}(a)$ and $\Phi'_{-}(b)$

$$\begin{aligned}&\frac{1}{b-a}\int_a^b\Phi(t)\,dt-\Phi\left(\frac{a+b}{2}\right) \qquad (2.13)\\&\leq\frac{1}{2}\left[\frac{\Phi(a)+\Phi(b)}{2}-\Phi\left(\frac{a+b}{2}\right)\right]\leq\frac{1}{8}\left[\Phi'_{-}(b)-\Phi'_{+}(a)\right](b-a).\end{aligned}$$

We observe that the convex function $\Phi(t)=\left|t-\frac{a+b}{2}\right|$ has finite lateral derivatives

$$\Phi'_{-}(b)=1\text{ and }\Phi'_{+}(a)=-1$$

and replacing this function in (2.13) we get in all terms the same quantity $\frac{1}{4}(b-a)$.

This proves that the constant $\frac{1}{2}$ in the second and third inequality from (2.12) is best possible. □

Remark 6 Let $\Phi:I\to\mathbb{R}$ be a continuous convex function on the interval of real numbers I and $m,M\in\mathbb{R}$, $m<M$ with $[m,M]\subset\mathring{I}$, $\mathring{I}$ the interior of I. Let $\bar{\mathbf{a}}=(a_1,\ldots,a_n)$, $\bar{\mathbf{p}}=(p_1,\ldots,p_n)$ be n-tuples of real numbers with $p_i\geq 0$ $(i\in\{1,\ldots,n\})$ and $\sum_{i=1}^n p_i=1$. If $m\leq a_i\leq M$, $i\in\{1,\ldots,n\}$, with $\sum_{i=1}^n p_ia_i\neq m,M$, then

$$\left|\sum_{i=1}^n p_i\left[|\Phi(a_i)|-\left|\Phi\left(\sum_{i=1}^n p_ia_i\right)\right|\right]\operatorname{sgn}\left|a_i-\sum_{j=1}^n p_ja_j\right|\right| \qquad (2.14)$$

$$\leq \sum_{i=1}^{n} p_i \Phi(a_i) - \Phi\left(\sum_{i=1}^{n} p_i a_i\right)$$
$$\leq \frac{1}{2}\left(\left[\sum_{i=1}^{n} p_i a_i, M; \Phi\right] - \left[m, \sum_{i=1}^{n} p_i a_i; \Phi\right]\right) \sum_{i=1}^{n} p_i \left|a_i - \sum_{j=1}^{n} p_j a_j\right|.$$

If the lateral derivatives $\Phi'_+(m)$ and $\Phi'_-(M)$ are finite, then we also have the inequalities

$$0 \leq \sum_{i=1}^{n} p_i \Phi(a_i) - \Phi\left(\sum_{i=1}^{n} p_i a_i\right) \tag{2.15}$$
$$\leq \frac{1}{2}\left(\left[\sum_{i=1}^{n} p_i a_i, M; \Phi\right] - \left[m, \sum_{i=1}^{n} p_i a_i; \Phi\right]\right) \sum_{i=1}^{n} p_i \left|a_i - \sum_{j=1}^{n} p_j a_j\right|$$
$$\leq \frac{1}{2}\left(\Phi'_-(M) - \Phi'_+(m)\right) \sum_{i=1}^{n} p_i \left|a_i - \sum_{j=1}^{n} p_j a_j\right|.$$

Remark 7 Define the weighted arithmetic mean of the positive n-tuple $x = (x_1, \ldots, x_n)$ with the nonnegative weights $w = (w_1, \ldots, w_n)$ by

$$A_n(w, x) := \frac{1}{W_n} \sum_{i=1}^{n} w_i x_i$$

where $W_n := \sum_{i=1}^{n} w_i > 0$ and the weighted geometric mean of the same n-tuple, by

$$G_n(w, x) := \left(\prod_{i=1}^{n} x_i^{w_i}\right)^{1/W_n}.$$

It is well know that the following arithmetic mean-geometric mean inequality holds

$$A_n(w, x) \geq G_n(w, x).$$

Applying the inequality (2.15) for the convex function $\Phi(t) = -\ln t$, $t > 0$ we have the following reverse of the arithmetic mean-geometric mean inequality

$$1 \leq \frac{A_n(w, x)}{G_n(w, x)} \tag{2.16}$$

$$\leq \left[\frac{\left(\frac{A_n(w,x)}{m}\right)^{A_n(w,x)-m}}{\left(\frac{M}{A_n(w,x)}\right)^{M-A_n(w,x)}} \right]^{\frac{1}{2}A_n(w,|x-A_n(w,x)|)}$$

$$\leq \exp\left[\frac{1}{2}\frac{M-m}{mM} A_n\left(w, \left|x - A_n\left(w, x\right)\right|\right)\right],$$

provided that $0 < m \leq x_i \leq M < \infty$ for $i \in \{1, \ldots, n\}$.

2.2 Applications for the Hölder Inequality

It is well known that if $f \in L_p(\Omega, \mu)$, $p > 1$, where the Lebesgue space $L_p(\Omega, \mu)$ is defined by

$$L_p(\Omega, \mu) := \{f : \Omega \to \mathbb{R},\ f \text{ is } \mu\text{-measurable and } \int_\Omega |f(x)|^p \, d\mu(x) < \infty\}$$

and $g \in L_q(\Omega, \mu)$ with $\frac{1}{p} + \frac{1}{q} = 1$ then $fg \in L(\Omega, \mu) := L_1(\Omega, \mu)$ and the *Hölder inequality* holds true

$$\int_\Omega |fg| \, d\mu \leq \left(\int_\Omega |f|^p \, d\mu\right)^{1/p} \left(\int_\Omega |g|^p \, d\mu\right)^{1/q}.$$

Assume that $p > 1$. If $h : \Omega \to \mathbb{R}$ is μ-measurable, satisfies the bounds

$$-\infty < m \leq |h(x)| \leq M < \infty \text{ for } \mu\text{-a.e. } x \in \Omega$$

and is such that $h, |h|^p \in L_w(\Omega, \mu)$, for a μ-measurable function $w : \Omega \to \mathbb{R}$, with $w(x) \geq 0$ for μ-a.e. $x \in \Omega$ and $\int_\Omega w d\mu > 0$, then from (2.2) we have

$$\left|\int_\Omega \left||h|^p - \overline{|h|}^p_{\Omega,w}\right| \operatorname{sgn}\left[|h| - \overline{|h|}_{\Omega,w}\right] w d\mu\right| \tag{2.17}$$

$$\leq \frac{\int_\Omega |h|^p \, w d\mu}{\int_\Omega w d\mu} - \left(\frac{\int_\Omega |h| \, w d\mu}{\int_\Omega w d\mu}\right)^p$$

$$\leq \frac{1}{2}\left(\left[\overline{|h|}_{\Omega,w}, M; (\cdot)^p\right] - \left[m, \overline{|h|}_{\Omega,w}; (\cdot)^p\right]\right) \tilde{D}_w(|h|)$$

$$\leq \frac{1}{2}\left(\left[\overline{|h|}_{\Omega,w}, M; (\cdot)^p\right] - \left[m, \overline{|h|}_{\Omega,w}; (\cdot)^p\right]\right) \tilde{D}_{w,2}(|h|)$$

$$\leq \frac{1}{4}\left(\left[\overline{|h|}_{\Omega,w}, M; (\cdot)^p\right] - \left[m, \overline{|h|}_{\Omega,w}; (\cdot)^p\right]\right)(M - m),$$

where $\overline{|h|}_{\Omega,w} := \frac{\int_\Omega |h| w d\mu}{\int_\Omega w d\mu} \in [m, M]$ and

$$\tilde{D}_w(|h|) := \frac{1}{\int_\Omega w d\mu} \int_\Omega w \left| |h| - \frac{\int_\Omega |h| \, w d\mu}{\int_\Omega w d\mu} \right| d\mu$$

while

$$\tilde{D}_{w,2}(|h|) = \left[\frac{\int_\Omega w \, |h|^2 \, d\mu}{\int_\Omega w d\mu} - \left(\frac{\int_\Omega |h| \, w d\mu}{\int_\Omega w d\mu} \right)^2 \right]^{\frac{1}{2}}.$$

The following result related to the Hölder inequality holds:

Proposition 1 (Dragomir [23]) *If $f \in L_p(\Omega, \mu)$, $g \in L_q(\Omega, \mu)$ with $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$ and there exists the constants γ, $\Gamma > 0$ and such that*

$$\gamma \le \frac{|f|}{|g|^{q-1}} \le \Gamma \ \mu\text{-a.e. on } \Omega,$$

then we have

$$\begin{aligned} &\left| \int_\Omega \left| \frac{|f|^p}{|g|^q} - \left(\frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu} \right)^p \right| \operatorname{sgn} \left[\frac{|f|}{|g|^{q-1}} - \frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu} \right] |g|^q \, d\mu \right| \qquad (2.18) \\ &\le \frac{\int_\Omega |f|^p \, d\mu}{\int_\Omega |g|^q \, d\mu} - \left(\frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu} \right)^p \\ &\le \frac{1}{2} \left(\left[\frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu}, \Gamma; (\cdot)^p \right] - \left[\gamma, \frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu}; (\cdot)^p \right] \right) \tilde{D}_{|g|^q} \left(\frac{|f|}{|g|^{q-1}} \right) \\ &\le \frac{1}{2} \left(\left[\frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu}, \Gamma; (\cdot)^p \right] - \left[\gamma, \frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu}; (\cdot)^p \right] \right) \tilde{D}_{|g|^q,2} \left(\frac{|f|}{|g|^{q-1}} \right) \\ &\le \frac{1}{4} \left(\left[\frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu}, \Gamma; (\cdot)^p \right] - \left[\gamma, \frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu}; (\cdot)^p \right] \right) (\Gamma - \gamma), \end{aligned}$$

where

$$\tilde{D}_{|g|^q} \left(\frac{|f|}{|g|^{q-1}} \right) = \frac{1}{\int_\Omega |g|^q \, d\mu} \int_\Omega |g|^q \left| \frac{|f|}{|g|^{q-1}} - \frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu} \right| d\mu$$

and

$$\tilde{D}_{|g|^q,2} \left(\frac{|f|}{|g|^{q-1}} \right) = \left[\frac{1}{\int_\Omega |g|^q \, d\mu} \int_\Omega \frac{|f|^2}{|g|^{q-2}} d\mu - \left(\frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu} \right)^2 \right]^{\frac{1}{2}}.$$

Proof The inequalities (2.19) follow from (2.17) by choosing

$$h=\frac{|f|}{|g|^{q-1}} \text{ and } w=|g|^{q}.$$

The details are omitted. □

Remark 8 We observe that for $p=q=2$ we have from the first inequality in (2.18) the following reverse of the Cauchy-Bunyakovsky-Schwarz inequality

$$\left|\int_{\Omega}\left|\frac{|f|^2}{|g|^2}-\left(\frac{\int_{\Omega}|fg|\,d\mu}{\int_{\Omega}|g|^2\,d\mu}\right)^2\right|\operatorname{sgn}\left[\frac{|f|}{|g|}-\frac{\int_{\Omega}|fg|\,d\mu}{\int_{\Omega}|g|^2\,d\mu}\right]|g|^2\,d\mu\right| \tag{2.19}$$
$$\le\frac{\int_{\Omega}|f|^2\,d\mu}{\int_{\Omega}|g|^2\,d\mu}-\left(\frac{\int_{\Omega}|fg|\,d\mu}{\int_{\Omega}|g|^2\,d\mu}\right)^2$$
$$\le\frac{1}{2}(\Gamma-\gamma)\frac{1}{\int_{\Omega}|g|^2\,d\mu}\int_{\Omega}|g|^2\left|\frac{|f|}{|g|}-\frac{\int_{\Omega}|fg|\,d\mu}{\int_{\Omega}|g|^2\,d\mu}\right|d\mu$$
$$\le\frac{1}{2}(\Gamma-\gamma)\left[\frac{1}{\int_{\Omega}|g|^2\,d\mu}\int_{\Omega}|f|^2\,d\mu-\left(\frac{\int_{\Omega}|fg|\,d\mu}{\int_{\Omega}|g|^2\,d\mu}\right)^2\right]^{\frac{1}{2}}$$
$$\le\frac{1}{4}(\Gamma-\gamma)^2,$$

provided that $f, g\in L_2(\Omega,\mu)$, and there exists the constants $\gamma,\Gamma>0$ such that

$$\gamma\le\frac{|f|}{|g|}\le\Gamma\ \mu\text{-a.e. on }\Omega.$$

2.3 Applications for *f*-Divergence

One of the important issues in many applications of Probability Theory is finding an appropriate measure of *distance* (or *difference* or *discrimination*) between two probability distributions. A number of divergence measures for this purpose have been proposed and extensively studied by Jeffreys [31], Kullback and Leibler [36], Rényi [44], Havrda and Charvat [28], Kapur [34], Sharma and Mittal [47], Burbea and Rao [4], Rao [43], Lin [37], Csiszár [9], Ali and Silvey [1], Vajda [54], Shioya and Da-te [48] and others (see for example [39] and the references therein).

These measures have been applied in a variety of fields such as: anthropology [43], genetics [39], finance, economics, and political science [46, 51, 52], biology [42], the analysis of contingency tables [27], approximation of probability distributions [8, 35], signal processing [32, 33] and pattern recognition [2, 6]. A number of these measures of distance are specific cases of Csiszár f-divergence and so further exploration of this concept will have a flow on effect to other measures of distance and to areas in which they are applied.

Assume that a set Ω and the σ-finite measure μ are given. Consider the set of all probability densities on μ to be $\mathcal{P} := \left\{p|p : \Omega \to \mathbb{R},\ p(x) \geq 0,\ \int_{\Omega} p(x)\, d\mu(x) = 1\right\}$.

Csiszár f-divergence is defined as follows [10]

$$I_f(p,q) := \int_{\Omega} p(x)\, f\left[\frac{q(x)}{p(x)}\right] d\mu(x), \quad p,\ q \in \mathcal{P}, \tag{2.20}$$

where f is convex on $(0, \infty)$. It is assumed that $f(u)$ is zero and strictly convex at $u = 1$. By appropriately defining this convex function, various divergences are derived.

The *Kullback-Leibler divergence* [36] is well known among the information divergences. It is defined as:

$$D_{KL}(p,q) := \int_{\Omega} p(x) \ln\left[\frac{p(x)}{q(x)}\right] d\mu(x), \quad p,\ q \in \mathcal{P}, \tag{2.21}$$

where ln is to base e.

In Information Theory and Statistics, various divergences are applied in addition to the Kullback-Leibler divergence. These are the: *variation distance D_v*, *Hellinger distance D_H* [29], *χ^2-divergence D_{χ^2}*, *α-divergence D_α*, *Bhattacharyya distance D_B* [3], *Harmonic distance D_{Ha}*, *Jeffrey's distance D_J* [31], *triangular discrimination D_Δ* [53], etc... They are defined as follows:

$$D_v(p,q) := \int_{\Omega} |p(x) - q(x)|\, d\mu(x), \quad p,\ q \in \mathcal{P}; \tag{2.22}$$

$$D_H(p,q) := \int_{\Omega} \left|\sqrt{p(x)} - \sqrt{q(x)}\right| d\mu(x), \quad p,\ q \in \mathcal{P}; \tag{2.23}$$

$$D_{\chi^2}(p,q) := \int_{\Omega} p(x)\left[\left(\frac{q(x)}{p(x)}\right)^2 - 1\right] d\mu(x), \quad p,\ q \in \mathcal{P}; \tag{2.24}$$

$$D_\alpha(p,q) := \frac{4}{1-\alpha^2}\left[1 - \int_{\Omega} [p(x)]^{\frac{1-\alpha}{2}} [q(x)]^{\frac{1+\alpha}{2}}\, d\mu(x)\right], \quad p,\ q \in \mathcal{P}; \tag{2.25}$$

$$D_B(p,q) := \int_{\Omega} \sqrt{p(x)\, q(x)} d\mu(x), \quad p,\ q \in \mathcal{P}; \tag{2.26}$$

$$D_{Ha}(p,q) := \int_{\Omega} \frac{2p(x)\, q(x)}{p(x) + q(x)} d\mu(x), \quad p,\ q \in \mathcal{P}; \tag{2.27}$$

$$D_J(p,q) := \int_\Omega [p(x) - q(x)] \ln\left[\frac{p(x)}{q(x)}\right] d\mu(x), \quad p,\ q \in \mathcal{P}; \tag{2.28}$$

$$D_\Delta(p,q) := \int_\Omega \frac{[p(x) - q(x)]^2}{p(x) + q(x)} d\mu(x), \quad p,\ q \in \mathcal{P}. \tag{2.29}$$

For other divergence measures, see the paper [34] by Kapur or the book on line [50] by Taneja.

Most of the above distances (2.21)–(2.29), are particular instances of Csiszár f-divergence. There are also many others which are not in this class (see for example [50]). For the basic properties of Csiszár f-divergence see [10, 11] and [54].

Before we apply the results obtained in the previous section we observe that, by employing the inequalities from (1.17) we can state the following theorem:

Proposition 2 (Dragomir [23]) *Let $f : (0,\infty) \to \mathbb{R}$ be a convex function with the property that $f(1) = 0$. Assume that $p, q \in \mathcal{P}$ and there exists the constants $0 < r < 1 < R < \infty$ such that*

$$r \le \frac{q(x)}{p(x)} \le R \text{ for } \mu\text{-a.e. } x \in \Omega. \tag{2.30}$$

Then we have

$$\begin{aligned} 0 \le I_f(p,q) &\le \frac{1}{2}\left[f'_-(R) - f'_+(r)\right] D_v(p,q) \\ &\le \frac{1}{2}\left[f'_-(R) - f'_+(r)\right]\left[D_{\chi^2}(p,q)\right]^{1/2} \\ &\le \frac{1}{4}(R - r)\left[f'_-(R) - f'_+(r)\right]. \end{aligned} \tag{2.31}$$

Proof From (1.17) we have

$$\begin{aligned} &\int_\Omega p(x) f\left(\frac{q(x)}{p(x)}\right) d\mu(x) - f\left(\int_\Omega q(x)\, d\mu(x)\right) \\ &\le \frac{1}{2}\left[f'_-(R) - f'_+(r)\right] \\ &\times \int_\Omega p(x)\left|\frac{q(x)}{p(x)} - \int_\Omega q(y)\, d\mu(y)\right| d\mu(x) \\ &\le \frac{1}{2}\left[f'_-(R) - f'_+(r)\right] \\ &\times \left[\int_\Omega p(x)\left(\frac{q(x)}{p(x)}\right)^2 d\mu - \left(\int_\Omega q(x)\, d\mu\right)^2\right]^{\frac{1}{2}} \\ &\le \frac{1}{4}(R - r)\left[f'_-(R) - f'_+(r)\right], \end{aligned} \tag{2.32}$$

and since

$$\int_{\Omega} p(x) \left| \frac{q(x)}{p(x)} - \int_{\Omega} q(y)\, d\mu(y) \right| d\mu(x) = D_v(p, q)$$

and

$$\int_{\Omega} p(x) \left(\frac{q(x)}{p(x)} \right)^2 d\mu - \left(\int_{\Omega} q(x)\, d\mu \right)^2 = D_{\chi^2}(p, q),$$

then we get from (2.32) the desired result (2.31). □

Remark 9 The inequality

$$I_f(p, q) \le \frac{1}{4} (R - r) \left[f'_-(R) - f'_+(r) \right] \tag{2.33}$$

was obtained for the discrete divergence measures in 2000 by Dragomir, see [15].

Proposition 3 (Dragomir [23]) *With the assumptions in Proposition 2 we have*

$$\begin{aligned} \left| I_{|f|(\operatorname{sgn}(\cdot)-1)}(p, q) \right| &\le I_f(p, q) \qquad (2.34) \\ &\le \frac{1}{2} ([1, R; f] - [r, 1; f]) D_v(p, q) \\ &\le \frac{1}{2} ([1, R; f] - [r, 1; f]) \left[D_{\chi^2}(p, q) \right]^{1/2} \\ &\le \frac{1}{4} ([1, R; f] - [r, 1; f]) (R - r), \end{aligned}$$

where $I_{|f|(\operatorname{sgn}(\cdot)-1)}(p, q)$ is the generalized f-divergence for the non-necessarily convex function $|f| (\operatorname{sgn}(\cdot) - 1)$ and is defined by

$$I_{|f|(\operatorname{sgn}(\cdot)-1)}(p, q) := \int_{\Omega} \left| f\left(\frac{q(x)}{p(x)} \right) \right| \operatorname{sgn}\left[\frac{q(x)}{p(x)} - 1 \right] p(x)\, d\mu. \tag{2.35}$$

Proof From the inequality (2.2) we have

$$\begin{aligned} &\left| \int_{\Omega} \left| f\left(\frac{q(x)}{p(x)} \right) \right| \operatorname{sgn}\left[\frac{q(x)}{p(x)} - 1 \right] p(x)\, d\mu. \right| \qquad (2.36) \\ &\le \int_{\Omega} p(x) f\left(\frac{q(x)}{p(x)} \right) d\mu(x) - f\left(\int_{\Omega} q(x)\, d\mu(x) \right) \\ &\le \frac{1}{2} ([1, R; f] - [r, 1; f]) \end{aligned}$$

$$
\begin{aligned}
&\times \int_{\Omega} p(x) \left| \frac{q(x)}{p(x)} - \int_{\Omega} q(y)\, d\mu(y) \right| d\mu(x) \\
&\leq \frac{1}{2} ([1, R; f] - [r, 1; f]) \\
&\times \left[\int_{\Omega} p(x) \left(\frac{q(x)}{p(x)} \right)^2 d\mu - \left(\int_{\Omega} q(x)\, d\mu \right)^2 \right]^{\frac{1}{2}} \\
&\leq \frac{1}{4} ([1, R; f] - [r, 1; f]) (R - r),
\end{aligned}
$$

from where we get the desired result (2.34). □

The above results can be utilized to obtain various inequalities for the divergence measures in Information Theory that are particular instances of f-divergence.

Consider the *Kullback-Leibler divergence*

$$D_{KL}(p, q) := \int_{\Omega} p(x) \ln \left[\frac{p(x)}{q(x)} \right] d\mu(x), \quad p,\ q \in \mathcal{P},$$

which is an f-divergence for the convex function $f : (0, \infty) \to \mathbb{R}$, $f(t) = -\ln t$.

If $p, q \in \mathcal{P}$ such that there exists the constants $0 < r < 1 < R < \infty$ with

$$r \leq \frac{q(x)}{p(x)} \leq R \text{ for } \mu\text{-a.e. } x \in \Omega, \tag{2.37}$$

then we get from (2.31) that

$$
\begin{aligned}
D_{KL}(p, q) &\leq \frac{R - r}{2rR} D_v(p, q) \\
&\leq \frac{R - r}{2rR} \left[D_{\chi^2}(p, q) \right]^{1/2} \leq \frac{(R - r)^2}{4rR}
\end{aligned}
\tag{2.38}
$$

and from (2.34) that

$$
\begin{aligned}
D_{KL}(p, q) &\leq \frac{1}{2} D_v(p, q) \ln \left(\frac{1}{R^{R-1} r^{1-r}} \right) \\
&\leq \frac{1}{2} \left[D_{\chi^2}(p, q) \right]^{1/2} \ln \left(\frac{1}{R^{R-1} r^{1-r}} \right) \\
&\leq \frac{1}{4} (R - r) \ln \left(\frac{1}{R^{R-1} r^{1-r}} \right).
\end{aligned}
\tag{2.39}
$$

The interested reader can obtain other similar results by considering f-divergence measures generated by other convex functions such as the *Jeffrey's distance* D_J or the *triangular discrimination* D_Δ. The details are omitted.

3 Reverse Inequalities in Terms of First Derivative

3.1 General Results

The following reverse of the Jensen's inequality holds:

Theorem 4 (Dragomir [21]) *Let $\Phi : I \to \mathbb{R}$ be a continuous convex function on the interval of real numbers I and $m, M \in \mathbb{R}$, $m < M$ with $[m, M] \subset \overset{\circ}{I}$, where $\overset{\circ}{I}$ is the interior of I. If $f : \Omega \to \mathbb{R}$ is μ-measurable, satisfies the bounds*

$$-\infty < m \le f(x) \le M < \infty \text{ for } \mu\text{-a.e. } x \in \Omega$$

and such that $f, \Phi \circ f \in L_w(\Omega, \mu)$, then

$$\begin{aligned} 0 &\le \int_\Omega w(x)\,\Phi(f(x))\,d\mu(x) - \Phi\left(\bar{f}_{\Omega,w}\right) \qquad (3.1)\\ &\le \left(M - \bar{f}_{\Omega,w}\right)\left(\bar{f}_{\Omega,w} - m\right)\frac{\Phi'_-(M) - \Phi'_+(m)}{M - m}\\ &\le \frac{1}{4}(M - m)\left[\Phi'_-(M) - \Phi'_+(m)\right], \end{aligned}$$

where $\bar{f}_{\Omega,w} := \int_\Omega w(x) f(x)\,d\mu(x) \in [m, M]$, Φ'_- is the left and Φ'_+ is the right derivative of the convex function Φ.

Proof By the convexity of Φ we have that

$$\begin{aligned} &\int_\Omega w(x)\,\Phi(f(x))\,d\mu(x) - \Phi\left(\bar{f}_{\Omega,w}\right) \qquad (3.2)\\ &= \int_\Omega w(x)\,\Phi\left[\frac{m(M - f(x)) + M(f(x) - m)}{M - m}\right]d\mu(x)\\ &- \Phi\left(\int_\Omega w(x)\left[\frac{m(M - f(x)) + M(f(x) - m)}{M - m}\right]d\mu(x)\right)\\ &\le \int_\Omega \frac{(M - f(x))\,\Phi(m) + (f(x) - m)\,\Phi(M)}{M - m}w(x)\,d\mu(x)\\ &- \Phi\left(\frac{m\left(M - \bar{f}_{\Omega,w}\right) + M\left(\bar{f}_{\Omega,w} - m\right)}{M - m}\right) \end{aligned}$$

$$= \frac{\left(M - \bar{f}_{\Omega,w}\right)\Phi(m) + \left(\bar{f}_{\Omega,w} - m\right)\Phi(M)}{M-m}$$
$$- \Phi\left(\frac{m\left(M - \bar{f}_{\Omega,w}\right) + M\left(\bar{f}_{\Omega,w} - m\right)}{M-m}\right) := B.$$

Then, by the convexity of Φ we have the gradient inequality

$$\Phi(t) - \Phi(M) \geq \Phi'_{-}(M)(t-M)$$

for any $t \in [m, M)$. If we multiply this inequality with $t - m \geq 0$, we deduce

$$(t-m)\Phi(t) - (t-m)\Phi(M) \geq \Phi'_{-}(M)(t-M)(t-m), \quad t \in [m, M]. \tag{3.3}$$

Similarly, using the other gradient inequality

$$\Phi(t) - \Phi(m) \geq \Phi'_{+}(m)(t-m)$$

for any $t \in (m, M]$, we also get

$$(M-t)\Phi(t) - (M-t)\Phi(m) \geq \Phi'_{+}(m)(t-m)(M-t), \quad t \in [m, M]. \tag{3.4}$$

Adding (3.3) to (3.4) and dividing by $M - m$, we deduce

$$\Phi(t) - \frac{(t-m)\Phi(M) + (M-t)\Phi(m)}{M-m} \geq \frac{(t-M)(t-m)}{M-m}\left[\Phi'_{-}(M) - \Phi'_{+}(m)\right],$$

for any $t \in (m, M)$.

By denoting

$$\Delta_{\Phi}(t; m, M) := \frac{(t-m)\Phi(M) + (M-t)\Phi(m)}{M-m} - \Phi(t), \quad t \in [m, M]$$

we then get the following inequality of interest

$$0 \leq \Delta_{\Phi}(t; m, M) \leq \frac{(M-t)(t-m)}{M-m}\left[\Phi'_{-}(M) - \Phi'_{+}(m)\right] \tag{3.5}$$
$$\leq \frac{1}{4}(M-m)\left[\Phi'_{-}(M) - \Phi'_{+}(m)\right]$$

for any $t \in (m, M)$.

Now, since with the above notations we have $B = \Delta_{\Phi}\left(\bar{f}_{\Omega,w}; m, M\right)$, then by (3.5) we have

$$\begin{aligned} B &\leq \frac{\left(M - \bar{f}_{\Omega,w}\right)\left(\bar{f}_{\Omega,w} - m\right)}{M - m}\left[\Phi'_{-}(M) - \Phi'_{+}(m)\right] \\ &\leq \frac{1}{4}(M - m)\left[\Phi'_{-}(M) - \Phi'_{+}(m)\right], \end{aligned}$$

and the proof is completed. □

Corollary 4 *Let $\Phi : I \to \mathbb{R}$ be a continuous convex function on the interval of real numbers I and m, $M \in \mathbb{R}$, $m < M$ with $[m, M] \subset \overset{\circ}{I}$. If $x_i \in I$ and $p_i \geq 0$ for $i \in \{1, \ldots, n\}$ with $\sum_{i=1}^{n} p_i = 1$, then we have the inequality*

$$\begin{aligned} 0 &\leq \sum_{i=1}^{n} p_i \Phi(x_i) - \Phi\left(\bar{x}_p\right) \\ &\leq \left(M - \bar{x}_p\right)\left(\bar{x}_p - m\right)\frac{\Phi'_{-}(M) - \Phi'_{+}(m)}{M - m} \\ &\leq \frac{1}{4}(M - m)\left[\Phi'_{-}(M) - \Phi'_{+}(m)\right], \end{aligned} \tag{3.6}$$

where $\bar{x}_p := \sum_{i=1}^{n} p_i x_i \in I$.

Remark 10 Consider the positive n-tuple $x = (x_1, \ldots, x_n)$ with the nonnegative weights $w = (w_1, \ldots, w_n)$ where $W_n := \sum_{i=1}^{n} w_i > 0$. Applying the inequality (3.6) for the convex function $\Phi(t) = -\ln t$, $t > 0$ we have

$$\begin{aligned} 1 &\leq \frac{A_n(w, x)}{G_n(w, x)} \leq \exp\left[\frac{1}{Mm}(M - A_n(w, x))(A_n(w, x) - m)\right] \\ &\leq \exp\left[\frac{1}{4}\frac{(M - m)^2}{mM}\right], \end{aligned} \tag{3.7}$$

provided that $0 < m \leq x_i \leq M < \infty$ for $i \in \{1, \ldots, n\}$.

For the Lebesgue measurable function $g : [\alpha, \beta] \to \mathbb{R}$ we introduce the *Lebesgue p-norms* defined as

$$\|g\|_{[\alpha,\beta],p} := \left(\int_{\alpha}^{\beta} |g(t)|^p \, dt\right)^{1/p} \text{ if } g \in L_p[\alpha, \beta],$$

for $p \geq 1$ and

$$\|g\|_{[\alpha,\beta],\infty} := \operatorname*{essup}_{t \in [\alpha,\beta]} |g(t)| \text{ if } g \in L_\infty[\alpha, \beta],$$

for $p = \infty$.

The following result also holds:

Theorem 5 (Dragomir [21]) *With the assumptions in Theorem 4, we have the inequalities*

$$\begin{aligned}0 &\le \int_\Omega w(x)\,\Phi(f(x))\,d\mu(x) - \Phi\left(\bar{f}_{\Omega,w}\right) \\ &\le \frac{\left(M-\bar{f}_{\Omega,w}\right)\int_m^{\bar{f}_{\Omega,w}}\left|\Phi'(t)\right|dt + \left(\bar{f}_{\Omega,w}-m\right)\int_{\bar{f}_{\Omega,w}}^M\left|\Phi'(t)\right|dt}{M-m} \\ &:= \Lambda_\Phi\left(\bar{f}_{\Omega,w}; m, M\right),\end{aligned} \tag{3.8}$$

where the integral in the second term of the inequality is taken in the Lebesgue sense.

We also have the bounds:

$$\begin{aligned}&\Lambda_\Phi\left(\bar{f}_{\Omega,w}; m, M\right) \\ &\le \begin{cases}\left[\frac{1}{2}+\frac{\left|\bar{f}_{\Omega,w}-\frac{m+M}{2}\right|}{M-m}\right]\int_m^M\left|\Phi'(t)\right|dt, \\ \\ \left[\frac{1}{2}\int_m^M\left|\Phi'(t)\right|dt + \frac{1}{2}\left|\int_{\bar{f}_{\Omega,w}}^M\left|\Phi'(t)\right|dt - \int_m^{\bar{f}_{\Omega,w}}\left|\Phi'(t)\right|dt\right|\right]\end{cases}\end{aligned} \tag{3.9}$$

and

$$\begin{aligned}&\Lambda_\Phi\left(\bar{f}_{\Omega,w}; m, M\right) \\ &\le \frac{\left(\bar{f}_{\Omega,w}-m\right)\left(M-\bar{f}_{\Omega,w}\right)}{M-m}\left[\left\|\Phi'\right\|_{[\bar{f}_{\Omega,w},M],\infty} + \left\|\Phi'\right\|_{[m,\bar{f}_{\Omega,w}],\infty}\right] \\ &\le \frac{1}{2}(M-m)\frac{\left\|\Phi'\right\|_{[\bar{f}_{\Omega,w},M],\infty} + \left\|\Phi'\right\|_{[m,\bar{f}_{\Omega,w}],\infty}}{2} \le \frac{1}{2}(M-m)\left\|\Phi'\right\|_{[m,M],\infty}\end{aligned} \tag{3.10}$$

and

$$\begin{aligned}\Lambda_\Phi\left(\bar{f}_{\Omega,w}; m, M\right) &\le \frac{1}{M-m}\left[\left(\bar{f}_{\Omega,w}-m\right)\left(M-\bar{f}_{\Omega,w}\right)^{1/q}\left\|\Phi'\right\|_{[\bar{f}_{\Omega,w},M],p}\right. \\ &\quad \left. + \left(M-\bar{f}_{\Omega,w}\right)\left(\bar{f}_{\Omega,w}-m\right)^{1/q}\left\|\Phi'\right\|_{[m,\bar{f}_{\Omega,w}],p}\right] \\ &\le \frac{1}{M-m}\left[\left(\bar{f}_{\Omega,w}-m\right)^q\left(M-\bar{f}_{\Omega,w}\right)\right. \\ &\quad \left. + \left(M-\bar{f}_{\Omega,w}\right)^q\left(\bar{f}_{\Omega,w}-m\right)\right]^{1/q}\left\|\Phi'\right\|_{[m,M],p}\end{aligned} \tag{3.11}$$

where $p>1$, $\frac{1}{p}+\frac{1}{q}=1$.

Proof Observe that, with the above notations we have

$$\begin{aligned}\Lambda_\Phi(t;m,M) &= \frac{(t-m)\,\Phi(M)+(M-t)\,\Phi(m)}{M-m}-\Phi(t) \qquad (3.12)\\ &= \frac{(t-m)\,\Phi(M)+(M-t)\,\Phi(m)-(M-m)\,\Phi(t)}{M-m}\\ &= \frac{(t-m)\,\Phi(M)+(M-t)\,\Phi(m)-(M-t+t-m)\,\Phi(t)}{M-m}\\ &= \frac{(t-m)\,[\Phi(M)-\Phi(t)]-(M-t)\,[\Phi(t)-\Phi(m)]}{M-m}\end{aligned}$$

for any $t\in[m,M]$.

Taking the modulus on (3.12) and noticing that $\Lambda_\Phi(t;m,M)\geq 0$ for any $t\in[m,M]$, we have that

$$\begin{aligned}\Lambda_\Phi(t;m,M) &\leq \frac{(t-m)\,|\Phi(M)-\Phi(t)|+(M-t)\,|\Phi(t)-\Phi(m)|}{M-m} \qquad (3.13)\\ &= \frac{(t-m)\left|\int_t^M \Phi'(s)\,ds\right|+(M-t)\left|\int_m^t \Phi'(s)\,ds\right|}{M-m}\\ &\leq \frac{(t-m)\int_t^M\left|\Phi'(s)\right|ds+(M-t)\int_m^t\left|\Phi'(s)\right|ds}{M-m}\end{aligned}$$

for any $t\in[m,M]$.

Finally, if we write the inequality (3.13) for $t=\bar{f}_{\Omega,w}\in[m,M]$ and utilize the inequality (3.2), we deduce the desired result (3.8).

Now, we observe that

$$\begin{aligned}&\frac{(t-m)\int_t^M\left|\Phi'(s)\right|ds+(M-t)\int_m^t\left|\Phi'(s)\right|ds}{M-m} \qquad (3.14)\\ &\leq \begin{cases}\max\{t-m,M-t\}\int_m^M\left|\Phi'(t)\right|dt\\ \\ \max\left\{\int_t^M\left|\Phi'(s)\right|ds,\int_m^t\left|\Phi'(s)\right|ds\right\}(M-m)\end{cases}\\ &= \begin{cases}\left[\frac{1}{2}(M-m)+\left|t-\frac{m+M}{2}\right|\right]\int_m^M\left|\Phi'(t)\right|dt\\ \\ \left[\frac{1}{2}\int_m^M\left|\Phi'(s)\right|ds+\frac{1}{2}\left|\int_t^M\left|\Phi'(s)\right|ds-\int_m^t\left|\Phi'(s)\right|ds\right|\right](M-m)\end{cases}\end{aligned}$$

for any $t\in[m,M]$. This proves the inequality (3.9).

By the Hölder's inequality we have

$$\int_t^M \left|\Phi'(s)\right| ds \leq \begin{cases} (M-t)\left\|\Phi'\right\|_{[t,M],\infty} \\ \\ (M-t)^{1/q}\left\|\Phi'\right\|_{[t,M],p} \text{ if } p>1,\ \frac{1}{p}+\frac{1}{q}=1 \end{cases}$$

and

$$\int_m^t \left|\Phi'(s)\right| ds \leq \begin{cases} (t-m)\left\|\Phi'\right\|_{[m,t],\infty} \\ \\ (t-m)^{1/q}\left\|\Phi'\right\|_{[m,t],p} \text{ if } p>1,\ \frac{1}{p}+\frac{1}{q}=1 \end{cases}$$

which give that

$$\begin{aligned} &\frac{(t-m)\int_t^M \left|\Phi'(s)\right| ds + (M-t)\int_m^t \left|\Phi'(s)\right| ds}{M-m} \qquad (3.15)\\ &\leq \frac{(t-m)(M-t)\left\|\Phi'\right\|_{[t,M],\infty} + (M-t)(t-m)\left\|\Phi'\right\|_{[m,t],\infty}}{M-m} \\ &= \frac{(t-m)(M-t)}{M-m}\left[\left\|\Phi'\right\|_{[t,M],\infty} + \left\|\Phi'\right\|_{[m,t],\infty}\right] \\ &\leq \frac{1}{2}(M-m)\frac{\left\|\Phi'\right\|_{[t,M],\infty} + \left\|\Phi'\right\|_{[m,t],\infty}}{2} \\ &\leq \frac{1}{2}(M-m)\max\left\{\left\|\Phi'\right\|_{[t,M],\infty}, \left\|\Phi'\right\|_{[m,t],\infty}\right\} = \frac{1}{2}(M-m)\left\|\Phi'\right\|_{[m,M],\infty} \end{aligned}$$

and

$$\begin{aligned} &\frac{(t-m)\int_t^M \left|\Phi'(s)\right| ds + (M-t)\int_m^t \left|\Phi'(s)\right| ds}{M-m} \qquad (3.16)\\ &\leq \frac{(t-m)(M-t)^{1/q}\left\|\Phi'\right\|_{[t,M],p} + (M-t)(t-m)^{1/q}\left\|\Phi'\right\|_{[m,t],p}}{M-m} \\ &\leq \frac{1}{M-m}\left[\left((t-m)(M-t)^{1/q}\right)^q + \left((M-t)(t-m)^{1/q}\right)^q\right]^{1/q} \\ &\times \left[\left\|\Phi'\right\|_{[t,M],p}^p + \left\|\Phi'\right\|_{[m,t],p}^p\right]^{1/p} \\ &= \frac{1}{M-m}\left[(t-m)^q(M-t) + (M-t)^q(t-m)\right]^{1/q}\left\|\Phi'\right\|_{[m,M],p} \end{aligned}$$

for any $t \in [m, M]$.

These prove the desired inequalities (3.10) and (3.11). □

The discrete case is as follows:

Corollary 5 *Let $\Phi : I \to \mathbb{R}$ be a continuous convex function on the interval of real numbers I and $m, M \in \mathbb{R}$, $m < M$ with $[m, M] \subset \overset{\circ}{I}$, $\overset{\circ}{I}$ is the interior of I. If $x_i \in I$ and $p_i \geq 0$ for $i \in \{1, \dots, n\}$ with $\sum_{i=1}^{n} p_i = 1$, then we have the inequality*

$$\begin{aligned} 0 &\leq \sum_{i=1}^{n} p_i \Phi(x_i) - \Phi(\bar{x}_p) &(3.17)\\ &\leq \frac{(M - \bar{x}_p) \int_m^{\bar{x}_p} |\Phi'(t)|\, dt + (\bar{x}_p - m) \int_{\bar{x}_p}^{M} |\Phi'(t)|\, dt}{M - m} \\ &:= \Lambda_\Phi(\bar{x}_p; m, M), \end{aligned}$$

where $\Lambda_\Phi(\bar{x}_p; m, M)$ satisfies the bounds

$$\begin{aligned} &\Lambda_\Phi(\bar{x}_p; m, M) &(3.18)\\ &\leq \begin{cases} \left[\frac{1}{2} + \frac{\left|\bar{x}_p - \frac{m+M}{2}\right|}{M-m}\right] \int_m^M |\Phi'(t)|\, dt, \\ \\ \left[\frac{1}{2} \int_m^M |\Phi'(t)|\, dt + \frac{1}{2} \left| \int_{\bar{x}_p}^M |\Phi'(t)|\, dt - \int_m^{\bar{x}_p} |\Phi'(t)|\, dt \right| \right] \end{cases} \end{aligned}$$

and

$$\begin{aligned} &\Lambda_\Phi(\bar{x}_p; m, M) &(3.19)\\ &\leq \frac{(\bar{x}_p - m)(M - \bar{x}_p)}{M - m} \left[\|\Phi'\|_{[\bar{x}_p, M], \infty} + \|\Phi'\|_{[m, \bar{x}_p], \infty} \right] \\ &\leq \frac{1}{2}(M - m) \frac{\|\Phi'\|_{[\bar{x}_p, M], \infty} + \|\Phi'\|_{[m, \bar{x}_p], \infty}}{2} \leq \frac{1}{2}(M - m) \|\Phi'\|_{[m, M], \infty} \end{aligned}$$

and

$$\begin{aligned} \Lambda_\Phi(\bar{x}_p; m, M) &\leq \frac{1}{M - m} \left[(\bar{x}_p - m)(M - \bar{x}_p)^{1/q} \|\Phi'\|_{[\bar{x}_p, M], p} \right. &(3.20)\\ &\quad \left. + (M - \bar{x}_p)(\bar{x}_p - m)^{1/q} \|\Phi'\|_{[m, \bar{x}_p], p} \right] \\ &\leq \frac{1}{M - m} \left[(\bar{x}_p - m)^q (M - \bar{x}_p) \right. \\ &\quad \left. + (M - \bar{x}_p)^q (\bar{x}_p - m) \right]^{1/q} \|\Phi'\|_{[m, M], p}. \end{aligned}$$

Remark 11 Under the assumptions of Remark 10 , on applying the inequality (3.17) for the convex function $\Phi(t) = -\ln t$, we have the following reverse of the arithmetic mean-geometric mean inequality

$$1 \leq \frac{A_n(w, x)}{G_n(w, x)} \leq \left(\frac{A_n(w, x)}{m} \right)^{M - A_n(w, x)} \left(\frac{M}{A_n(w, x)} \right)^{A_n(w, x) - m}. \qquad (3.21)$$

3.2 Applications for the Hölder Inequality

Assume that $p > 1$. If $h : \Omega \to \mathbb{R}$ is μ-measurable, satisfies the bounds

$$-\infty < m \le |h(x)| \le M < \infty \text{ for } \mu\text{-a.e. } x \in \Omega$$

and is such that $h,\ |h|^p \in L_w(\Omega, \mu)$, for a μ-measurable function $w : \Omega \to \mathbb{R}$, with $w(x) \ge 0$ for μ-a.e. $x \in \Omega$ and $\int_\Omega w d\mu > 0$, then from (3.1) we have

$$\begin{aligned} 0 &\le \frac{\int_\Omega |h|^p w d\mu}{\int_\Omega w d\mu} - \left(\frac{\int_\Omega |h| w d\mu}{\int_\Omega w d\mu}\right)^p \\ &\le p\frac{M^{p-1} - m^{p-1}}{M - m}\left(M - \overline{|h|}_{\Omega,w}\right)\left(\overline{|h|}_{\Omega,w} - m\right) \\ &\le \frac{1}{4}p(M - m)\left(M^{p-1} - m^{p-1}\right), \end{aligned} \tag{3.22}$$

where $\overline{|h|}_{\Omega,w} := \frac{\int_\Omega |h| w d\mu}{\int_\Omega w d\mu} \in [m, M]$.

Proposition 4 (Dragomir [21]) *If $f \in L_p(\Omega, \mu)$, $g \in L_q(\Omega, \mu)$ with $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$ and there exists the constants $\gamma, \Gamma > 0$ and such that*

$$\gamma \le \frac{|f|}{|g|^{q-1}} \le \Gamma \ \mu\text{-a.e. on } \Omega$$

then we have

$$\begin{aligned} 0 &\le \frac{\int_\Omega |f|^p d\mu}{\int_\Omega |g|^q d\mu} - \left(\frac{\int_\Omega |fg| d\mu}{\int_\Omega |g|^q d\mu}\right)^p \\ &\le p\frac{\Gamma^{p-1} - \gamma^{p-1}}{\Gamma - \gamma}\left(\Gamma - \frac{\int_\Omega |fg| d\mu}{\int_\Omega |g|^q d\mu}\right)\left(\frac{\int_\Omega |fg| d\mu}{\int_\Omega |g|^q d\mu} - \gamma\right) \\ &\le \frac{1}{4}p(\Gamma - \gamma)\left(\Gamma^{p-1} - \gamma^{p-1}\right). \end{aligned} \tag{3.23}$$

Proof The inequalities (3.23) follow from (3.22) by choosing

$$h = \frac{|f|}{|g|^{q-1}} \text{ and } w = |g|^q .$$

The details are omitted. □

Remark 12 We observe that for $p = q = 2$ we have from the first inequality in (3.23) the following reverse of the Cauchy-Bunyakovsky-Schwarz inequality

$$
\begin{aligned}
0 &\le \int_\Omega |g|^2 \, d\mu \int_\Omega |f|^2 \, d\mu - \left(\int_\Omega |fg| \, d\mu\right)^2 \qquad (3.24)\\
&\le \left(\Gamma - \frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^2 \, d\mu}\right)\left(\frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^2 \, d\mu} - \gamma\right)\left(\int_\Omega |g|^2 \, d\mu\right)^2\\
&\le \frac{1}{4}(\Gamma-\gamma)^2\left(\int_\Omega |g|^2 \, d\mu\right)^2,
\end{aligned}
$$

provided that $f, g \in L_2(\Omega,\mu)$ and there exists the constants $\gamma, \Gamma > 0$ such that

$$
\gamma \le \frac{|f|}{|g|} \le \Gamma \ \mu\text{-a.e. on } \Omega.
$$

Corollary 6 *With the assumptions of Proposition 4 we have the following additive reverses of the Hölder inequality:*

$$
\begin{aligned}
0 &\le \left(\int_\Omega |f|^p \, d\mu\right)^{1/p}\left(\int_\Omega |g|^q \, d\mu\right)^{1/q} - \int_\Omega |fg| \, d\mu \qquad (3.25)\\
&\le p^{1/p}\left(\frac{\Gamma^{p-1}-\gamma^{p-1}}{\Gamma-\gamma}\right)^{\frac{1}{p}}\left(\Gamma - \frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu}\right)^{\frac{1}{p}}\left(\frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu} - \gamma\right)^{\frac{1}{p}}\int_\Omega |g|^q \, d\mu\\
&\le \frac{1}{4^{1/p}}p^{1/p}(\Gamma-\gamma)^{1/p}\left(\Gamma^{p-1}-\gamma^{p-1}\right)^{1/p}\int_\Omega |g|^q \, d\mu
\end{aligned}
$$

where $p > 1$ *and* $\frac{1}{p}+\frac{1}{q}=1$.

Proof By multiplying in (3.23) with $\left(\int_\Omega |g|^q \, d\mu\right)^p$ we have

$$
\begin{aligned}
&\int_\Omega |f|^p \, d\mu \left(\int_\Omega |g|^q \, d\mu\right)^{p-1} - \left(\int_\Omega |fg| \, d\mu\right)^p\\
&\le p\frac{\Gamma^{p-1}-\gamma^{p-1}}{\Gamma-\gamma}\left(\Gamma - \frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu}\right)\left(\frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu} - \gamma\right)\left(\int_\Omega |g|^q \, d\mu\right)^p\\
&\le \frac{1}{4}p(\Gamma-\gamma)\left(\Gamma^{p-1}-\gamma^{p-1}\right)\left(\int_\Omega |g|^q \, d\mu\right)^p,
\end{aligned}
$$

which is equivalent with

$$
\begin{aligned}
&\int_\Omega |f|^p \, d\mu \left(\int_\Omega |g|^q \, d\mu\right)^{p-1} \qquad (3.26)\\
&\le \left(\int_\Omega |fg| \, d\mu\right)^p + p\left(\Gamma - \frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu}\right)\left(\frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu} - \gamma\right)
\end{aligned}
$$

$$\times\left(\int_\Omega |g|^q\,d\mu\right)^p \frac{\Gamma^{p-1}-\gamma^{p-1}}{\Gamma-\gamma}$$
$$\le\left(\int_\Omega |fg|\,d\mu\right)^p+\frac{1}{4}p\,(\Gamma-\gamma)\left(\Gamma^{p-1}-\gamma^{p-1}\right)\left(\int_\Omega |g|^q\,d\mu\right)^p.$$

Taking the power $1/p$ with $p>1$ and employing the following elementary inequality that state that for $p>1$ and $\alpha,\beta>0$,

$$(\alpha+\beta)^{1/p}\le\alpha^{1/p}+\beta^{1/p}$$

we have from the first part of (3.26) that

$$\int_\Omega |f|^p\,d\mu\left(\int_\Omega |g|^q\,d\mu\right)^{1-\frac{1}{p}} \tag{3.27}$$
$$\le\int_\Omega |fg|\,d\mu$$
$$+\left[p\left(\Gamma-\frac{\int_\Omega |fg|\,d\mu}{\int_\Omega |g|^q\,d\mu}\right)\left(\frac{\int_\Omega |fg|\,d\mu}{\int_\Omega |g|^q\,d\mu}-\gamma\right)\left(\int_\Omega |g|^q\,d\mu\right)^p\frac{\Gamma^{p-1}-\gamma^{p-1}}{\Gamma-\gamma}\right]^{1/p}.$$

Since $1-\frac{1}{p}=\frac{1}{q}$, we get from (3.27) the first inequality in (3.25). The rest is obvious. □

If $h:\Omega\to\mathbb{R}$ is μ-measurable, satisfies the bounds

$$-\infty<m\le|h(x)|\le M<\infty\text{ for }\mu\text{-a.e. }x\in\Omega$$

and is such that $h,\ |h|^p\in L_w(\Omega,\mu)$, for a μ-measurable function $w:\Omega\to\mathbb{R}$, with $w(x)\ge0$ for μ-a.e. $x\in\Omega$ and $\int_\Omega w\,d\mu>0$, then from Theorem 5 we have amongst other the following inequality

$$0\le\frac{\int_\Omega |h|^p\,w\,d\mu}{\int_\Omega w\,d\mu}-\left(\frac{\int_\Omega |h|\,w\,d\mu}{\int_\Omega w\,d\mu}\right)^p \tag{3.28}$$
$$\le\left(M^p-m^p\right)\left[\frac{1}{2}+\frac{1}{M-m}\left|\frac{\int_\Omega |h|\,w\,d\mu}{\int_\Omega w\,d\mu}-\frac{m+M}{2}\right|\right].$$

From this inequality we can state that:

Proposition 5 (Dragomir [21]) *With the assumptions of Proposition 4 we have*

$$0\le\frac{\int_\Omega |f|^p\,d\mu}{\int_\Omega |g|^q\,d\mu}-\left(\frac{\int_\Omega |fg|\,d\mu}{\int_\Omega |g|^q\,d\mu}\right)^p \tag{3.29}$$
$$\le\left(\Gamma^p-\gamma^p\right)\left[\frac{1}{2}+\frac{1}{\Gamma-\gamma}\left|\frac{\int_\Omega |fg|\,d\mu}{\int_\Omega |g|^q\,d\mu}-\frac{\gamma+\Gamma}{2}\right|\right].$$

Finally, the following additive reverse of the Hölder inequality can be stated as well:

Corollary 7 *With the assumptions of Proposition 4 we have*

$$\left(\int_{\Omega}|f|^{p}\,d\mu\right)^{1/p}\left(\int_{\Omega}|g|^{q}\,d\mu\right)^{1/q}-\int_{\Omega}|fg|\,d\mu \tag{3.30}$$
$$\leq\left(\Gamma^{p}-\gamma^{p}\right)^{1/p}\left[\frac{1}{2}+\frac{1}{\Gamma-\gamma}\left|\frac{\int_{\Omega}|fg|\,d\mu}{\int_{\Omega}|g|^{q}\,d\mu}-\frac{\gamma+\Gamma}{2}\right|\right]^{1/p}\int_{\Omega}|g|^{q}\,d\mu.$$

Remark 13 We observe that for $p=q=2$ we have from the first inequality in (3.29) the following reverse of the Cauchy-Bunyakovsky-Schwarz inequality

$$\int_{\Omega}|g|^{2}\,d\mu\int_{\Omega}|f|^{2}\,d\mu-\left(\int_{\Omega}|fg|\,d\mu\right)^{2} \tag{3.31}$$
$$\leq\left(\Gamma^{2}-\gamma^{2}\right)\left[\frac{1}{2}+\frac{1}{\Gamma-\gamma}\left|\frac{\int_{\Omega}|fg|\,d\mu}{\int_{\Omega}|g|^{2}\,d\mu}-\frac{\gamma+\Gamma}{2}\right|\right]\left(\int_{\Omega}|g|^{2}\,d\mu\right)^{2}$$

provided that $f,\ g\in L_{2}(\Omega,\mu)$ and there exists the constants $\gamma,\ \Gamma>0$ such that

$$\gamma\leq\frac{|f|}{|g|}\leq\Gamma\ \mu\text{-a.e. on }\Omega.$$

One can easily observe that the bound provided by (3.31) is not as good as the one given by (3.24). The details are omitted.

3.3 Applications for f-Divergence

The following result holds:

Proposition 6 (Dragomir [21]) *Let $f:(0,\infty)\to\mathbb{R}$ be a convex function with the property that $f(1)=0$. Assume that $p,\ q\in\mathcal{P}$ and there exists the constants $0<r<1<R<\infty$ such that*

$$r\leq\frac{q(x)}{p(x)}\leq R\ \textit{for}\ \mu\textit{-a.e.}\ x\in\Omega. \tag{3.32}$$

Then we have the inequalities

$$0\leq I_{f}(p,q)\leq(R-1)(1-r)\,\frac{f'_{-}(R)-f'_{+}(r)}{R-r} \tag{3.33}$$
$$\leq\frac{1}{4}(R-r)\left[f'_{-}(R)-f'_{+}(r)\right].$$

Proof Utilising Theorem 4 we can write that

$$\begin{aligned}&\int_\Omega p(x) f\left(\frac{q(x)}{p(x)}\right) d\mu(x) - f\left(\int_\Omega q(x)\, d\mu(x)\right) \qquad (3.34)\\ &\le \left(R - \int_\Omega q(x)\, d\mu(x)\right)\left(\int_\Omega q(x)\, d\mu(x) - r\right)\frac{f'_-(R) - f'_+(r)}{R-r}\\ &\le \frac{1}{4}(R-r)\left[f'_-(R) - f'_+(r)\right],\end{aligned}$$

for $p, q \in \mathcal{P}$ satisfying (3.32) and since $f\left(\int_\Omega q(x)\, d\mu(x)\right) = f(1) = 0$ we get from (3.34) the desired result (3.33). □

By the use of Theorem 5 we can also state the following result:

Proposition 7 (Dragomir [21]) *With the assumptions in Proposition 6, we have the inequalities*

$$0 \le I_f(p, q) \le B_f(r, R) \qquad (3.35)$$

where

$$B_f(r, R) := \frac{(R-1)\int_r^1 |f'(t)|\, dt + (1-r)\int_1^R |f'(t)|\, dt}{R-r}. \qquad (3.36)$$

Moreover, we have the following bounds for $B_f(r, R)$,

$$\begin{aligned}&B_f(r, R) \qquad (3.37)\\ &\le \begin{cases}\left[\frac{1}{2} + \frac{\left|1-\frac{r+R}{2}\right|}{R-r}\right]\int_r^R |f'(t)|\, dt,\\ \left[\frac{1}{2}\int_r^R |f'(t)|\, dt + \frac{1}{2}\left|\int_1^R |f'(t)|\, dt - \int_r^1 |f'(t)|\, dt\right|\right]\end{cases}\end{aligned}$$

and

$$\begin{aligned}&B_f(r, R) \qquad (3.38)\\ &\le \frac{(1-r)(R-1)}{R-r}\left[\|f'\|_{[1,R],\infty} + \|f'\|_{[r,1],\infty}\right]\\ &\le \frac{1}{2}(R-r)\frac{\|f'\|_{[1,R],\infty} + \|f'\|_{[r,1],\infty}}{2} \le \frac{1}{2}(R-r)\|f'\|_{[r,R],\infty}\end{aligned}$$

and

$$
\begin{aligned}
& B_f\,(r, R) \qquad (3.39)\\
& \le \frac{1}{R-r}\left[(1-r)\,(R-1)^{1/q}\left\|f'\right\|_{[1,R],p} + (R-1)\,(1-r)^{1/q}\left\|f'\right\|_{[r,1],p}\right]\\
& \le \frac{1}{R-r}\left[(1-r)^q\,(R-1) + (R-1)^q\,(1-r)\right]^{1/q}\left\|f'\right\|_{[r,R],p}
\end{aligned}
$$

where $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$.

The above results can be utilized to obtain various inequalities for the divergence measures in information theory that are particular instances of f-divergences.

Consider, for example, the Kullback-Leibler divergence measure

$$D_{KL}\,(p, q) := \int_\Omega p\,(x)\ln\left[\frac{p\,(x)}{q\,(x)}\right] d\mu\,(x)\,, \quad p,\ q \in \mathcal{P},$$

which is an f-divergence for the convex function $f : (0, \infty) \to \mathbb{R}$, $f\,(t) = -\ln t$.

If $p, q \in \mathcal{P}$ such that there exists the constants $0 < r < 1 < R < \infty$ with

$$r \le \frac{q\,(x)}{p\,(x)} \le R \text{ for } \mu\text{-a.e. } x \in \Omega, \qquad (3.40)$$

then we get from (3.33) that

$$D_{KL}\,(p, q) \le \frac{(R-1)\,(1-r)}{rR} \qquad (3.41)$$

and from (3.35) that

$$D_{KL}\,(p, q) \le \ln\left(\frac{R^{1-r}}{r^{R-1}}\right)^{\frac{1}{R-r}}.$$

The interested reader can obtain similar results for other divergence measures as listed above. However, the details are omitted.

4 More Reverse Inequalities

4.1 General Results

The following reverse of the Jensen's inequality that provides a refinement and an alternative for the inequality in Theorem 4 holds:

Theorem 6 (Dragomir [20]) *Let $\Phi : I \to \mathbb{R}$ be a continuous convex function on the interval of real numbers I and $m, M \in \mathbb{R}$, $m < M$ with $[m, M] \subset \mathring{I}$, $\mathring{I}$ is the interior of I. If $f : \Omega \to \mathbb{R}$ is μ-measurable, satisfies the bounds*

$$-\infty < m \leq f(x) \leq M < \infty \, \textit{for } \mu\textit{-a.e. } x \in \Omega$$

and such that $f, \Phi \circ f \in L_w(\Omega, \mu)$, where $w \geq 0$ μ-a.e. on Ω with $\int_\Omega w d\mu = 1$, then

$$\begin{aligned} 0 &\leq \int_\Omega w (\Phi \circ f) \, d\mu - \Phi\left(\bar{f}_{\Omega,w}\right) \\ &\leq \frac{\left(M - \bar{f}_{\Omega,w}\right)\left(\bar{f}_{\Omega,w} - m\right)}{M - m} \sup_{t \in (m,M)} \Psi_\Phi(t; m, M) \\ &\leq \left(M - \bar{f}_{\Omega,w}\right)\left(\bar{f}_{\Omega,w} - m\right) \frac{\Phi'_-(M) - \Phi'_+(m)}{M - m} \\ &\leq \frac{1}{4}(M - m)\left[\Phi'_-(M) - \Phi'_+(m)\right], \end{aligned} \tag{4.1}$$

where $\bar{f}_{\Omega,w} := \int_\Omega w(x) f(x) \, d\mu(x) \in [m, M]$ and $\Psi_\Phi(\cdot; m, M) : (m, M) \to \mathbb{R}$ is defined by

$$\Psi_\Phi(t; m, M) = \frac{\Phi(M) - \Phi(t)}{M - t} - \frac{\Phi(t) - \Phi(m)}{t - m}.$$

We also have the inequality

$$\begin{aligned} 0 \leq \int_\Omega w (\Phi \circ f) \, d\mu - \Phi\left(\bar{f}_{\Omega,w}\right) &\leq \frac{1}{4}(M - m) \Psi_\Phi\left(\bar{f}_{\Omega,w}; m, M\right) \\ \leq \frac{1}{4}(M - m)\left[\Phi'_-(M) - \Phi'_+(m)\right], \end{aligned} \tag{4.2}$$

provided that $\bar{f}_{\Omega,w} \in (m, M)$.

Proof By the convexity of Φ we have that

$$\begin{aligned} &\int_\Omega w(x) \Phi(f(x)) \, d\mu(x) - \Phi\left(\bar{f}_{\Omega,w}\right) \\ &= \int_\Omega w(x) \Phi\left[\frac{m(M - f(x)) + M(f(x) - m)}{M - m}\right] d\mu(x) \\ &- \Phi\left(\int_\Omega w(x)\left[\frac{m(M - f(x)) + M(f(x) - m)}{M - m}\right] d\mu(x)\right) \end{aligned} \tag{4.3}$$

$$\begin{aligned}&\leq \int_{\Omega} \frac{(M-f(x))\,\Phi(m)+(f(x)-m)\,\Phi(M)}{M-m} w(x)\,d\mu(x)\\&-\Phi\left(\frac{m\left(M-\bar{f}_{\Omega,w}\right)+M\left(\bar{f}_{\Omega,w}-m\right)}{M-m}\right)\\&=\frac{\left(M-\bar{f}_{\Omega,w}\right)\Phi(m)+\left(\bar{f}_{\Omega,w}-m\right)\Phi(M)}{M-m}\\&-\Phi\left(\frac{m\left(M-\bar{f}_{\Omega,w}\right)+M\left(\bar{f}_{\Omega,w}-m\right)}{M-m}\right):=B.\end{aligned}$$

By denoting

$$\Delta_{\Phi}(t;m,M):=\frac{(t-m)\,\Phi(M)+(M-t)\,\Phi(m)}{M-m}-\Phi(t)\,,\quad t\in[m,M]$$

we have

$$\begin{aligned}\Delta_{\Phi}(t;m,M)&=\frac{(t-m)\,\Phi(M)+(M-t)\,\Phi(m)-(M-m)\,\Phi(t)}{M-m}\qquad(4.4)\\&=\frac{(t-m)\,\Phi(M)+(M-t)\,\Phi(m)-(M-t+t-m)\,\Phi(t)}{M-m}\\&=\frac{(t-m)\,[\Phi(M)-\Phi(t)]-(M-t)\,[\Phi(t)-\Phi(m)]}{M-m}\\&=\frac{(M-t)\,(t-m)}{M-m}\Psi_{\Phi}(t;m,M)\end{aligned}$$

for any $t\in(m,M)$.

Therefore we have the equality

$$B=\frac{\left(M-\bar{f}_{\Omega,w}\right)\left(\bar{f}_{\Omega,w}-m\right)}{M-m}\Psi_{\Phi}\left(\bar{f}_{\Omega,w};m,M\right)\qquad(4.5)$$

provided that $\bar{f}_{\Omega,w}\in(m,M)$.

For $\bar{f}_{\Omega,w}=m$ or $\bar{f}_{\Omega,w}=M$ the inequality (4.1) is obvious. If $\bar{f}_{\Omega,w}\in(m,M)$, then

$$\begin{aligned}\Psi_{\Phi}\left(\bar{f}_{\Omega,w};m,M\right)&\leq\sup_{t\in(m,M)}\Psi_{\Phi}(t;m,M)\\&=\sup_{t\in(m,M)}\left[\frac{\Phi(M)-\Phi(t)}{M-t}-\frac{\Phi(t)-\Phi(m)}{t-m}\right]\end{aligned}$$

$$\begin{aligned}
&\leq \sup_{t\in(m,M)}\left[\frac{\Phi(M)-\Phi(t)}{M-t}\right]+\sup_{t\in(m,M)}\left[-\frac{\Phi(t)-\Phi(m)}{t-m}\right]\\
&=\sup_{t\in(m,M)}\left[\frac{\Phi(M)-\Phi(t)}{M-t}\right]-\inf_{t\in(m,M)}\left[\frac{\Phi(t)-\Phi(m)}{t-m}\right]\\
&=\Phi'_-(M)-\Phi'_+(m),
\end{aligned}$$

which by (4.3) and (4.5) produces the desired result (4.1).

Since, obviously

$$\frac{\left(M-\bar{f}_{\Omega,w}\right)\left(\bar{f}_{\Omega,w}-m\right)}{M-m}\leq\frac{1}{4}(M-m),$$

then by (4.3) and (4.5) we deduce the first inequality (4.2). The second part is clear. □

Corollary 8 *Let $\Phi : I \to \mathbb{R}$ be a continuous convex function on the interval of real numbers I and $m, M \in \mathbb{R}$, $m < M$ with $[m, M] \subset \mathring{I}$. If $x_i \in [m, M]$ and $p_i \geq 0$ for $i \in \{1,\ldots,n\}$ with $\sum_{i=1}^n p_i = 1$, then we have the inequalities*

$$\begin{aligned}
0&\leq\sum_{i=1}^n p_i\Phi(x_i)-\Phi\left(\bar{x}_p\right) \qquad (4.6)\\
&\leq\frac{\left(M-\bar{x}_p\right)\left(\bar{x}_p-m\right)}{M-m}\sup_{t\in(m,M)}\Psi_\Phi(t;m,M)\\
&\leq\left(M-\bar{x}_p\right)\left(\bar{x}_p-m\right)\frac{\Phi'_-(M)-\Phi'_+(m)}{M-m}\\
&\leq\frac{1}{4}(M-m)\left[\Phi'_-(M)-\Phi'_+(m)\right],
\end{aligned}$$

and

$$\begin{aligned}
0&\leq\sum_{i=1}^n p_i\Phi(x_i)-\Phi\left(\bar{x}_p\right)\leq\frac{1}{4}(M-m)\Psi_\Phi\left(\bar{x}_p;m,M\right) \qquad (4.7)\\
&\leq\frac{1}{4}(M-m)\left[\Phi'_-(M)-\Phi'_+(m)\right],
\end{aligned}$$

where $\bar{x}_p := \sum_{i=1}^n p_i x_i \in (m, M)$.

Remark 14 Consider the positive n-tuple $x = (x_1, \ldots, x_n)$ with the nonnegative weights $w = (w_1, \ldots, w_n)$ where $W_n := \sum_{i=1}^n w_i > 0$. Applying the inequality between the first and third term in (4.6) for the convex function $\Phi(t) = -\ln t$, $t > 0$ we have

$$
\begin{aligned}
1 &\le \frac{A_n(w,x)}{G_n(w,x)} \le \exp\left[\frac{1}{Mm}(M - A_n(w,x))(A_n(w,x) - m)\right] \qquad (4.8)\\
&\le \exp\left[\frac{1}{4}\frac{(M-m)^2}{mM}\right],
\end{aligned}
$$

provided that $0 < m \le x_i \le M < \infty$ for $i \in \{1, \ldots, n\}$.

Also, if we apply the inequality (4.7) for the same function Φ we get that

$$
\begin{aligned}
1 &\le \frac{A_n(w,x)}{G_n(w,x)} \qquad (4.9)\\
&\le \left[\left(\frac{M}{A_n(w,x)}\right)^{M - A_n(w,x)} \left(\frac{m}{A_n(w,x)}\right)^{A_n(w,x) - m}\right]^{-\frac{1}{4}(M-m)}\\
&\le \exp\left[\frac{1}{4}\frac{(M-m)^2}{mM}\right].
\end{aligned}
$$

The following result also holds:

Theorem 7 (Dragomir [20]) *With the assumptions of Theorem 6, we have the inequalities*

$$
\begin{aligned}
0 &\le \int_\Omega w(\Phi \circ f)\, d\mu(x) - \Phi\left(\bar{f}_{\Omega,w}\right) \qquad (4.10)\\
&\le 2\max\left\{\frac{M - \bar{f}_{\Omega,w}}{M-m}, \frac{\bar{f}_{\Omega,w} - m}{M-m}\right\}\left[\frac{\Phi(m) + \Phi(M)}{2} - \Phi\left(\frac{m+M}{2}\right)\right]\\
&\le \frac{1}{2}\max\left\{M - \bar{f}_{\Omega,w}, \bar{f}_{\Omega,w} - m\right\}\left[\Phi'_-(M) - \Phi'_+(m)\right].
\end{aligned}
$$

Proof First of all, we recall the following result obtained by the author in [16] that provides a refinement and a reverse for the weighted Jensen's discrete inequality:

$$
\begin{aligned}
& n \min_{i \in \{1,\ldots,n\}} \{p_i\}\left[\frac{1}{n}\sum_{i=1}^n \Phi(x_i) - \Phi\left(\frac{1}{n}\sum_{i=1}^n x_i\right)\right] \qquad (4.11)\\
&\le \frac{1}{P_n}\sum_{i=1}^n p_i \Phi(x_i) - \Phi\left(\frac{1}{P_n}\sum_{i=1}^n p_i x_i\right)\\
& n \max_{i \in \{1,\ldots,n\}} \{p_i\}\left[\frac{1}{n}\sum_{i=1}^n \Phi(x_i) - \Phi\left(\frac{1}{n}\sum_{i=1}^n x_i\right)\right],
\end{aligned}
$$

where $\Phi : C \to \mathbb{R}$ is a convex function defined on the convex subset C of the linear space X, $\{x_i\}_{i\in\{1,\dots,n\}} \subset C$ are vectors and $\{p_i\}_{i\in\{1,\dots,n\}}$ are nonnegative numbers with $P_n := \sum_{i=1}^{n} p_i > 0$.

For $n = 2$ we deduce from (4.11) that

$$\begin{aligned}
&2\min\{t, 1-t\}\left[\frac{\Phi(x)+\Phi(y)}{2} - \Phi\left(\frac{x+y}{2}\right)\right] \qquad (4.12)\\
&\le t\Phi(x) + (1-t)\Phi(y) - \Phi(tx+(1-t)y)\\
&\le 2\max\{t, 1-t\}\left[\frac{\Phi(x)+\Phi(y)}{2} - \Phi\left(\frac{x+y}{2}\right)\right]
\end{aligned}$$

for any $x, y \in C$ and $t \in [0,1]$.

If we use the second inequality in (4.12) for the convex function $\Phi : I \to \mathbb{R}$ and $m, M \in \mathbb{R}, m < M$ with $[m, M] \subset \mathring{I}$, we have for $t = \frac{M-\bar{f}_{\Omega,w}}{M-m}$ that

$$\begin{aligned}
&\frac{\left(M-\bar{f}_{\Omega,w}\right)\Phi(m) + \left(\bar{f}_{\Omega,w}-m\right)\Phi(M)}{M-m} \qquad (4.13)\\
&-\Phi\left(\frac{m\left(M-\bar{f}_{\Omega,w}\right) + M\left(\bar{f}_{\Omega,w}-m\right)}{M-m}\right)\\
&\le 2\max\left\{\frac{M-\bar{f}_{\Omega,w}}{M-m}, \frac{\bar{f}_{\Omega,w}-m}{M-m}\right\}\\
&\times\left[\frac{\Phi(m)+\Phi(M)}{2} - \Phi\left(\frac{m+M}{2}\right)\right].
\end{aligned}$$

Utilizing the inequality (4.3) and (4.13) we deduce the first inequality in (4.10).

Since

$$\begin{aligned}
&\frac{\frac{\Phi(m)+\Phi(M)}{2} - \Phi\left(\frac{m+M}{2}\right)}{M-m}\\
&= \frac{1}{4}\left[\frac{\Phi(M)-\Phi\left(\frac{m+M}{2}\right)}{M-\frac{m+M}{2}} - \frac{\Phi\left(\frac{m+M}{2}\right)-\Phi(m)}{\frac{m+M}{2}-m}\right]
\end{aligned}$$

and, by the gradient inequality, we have that

$$\frac{\Phi(M)-\Phi\left(\frac{m+M}{2}\right)}{M-\frac{m+M}{2}} \le \Phi'_{-}(M)$$

and

$$\frac{\Phi\left(\frac{m+M}{2}\right)-\Phi(m)}{\frac{m+M}{2}-m} \ge \Phi'_{+}(m),$$

then we get

$$\frac{\frac{\Phi(m)+\Phi(M)}{2}-\Phi\left(\frac{m+M}{2}\right)}{M-m}\leq\frac{1}{4}\left[\Phi'_{-}(M)-\Phi'_{+}(m)\right]. \tag{4.14}$$

On making use of (4.13) and (4.14) we deduce the last part of (4.10). □

Corollary 9 *With the assumptions in Corollary 8, we have the inequalities*

$$\begin{aligned}0&\leq\sum_{i=1}^{n}p_i\Phi(x_i)-\Phi(\bar{x}_p) \qquad (4.15)\\&\leq 2\max\left\{\frac{M-\bar{x}_p}{M-m},\frac{\bar{x}_p-m}{M-m}\right\}\left[\frac{\Phi(m)+\Phi(M)}{2}-\Phi\left(\frac{m+M}{2}\right)\right]\\&\leq\frac{1}{2}\max\left\{M-\bar{x}_p,\bar{x}_p-m\right\}\left[\Phi'_{-}(M)-\Phi'_{+}(m)\right].\end{aligned}$$

Remark 15 Since, obviously,

$$\frac{M-\bar{f}_{\Omega,w}}{M-m},\frac{\bar{f}_{\Omega,w}-m}{M-m}\leq 1$$

then we obtain from the first inequality in (4.10) the simpler, however coarser inequality

$$\begin{aligned}0&\leq\int_{\Omega}w(\Phi\circ f)\,d\mu(x)-\Phi\left(\bar{f}_{\Omega,w}\right) \qquad (4.16)\\&\leq 2\left[\frac{\Phi(m)+\Phi(M)}{2}-\Phi\left(\frac{m+M}{2}\right)\right].\end{aligned}$$

We notice that the discrete version of this result, namely

$$0\leq\sum_{i=1}^{n}p_i\Phi(x_i)-\Phi(\bar{x}_p)\leq 2\left[\frac{\Phi(m)+\Phi(M)}{2}-\Phi\left(\frac{m+M}{2}\right)\right] \tag{4.17}$$

was obtained in 2008 by Simić in [49].

Remark 16 With the assumptions in Remark 14 we have the following reverse of the arithmetic mean-geometric mean inequality

$$1\leq\frac{A_n(w,x)}{G_n(w,x)}\leq\left(\frac{A(m,M)}{G(m,M)}\right)^{2\max\left\{\frac{M-A_n(w,x)}{M-m},\frac{A_n(w,x)-m}{M-m}\right\}}, \tag{4.18}$$

where $A(m,M)$ is the arithmetic mean while $G(m,M)$ is the geometric mean of the positive numbers m and M.

4.2 Applications for the Hölder Inequality

Assume that $p > 1$. If $h : \Omega \to \mathbb{R}$ is μ-measurable, satisfies the bounds

$$0 < m \le |h(x)| \le M < \infty \text{ for } \mu\text{-a.e. } x \in \Omega$$

and is such that $h, |h|^p \in L_w(\Omega, \mu)$, for a μ-measurable function $w : \Omega \to \mathbb{R}$, with $w(x) \ge 0$ for μ-a.e. $x \in \Omega$ and $\int_\Omega w d\mu > 0$, then from (4.1) we have

$$\begin{aligned} 0 &\le \frac{\int_\Omega |h|^p w d\mu}{\int_\Omega w d\mu} - \left(\frac{\int_\Omega |h| w d\mu}{\int_\Omega w d\mu}\right)^p \\ &\le \frac{\left(M - \overline{|h|}_{\Omega,w}\right)\left(\overline{|h|}_{\Omega,w} - m\right)}{M - m} B_p(m, M) \\ &\le p \frac{M^{p-1} - m^{p-1}}{M - m}\left(M - \overline{|h|}_{\Omega,w}\right)\left(\overline{|h|}_{\Omega,w} - m\right) \\ &\le \frac{1}{4} p (M - m)\left(M^{p-1} - m^{p-1}\right), \end{aligned} \tag{4.19}$$

where $\overline{|h|}_{\Omega,w} := \frac{\int_\Omega |h| w d\mu}{\int_\Omega w d\mu} \in [m, M]$ and $\Psi_p(\cdot; m, M) : (m, M) \to \mathbb{R}$ is defined by

$$\Psi_p(t; m, M) = \frac{M^p - t^p}{M - t} - \frac{t^p - m^p}{t - m}$$

while

$$B_p(m, M) := \sup_{t \in (m,M)} \Psi_p(t; m, M). \tag{4.20}$$

From (4.2) we also have the inequality

$$\begin{aligned} 0 &\le \frac{\int_\Omega |h|^p w d\mu}{\int_\Omega w d\mu} - \left(\frac{\int_\Omega |h| w d\mu}{\int_\Omega w d\mu}\right)^p \le \frac{1}{4}(M - m)\Psi_p\left(\overline{|h|}_{\Omega,w}; m, M\right) \\ &\le \frac{1}{4} p (M - m)\left(M^{p-1} - m^{p-1}\right). \end{aligned} \tag{4.21}$$

Proposition 8 (Dragomir [20]) *If $f \in L_p(\Omega, \mu)$, $g \in L_q(\Omega, \mu)$ with $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$ and there exists the constants $\gamma, \Gamma > 0$ and such that*

$$\gamma \le \frac{|f|}{|g|^{q-1}} \le \Gamma \ \mu\text{-a.e. on } \Omega,$$

then we have

$$\begin{aligned}0 &\le \frac{\int_\Omega |f|^p\,d\mu}{\int_\Omega |g|^q\,d\mu} - \left(\frac{\int_\Omega |fg|\,d\mu}{\int_\Omega |g|^q\,d\mu}\right)^p \qquad (4.22)\\ &\le \frac{B_p(\gamma,\Gamma)}{\Gamma-\gamma}\left(\Gamma - \frac{\int_\Omega |fg|\,d\mu}{\int_\Omega |g|^q\,d\mu}\right)\left(\frac{\int_\Omega |fg|\,d\mu}{\int_\Omega |g|^q\,d\mu} - \gamma\right)\\ &\le p\frac{\Gamma^{p-1}-\gamma^{p-1}}{\Gamma-\gamma}\left(\Gamma - \frac{\int_\Omega |fg|\,d\mu}{\int_\Omega |g|^q\,d\mu}\right)\left(\frac{\int_\Omega |fg|\,d\mu}{\int_\Omega |g|^q\,d\mu} - \gamma\right)\\ &\le \frac{1}{4}p(\Gamma-\gamma)\left(\Gamma^{p-1}-\gamma^{p-1}\right),\end{aligned}$$

and

$$\begin{aligned}0 &\le \frac{\int_\Omega |f|^p\,d\mu}{\int_\Omega |g|^q\,d\mu} - \left(\frac{\int_\Omega |fg|\,d\mu}{\int_\Omega |g|^q\,d\mu}\right)^p \qquad (4.23)\\ &\le \frac{1}{4}(\Gamma-\gamma)\Psi_p\left(\frac{\int_\Omega |fg|\,d\mu}{\int_\Omega |g|^q\,d\mu};\gamma,\Gamma\right) \le \frac{1}{4}p(\Gamma-\gamma)\left(\Gamma^{p-1}-\gamma^{p-1}\right),\end{aligned}$$

where $B_p(\cdot,\cdot)$ and $\Psi_p(\cdot;\cdot,\cdot)$ are defined above.

Proof The inequalities (4.22) and (4.23) follow from (4.19) and (4.21) by choosing

$$h = \frac{|f|}{|g|^{q-1}} \text{ and } w = |g|^q .$$

The details are omitted. □

Remark 17 We observe that for $p = q = 2$ we have $\Psi_2(t;\gamma,\Gamma) = \Gamma-\gamma = B_2(\gamma,\Gamma)$ and then from the first inequality in (4.22) we get the following reverse of the Cauchy-Bunyakovsky-Schwarz inequality:

$$\begin{aligned}&\int_\Omega |g|^2\,d\mu \int_\Omega |f|^2\,d\mu - \left(\int_\Omega |fg|\,d\mu\right)^2 \qquad (4.24)\\ &\le \left(\Gamma - \frac{\int_\Omega |fg|\,d\mu}{\int_\Omega |g|^2\,d\mu}\right)\left(\frac{\int_\Omega |fg|\,d\mu}{\int_\Omega |g|^2\,d\mu} - \gamma\right)\left(\int_\Omega |g|^2\,d\mu\right)^2\end{aligned}$$

provided that $f, g \in L_2(\Omega,\mu)$, and there exists the constants $\gamma, \Gamma > 0$ such that

$$\gamma \le \frac{|f|}{|g|} \le \Gamma \ \mu\text{-a.e. on } \Omega.$$

Corollary 10 *With the assumptions of Proposition 8 we have the following additive reverses of the Hölder inequality*

$$
\begin{aligned}
0 &\le \left(\int_\Omega |f|^p \, d\mu\right)^{1/p} \left(\int_\Omega |g|^q \, d\mu\right)^{1/q} - \int_\Omega |fg| \, d\mu \qquad (4.25)\\
&\le \left[\frac{B_p(\gamma,\Gamma)}{\Gamma-\gamma}\right]^{1/p} \left(\Gamma - \frac{\int_\Omega |fg|\, d\mu}{\int_\Omega |g|^q \, d\mu}\right)^{1/p} \left(\frac{\int_\Omega |fg|\, d\mu}{\int_\Omega |g|^q \, d\mu} - \gamma\right)^{1/p}\\
&\times \int_\Omega |g|^q \, d\mu\\
&\le p^{1/p} \left(\frac{\Gamma^{p-1}-\gamma^{p-1}}{\Gamma-\gamma}\right)^{1/p} \left(\Gamma - \frac{\int_\Omega |fg|\, d\mu}{\int_\Omega |g|^q \, d\mu}\right)^{1/p} \left(\frac{\int_\Omega |fg|\, d\mu}{\int_\Omega |g|^q \, d\mu} - \gamma\right)^{1/p}\\
&\times \int_\Omega |g|^q \, d\mu\\
&\le \frac{1}{4^{1/p}} p^{1/p} (\Gamma-\gamma)^{1/p} \left(\Gamma^{p-1}-\gamma^{p-1}\right)^{1/p} \int_\Omega |g|^q \, d\mu
\end{aligned}
$$

and

$$
\begin{aligned}
0 &\le \left(\int_\Omega |f|^p \, d\mu\right)^{1/p} \left(\int_\Omega |g|^q \, d\mu\right)^{1/q} - \int_\Omega |fg| \, d\mu \qquad (4.26)\\
&\le \frac{1}{4^{1/p}} (\Gamma-\gamma)^{1/p} \Psi_p^{1/p} \left(\frac{\int_\Omega |fg|\, d\mu}{\int_\Omega |g|^q \, d\mu}; m, M\right) \int_\Omega |g|^q \, d\mu\\
&\le \frac{1}{4^{1/p}} p^{1/p} (\Gamma-\gamma)^{1/p} \left(\Gamma^{p-1}-\gamma^{p-1}\right)^{1/p} \int_\Omega |g|^q \, d\mu
\end{aligned}
$$

where $p > 1$ *and* $\frac{1}{p} + \frac{1}{q} = 1$.

Proof By multiplying in (4.22) with $\left(\int_\Omega |g|^q \, d\mu\right)^p$ we have

$$
\begin{aligned}
&\int_\Omega |f|^p \, d\mu \left(\int_\Omega |g|^q \, d\mu\right)^{p-1} - \left(\int_\Omega |fg| \, d\mu\right)^p\\
&\le \frac{B_p(\gamma,\Gamma)}{\Gamma-\gamma} \left(\Gamma - \frac{\int_\Omega |fg|\, d\mu}{\int_\Omega |g|^q \, d\mu}\right) \left(\frac{\int_\Omega |fg|\, d\mu}{\int_\Omega |g|^q \, d\mu} - \gamma\right) \left(\int_\Omega |g|^q \, d\mu\right)^p\\
&\le p \frac{\Gamma^{p-1}-\gamma^{p-1}}{\Gamma-\gamma} \left(\Gamma - \frac{\int_\Omega |fg|\, d\mu}{\int_\Omega |g|^q \, d\mu}\right) \left(\frac{\int_\Omega |fg|\, d\mu}{\int_\Omega |g|^q \, d\mu} - \gamma\right) \left(\int_\Omega |g|^q \, d\mu\right)^p\\
&\le \frac{1}{4} p (\Gamma-\gamma) \left(\Gamma^{p-1}-\gamma^{p-1}\right) \left(\int_\Omega |g|^q \, d\mu\right)^p,
\end{aligned}
$$

which is equivalent with

$$\begin{aligned}
&\int_\Omega |f|^p \, d\mu \left(\int_\Omega |g|^q \, d\mu \right)^{p-1} \qquad (4.27)\\
&\le \left(\int_\Omega |fg| \, d\mu \right)^p + \frac{B_p(\gamma, \Gamma)}{\Gamma - \gamma} \left(\Gamma - \frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu} \right) \left(\frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu} - \gamma \right)\\
&\times \left(\int_\Omega |g|^q \, d\mu \right)^p\\
&\le \left(\int_\Omega |fg| \, d\mu \right)^p + p \left(\Gamma - \frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu} \right) \left(\frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu} - \gamma \right)\\
&\times \left(\int_\Omega |g|^q \, d\mu \right)^p \frac{\Gamma^{p-1} - \gamma^{p-1}}{\Gamma - \gamma}\\
&\le \left(\int_\Omega |fg| \, d\mu \right)^p + \frac{1}{4} p (\Gamma - \gamma) \left(\Gamma^{p-1} - \gamma^{p-1} \right) \left(\int_\Omega |g|^q \, d\mu \right)^p .
\end{aligned}$$

Taking the power $1/p$ with $p > 1$ and employing the following elementary inequality that state that for $p > 1$ and $\alpha, \beta > 0$,

$$(\alpha + \beta)^{1/p} \le \alpha^{1/p} + \beta^{1/p}$$

we have from the first part of (4.27) that

$$\begin{aligned}
&\left(\int_\Omega |f|^p \right)^{1/p} d\mu \left(\int_\Omega |g|^q \, d\mu \right)^{1-\frac{1}{p}} \qquad (4.28)\\
&\le \int_\Omega |fg| \, d\mu + \left[\frac{B_p(\gamma, \Gamma)}{\Gamma - \gamma} \right]^{1/p} \left(\Gamma - \frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu} \right)^{1/p} \left(\frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu} - \gamma \right)^{1/p}\\
&\times \int_\Omega |g|^q \, d\mu
\end{aligned}$$

and since $1 - \frac{1}{p} = \frac{1}{q}$ we get from (4.28) the first inequality in (4.25). The rest is obvious.

The inequality (4.26) can be proved in a similar manner, however the details are omitted. □

If $h : \Omega \to \mathbb{R}$ is μ-measurable, satisfies the bounds

$$0 < m \le |h(x)| \le M < \infty \text{ for } \mu\text{-a.e. } x \in \Omega$$

and is such that $h, |h|^p \in L_w(\Omega, \mu)$, for a μ-measurable function $w : \Omega \to \mathbb{R}$, with $w(x) \ge 0$ for μ-a.e. $x \in \Omega$ and $\int_\Omega w d\mu > 0$, then from (4.10) we also have the inequality

$$0 \le \frac{\int_\Omega |h|^p \, w d\mu}{\int_\Omega w d\mu} - \left(\frac{\int_\Omega |h| \, w d\mu}{\int_\Omega w d\mu}\right)^p \tag{4.29}$$
$$\le 2\left[\frac{m^p + M^p}{2} - \left(\frac{m+M}{2}\right)^p\right] \max\left\{\frac{M - \overline{|h|}_{\Omega,w}}{M-m}, \frac{\overline{|h|}_{\Omega,w} - m}{M-m}\right\}$$
$$\le \frac{1}{2} p \left(M^{p-1} - m^{p-1}\right) \max\left\{M - \overline{|h|}_{\Omega,w}, \overline{|h|}_{\Omega,w} - m\right\}.$$

where, as above, $\overline{|h|}_{\Omega,w} := \frac{\int_\Omega |h| w d\mu}{\int_\Omega w d\mu} \in [m, M]$.

From the inequality (4.29) we can state:

Proposition 9 (Dragomir [20]) *With the assumptions of Proposition 8 we have*

$$0 \le \frac{\int_\Omega |f|^p \, d\mu}{\int_\Omega |g|^q \, d\mu} - \left(\frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu}\right)^p \tag{4.30}$$
$$\le 2 \cdot \frac{\frac{\gamma^p + \Gamma^p}{2} - \left(\frac{\gamma + \Gamma}{2}\right)^p}{\Gamma - \gamma} \max\left\{\Gamma - \frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu}, \frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu} - \gamma\right\}$$
$$\le \frac{1}{2} p \left(\Gamma^{p-1} - \gamma^{p-1}\right) \max\left\{\Gamma - \frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu}, \frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu} - \gamma\right\}.$$

Finally, the following additive reverse of the Hölder inequality can be stated as well:

Corollary 11 *With the assumptions of Proposition 8 we have*

$$0 \le \left(\int_\Omega |f|^p \, d\mu\right)^{1/p} \left(\int_\Omega |g|^q \, d\mu\right)^{1/q} - \int_\Omega |fg| \, d\mu \tag{4.31}$$
$$\le 2^{1/p} \cdot \left(\frac{\frac{\gamma^p + \Gamma^p}{2} - \left(\frac{\gamma + \Gamma}{2}\right)^p}{\Gamma - \gamma}\right)^{1/p}$$
$$\times \max\left\{\left(\Gamma - \frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu}\right)^{1/p}, \left(\frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu} - \gamma\right)^{1/p}\right\} \int_\Omega |g|^q \, d\mu$$
$$\le \frac{1}{2^{1/p}} p^{1/p} \max\left\{\left(\Gamma - \frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu}\right)^{1/p}, \left(\frac{\int_\Omega |fg| \, d\mu}{\int_\Omega |g|^q \, d\mu} - \gamma\right)^{1/p}\right\}$$
$$\times \left(\Gamma^{p-1} - \gamma^{p-1}\right)^{1/p} \int_\Omega |g|^q \, d\mu.$$

Remark 18 As a simpler, however coarser inequality we have the following result:

$$\begin{aligned}0 &\leq \left(\int_\Omega |f|^p \, d\mu\right)^{1/p} \left(\int_\Omega |g|^q \, d\mu\right)^{1/q} - \int_\Omega |fg| \, d\mu \\ &\leq 2^{1/p} \cdot \left[\frac{\gamma^p + \Gamma^p}{2} - \left(\frac{\gamma + \Gamma}{2}\right)^p\right]^{1/p} \int_\Omega |g|^q \, d\mu,\end{aligned}$$

where f and g are as above.

4.3 Applications for f-Divergence

The following result holds:

Proposition 10 (Dragomir [20]) *Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with the property that $f(1) = 0$. Assume that $p, q \in \mathcal{P}$ and there exists the constants $0 < r < 1 < R < \infty$ such that*

$$r \leq \frac{q(x)}{p(x)} \leq R \text{ for } \mu\text{-a.e. } x \in \Omega. \tag{4.32}$$

Then we have the inequalities

$$\begin{aligned}I_f(p, q) &\leq \frac{(R-1)(1-r)}{R-r} \sup_{t \in (r,R)} \Psi_f(t; r, R) \\ &\leq (R-1)(1-r) \frac{f'_-(R) - f'_+(r)}{R-r} \\ &\leq \frac{1}{4}(R-r)\left[f'_-(R) - f'_+(r)\right],\end{aligned} \tag{4.33}$$

and $\Psi_f(\cdot; r, R) : (r, R) \to \mathbb{R}$ is defined by

$$\Psi_f(t; r, R) = \frac{f(R) - f(t)}{R - t} - \frac{f(t) - f(r)}{t - r}.$$

We also have the inequality

$$\begin{aligned}I_f(p, q) &\leq \frac{1}{4}(R-r) \frac{f(R)(1-r) + f(r)(R-1)}{(R-1)(1-r)} \\ &\leq \frac{1}{4}(R-r)\left[f'_-(R) - f'_+(r)\right].\end{aligned} \tag{4.34}$$

The proof follows by Theorem 6 by choosing $w(x) = p(x)$, $f(x) = \frac{q(x)}{p(x)}$, $m = r$ and $M = R$ and performing the required calculations. The details are omitted.

Utilising the same approach and Theorem 7 we can also state that:

Proposition 11 (Dragomir [20]) *With the assumptions of Proposition 10 we have*

$$I_f(p,q) \le 2\max\left\{\frac{R-1}{R-r}, \frac{1-r}{R-r}\right\}\left[\frac{f(r)+f(R)}{2} - f\left(\frac{r+R}{2}\right)\right] \tag{4.35}$$
$$\le \frac{1}{2}\max\{R-1, 1-r\}\left[f'_-(R) - f'_+(r)\right].$$

The above results can be utilized to obtain various inequalities for the divergence measures in Information Theory that are particular instances of f-divergence.

Consider the Kullback-Leibler divergence

$$D_{KL}(p,q) := \int_\Omega p(x)\ln\left[\frac{p(x)}{q(x)}\right]d\mu(x), \quad p,\ q \in \mathcal{P},$$

which is an f-divergence for the convex function $f : (0,\infty) \to \mathbb{R}$, $f(t) = -\ln t$.

If $p, q \in \mathcal{P}$ such that there exists the constants $0 < r < 1 < R < \infty$ with

$$r \le \frac{q(x)}{p(x)} \le R \text{ for } \mu\text{-a.e. } x \in \Omega. \tag{4.36}$$

then we get from (4.33) that

$$D_{KL}(p,q) \le \frac{(R-1)(1-r)}{rR}, \tag{4.37}$$

from (4.34) that

$$D_{KL}(p,q) \le \frac{1}{4}(R-r)\ln\left[R^{-\frac{1}{R-1}}r^{-\frac{1}{1-r}}\right]$$

and from (4.35) that

$$D_{KL}(p,q) \le 2\max\left\{\frac{R-1}{R-r}, \frac{1-r}{R-r}\right\}\ln\left(\frac{A(r,R)}{G(r,R)}\right) \tag{4.38}$$
$$\le \frac{1}{2}\max\{R-1, 1-r\}\left(\frac{R-r}{rR}\right),$$

where $A(r,R)$ is the arithmetic mean and $G(r,R)$ is the geometric mean of the positive numbers r and R.

5 Superadditivity and Monotonicity Properties

5.1 General Results

For a μ-measurable function $w : \Omega \to \mathbb{R}$, with $w(x) \geq 0$ for μ-a.e. $x \in \Omega$ and $\int_\Omega w d\mu > 0$ we consider the functional

$$J(w; \Phi, f) := \int_\Omega w(\Phi \circ f) d\mu - \Phi\left(\frac{\int_\Omega wf d\mu}{\int_\Omega w d\mu}\right) \int_\Omega w d\mu \geq 0, \tag{5.1}$$

where $\Phi : I \to \mathbb{R}$ is a continuous convex function on the interval of real numbers I, $f : \Omega \to \mathbb{R}$ is μ-measurable and such that f, $\Phi \circ f \in L_w(\Omega, \mu)$.

Theorem 8 (Dragomir [17]) *Let $w_i : \Omega \to \mathbb{R}$, with $w_i(x) \geq 0$ for μ-a.e. $x \in \Omega$ and $\int_\Omega w_i d\mu > 0$, $i \in \{1, 2\}$. If $\Phi : I \to \mathbb{R}$ is a continuous convex function on the interval of real numbers I, $f : \Omega \to \mathbb{R}$ is μ-measurable and such that f, $\Phi \circ f \in L_{w_1}(\Omega, \mu) \cap L_{w_2}(\Omega, \mu)$, then*

$$J(w_1 + w_2; \Phi, f) \geq J(w_1; \Phi, f) + J(w_2; \Phi, f) \geq 0 \tag{5.2}$$

i.e., J is a superadditive functional of weights.

Moreover, if $w_2 \geq w_1 \geq 0$ μ-a.e. on Ω, then

$$J(w_2; \Phi, f) \geq J(w_1; \Phi, f) \geq 0, \tag{5.3}$$

i.e., J is a monotonic nondecreasing functional of weights.

Proof Utilising the convexity property of Φ we have successively

$$\begin{aligned}
&J(w_1 + w_2; \Phi, f) \\
&= \int_\Omega (w_1 + w_2)(\Phi \circ f) d\mu - \Phi\left(\frac{\int_\Omega (w_1 + w_2) f d\mu}{\int_\Omega (w_1 + w_2) d\mu}\right) \int_\Omega (w_1 + w_2) d\mu \\
&= \int_\Omega w_1 (\Phi \circ f) d\mu + \int_\Omega w_2 (\Phi \circ f) d\mu \\
&- \Phi\left(\frac{\int_\Omega w_1 d\mu \cdot \frac{\int_\Omega w_1 f d\mu}{\int_\Omega w_1 d\mu} + \int_\Omega w_2 d\mu \cdot \frac{\int_\Omega w_2 f d\mu}{\int_\Omega w_2 d\mu}}{\int_\Omega (w_1 + w_2) d\mu}\right) \int_\Omega (w_1 + w_2) d\mu \\
&\geq \int_\Omega w_1 (\Phi \circ f) d\mu + \int_\Omega w_2 (\Phi \circ f) d\mu \\
&- \left[\frac{\int_\Omega w_1 d\mu}{\int_\Omega (w_1 + w_2) d\mu} \Phi\left(\frac{\int_\Omega w_1 f d\mu}{\int_\Omega w_1 d\mu}\right) + \frac{\int_\Omega w_2 d\mu}{\int_\Omega (w_1 + w_2) d\mu} \Phi\left(\frac{\int_\Omega w_2 f d\mu}{\int_\Omega w_2 d\mu}\right)\right]
\end{aligned} \tag{5.4}$$

$$
\begin{aligned}
&\times \int_\Omega (w_1 + w_2)\, d\mu \\
&= \int_\Omega w_1 (\Phi \circ f)\, d\mu - \Phi \left(\frac{\int_\Omega w_1 f d\mu}{\int_\Omega w_1 d\mu} \right) \int_\Omega w_1 d\mu \\
&+ \int_\Omega w_2 (\Phi \circ f)\, d\mu - \Phi \left(\frac{\int_\Omega w_2 f d\mu}{\int_\Omega w_2 d\mu} \right) \int_\Omega w_2 d\mu \\
&= J (w_1; \Phi, f) + J (w_2; \Phi, f)
\end{aligned}
$$

which proves the superadditivity property.

Now, if $w_2 \geq w_1 \geq 0$, then on applying the superadditivity property we have

$$
\begin{aligned}
J (w_2; \Phi, f) = J (w_1 + (w_2 - w_1) ; \Phi, f) &\geq J (w_1; \Phi, f) + J (w_2 - w_1; \Phi, f) \\
&\geq J (w_1; \Phi, f)
\end{aligned}
$$

since by the Jensen's inequality for the positive weights we have $J (w_2 - w_1; \Phi, f) \geq 0$. □

The above theorem has a simple however interesting consequence that provides both a refinement and a reverse for the Jensen's integral inequality:

Corollary 12 *Let $w_i : \Omega \to \mathbb{R}$, with $w_i (x) \geq 0$ for μ-a.e. $x \in \Omega$, $\int_\Omega w_i d\mu > 0$, $i \in \{1, 2\}$ and there exists the nonnegative constants γ, Γ such that*

$$
0 \leq \gamma \leq \frac{w_2}{w_1} \leq \Gamma < \infty \ \mu\text{-a.e. on } \Omega. \tag{5.5}
$$

If $\Phi : I \to \mathbb{R}$ is a continuous convex function on the interval of real numbers I, $f : \Omega \to \mathbb{R}$ is μ-measurable and such that $f, \Phi \circ f \in L_{w_1} (\Omega, \mu) \cap L_{w_2} (\Omega, \mu)$, then

$$
\begin{aligned}
0 &\leq \gamma \left[\int_\Omega w_1 (\Phi \circ f)\, d\mu - \Phi \left(\frac{\int_\Omega w_1 f d\mu}{\int_\Omega w_1 d\mu} \right) \int_\Omega w_1 d\mu \right] \\
&\leq \int_\Omega w_2 (\Phi \circ f)\, d\mu - \Phi \left(\frac{\int_\Omega w_2 f d\mu}{\int_\Omega w_2 d\mu} \right) \int_\Omega w_2 d\mu \\
&\leq \Gamma \left[\int_\Omega w_1 (\Phi \circ f)\, d\mu - \Phi \left(\frac{\int_\Omega w_1 f d\mu}{\int_\Omega w_1 d\mu} \right) \int_\Omega w_1 d\mu \right]
\end{aligned} \tag{5.6}
$$

or, equivalently,

$$
0 \leq \gamma \frac{\int_\Omega w_1 d\mu}{\int_\Omega w_2 d\mu} \left[\frac{\int_\Omega w_1 (\Phi \circ f)\, d\mu}{\int_\Omega w_1 d\mu} - \Phi \left(\frac{\int_\Omega w_1 f d\mu}{\int_\Omega w_1 d\mu} \right) \right] \tag{5.7}
$$

$$\le \frac{\int_{\Omega} w_2 (\Phi \circ f)\, d\mu}{\int_{\Omega} w_2 d\mu} - \Phi\left(\frac{\int_{\Omega} w_2 f d\mu}{\int_{\Omega} w_2 d\mu}\right)$$
$$\le \Gamma \frac{\int_{\Omega} w_1 d\mu}{\int_{\Omega} w_2 d\mu}\left[\frac{\int_{\Omega} w_1 (\Phi \circ f)\, d\mu}{\int_{\Omega} w_1 d\mu} - \Phi\left(\frac{\int_{\Omega} w_1 f d\mu}{\int_{\Omega} w_1 d\mu}\right)\right].$$

Proof From (5.5) we have $\gamma w_1 \le w_2 \le \Gamma w_1 < \infty$ μ-a.e. on Ω and by the monotonicity property (5.3) we get

$$J(\Gamma w_1; \Phi, f) \ge J(w_2; \Phi, f) \ge J(\gamma w_1; \Phi, f). \tag{5.8}$$

Since the functional is positive homogeneous, namely $J(\alpha w; \Phi, f) = \alpha J(w; \Phi, f)$, then we get from (5.8) the desired result (5.6). □

Remark 19 Assume that $\mu(\Omega) < \infty$ and let $w : \Omega \to \mathbb{R}$, with $w(x) \ge 0$ for μ-a.e. $x \in \Omega$, $\int_{\Omega} w d\mu > 0$ and w is essentially bounded, i.e. $\operatorname{essinf}_{x\in\Omega} w(x)$ and $\operatorname{essup}_{x\in\Omega} w(x)$ are finite. If $\Phi : I \to \mathbb{R}$ is a continuous convex function on the interval of real numbers I, $f : \Omega \to \mathbb{R}$ is μ-measurable and such that f, $\Phi \circ f \in L_w(\Omega, \mu) \cap L(\Omega, \mu)$, then

$$0 \le \frac{\operatorname{essinf}_{x\in\Omega} w(x)}{\frac{1}{\mu(\Omega)}\int_{\Omega} w d\mu}\left[\frac{\int_{\Omega} (\Phi \circ f)\, d\mu}{\mu(\Omega)} - \Phi\left(\frac{\int_{\Omega} f d\mu}{\mu(\Omega)}\right)\right] \tag{5.9}$$
$$\le \frac{\int_{\Omega} w (\Phi \circ f)\, d\mu}{\int_{\Omega} w d\mu} - \Phi\left(\frac{\int_{\Omega} w f d\mu}{\int_{\Omega} w d\mu}\right)$$
$$\le \frac{\operatorname{essup}_{x\in\Omega} w(x)}{\frac{1}{\mu(\Omega)}\int_{\Omega} w d\mu}\left[\frac{\int_{\Omega} (\Phi \circ f)\, d\mu}{\mu(\Omega)} - \Phi\left(\frac{\int_{\Omega} f d\mu}{\mu(\Omega)}\right)\right].$$

This result can be used to provide the following result related to the Hermite-Hadamard inequality for convex functions that states that

$$\frac{1}{b-a}\int_a^b \Phi(t)\, dt \ge \Phi\left(\frac{a+b}{2}\right)$$

for any convex function $\Phi : [a, b] \to \mathbb{R}$.

Indeed , if $w : [a, b] \to [0, \infty)$ is Lebesgue integrable, then we have

$$0 \le \frac{\operatorname{essinf}_{x\in[a,b]} w(x)}{\frac{1}{b-a}\int_a^b w(t)\, dt}\left[\frac{1}{b-a}\int_a^b \Phi(t)\, dt - \Phi\left(\frac{a+b}{2}\right)\right] \tag{5.10}$$
$$\le \frac{\int_a^b w(t)\, \Phi(t)\, dt}{\int_a^b w(t)\, dt} - \Phi\left(\frac{\int_{\Omega} w(t)\, t dt}{\int_a^b w(t)\, dt}\right)$$
$$\le \frac{\operatorname{essup}_{x\in[a,b]} w(x)}{\frac{1}{b-a}\int_a^b w(t)\, dt}\left[\frac{1}{b-a}\int_a^b \Phi(t)\, dt - \Phi\left(\frac{a+b}{2}\right)\right].$$

Now we consider another functional depending on the weights

$$K\left(w;\Phi,f\right):=\frac{J\left(w;\Phi,f\right)}{\int_\Omega wd\mu}=\frac{\int_\Omega w\left(\Phi\circ f\right)d\mu}{\int_\Omega wd\mu}-\Phi\left(\frac{\int_\Omega wfd\mu}{\int_\Omega wd\mu}\right)\geq 0$$

and the composite functional

$$L\left(w;\Phi,f\right):=\left(\int_\Omega wd\mu\right)\ln\left[K\left(w;\Phi,f\right)+1\right]\geq 0,$$

where $\Phi:I\to\mathbb{R}$ is a continuous convex function on the interval of real numbers I and $f:\Omega\to\mathbb{R}$ is μ-measurable and such that $f,\ \Phi\circ f\in L_w\left(\Omega,\mu\right)$.

Theorem 9 (Dragomir [17]) *With the assumptions of Theorem 8, L is a superadditive and monotonic nondecreasing functional of weights.*

Proof Let $w_i:\Omega\to\mathbb{R}$, with $w_i\left(x\right)\geq 0$ for μ-a.e. $x\in\Omega$ and $\int_\Omega w_id\mu>0$, $i\in\{1,2\}$ such that $f,\ \Phi\circ f\in L_{w_1}\left(\Omega,\mu\right)\cap L_{w_2}\left(\Omega,\mu\right)$.

Utilising the superadditivity property of J we have

$$\begin{aligned}
&L\left(w_1+w_2;\Phi,f\right) \qquad (5.11)\\
&=\left(\int_\Omega\left(w_1+w_2\right)d\mu\right)\ln\left[K\left(w_1+w_2;\Phi,f\right)+1\right]\\
&=\left(\int_\Omega\left(w_1+w_2\right)d\mu\right)\ln\left[\frac{J\left(w_1+w_2;\Phi,f\right)}{\int_\Omega\left(w_1+w_2\right)d\mu}+1\right]\\
&\geq\left(\int_\Omega\left(w_1+w_2\right)d\mu\right)\ln\left[\frac{J\left(w_1;\Phi,f\right)+J\left(w_2;\Phi,f\right)}{\int_\Omega\left(w_1+w_2\right)d\mu}+1\right]\\
&=\left(\int_\Omega\left(w_1+w_2\right)d\mu\right)\\
&\times\ln\left[\frac{\int_\Omega w_1d\mu\cdot\frac{J(w_1;\Phi,f)}{\int_\Omega w_1d\mu}+\int_\Omega w_2d\mu\cdot\frac{J(w_2;\Phi,f)}{\int_\Omega w_2d\mu}}{\int_\Omega\left(w_1+w_2\right)d\mu}+1\right]\\
&=\left(\int_\Omega\left(w_1+w_2\right)d\mu\right)\\
&\times\ln\left[\frac{\int_\Omega w_1d\mu\cdot\left(\frac{J(w_1;\Phi,f)}{\int_\Omega w_1d\mu}+1\right)+\int_\Omega w_2d\mu\cdot\left(\frac{J(w_2;\Phi,f)}{\int_\Omega w_2d\mu}+1\right)}{\int_\Omega\left(w_1+w_2\right)d\mu}\right]\\
&:=A.
\end{aligned}$$

By the weighted arithmetic mean–geometric mean inequality we have

$$\frac{\int_\Omega w_1 d\mu \cdot \left(\frac{J(w_1;\Phi,f)}{\int_\Omega w_1 d\mu}+1\right)+\int_\Omega w_2 d\mu \cdot \left(\frac{J(w_2;\Phi,f)}{\int_\Omega w_2 d\mu}+1\right)}{\int_\Omega (w_1+w_2)\,d\mu}$$
$$\geq \left(\frac{J(w_1;\Phi,f)}{\int_\Omega w_1 d\mu}+1\right)^{\frac{\int_\Omega w_1 d\mu}{\int_\Omega (w_1+w_2) d\mu}}\left(\frac{J(w_2;\Phi,f)}{\int_\Omega w_2 d\mu}+1\right)^{\frac{\int_\Omega w_2 d\mu}{\int_\Omega (w_1+w_2) d\mu}},$$

therefore, by taking the logarithm and utilizing the definition of the functional K, we get the inequality

$$A \geq \left(\int_\Omega w_1 d\mu\right)\ln\left(K(w_1;\Phi,f)+1\right)+\left(\int_\Omega w_2 d\mu\right)\ln\left(K(w_2;\Phi,f)+1\right) \tag{5.12}$$
$$= L(w_1;\Phi,f)+L(w_2;\Phi,f).$$

Utilising (5.11) and (5.12) we deduce the superadditivity of the functional L as a function of weights.

Since $L(w;\Phi,f)\geq 0$ for any weight w and it is superadditive, by employing a similar argument to the one in the proof of Theorem 8 we conclude that it is also monotonic nondecreasing as a function of weights. □

The following result provides another refinement and reverse of the Jensen inequality:

Corollary 13 *Let $w_i:\Omega\to\mathbb{R}$ with $w_i(x)\geq 0$ for μ-a.e. $x\in\Omega$, $\int_\Omega w_i d\mu>0$, $i\in\{1,2\}$ and there exists the nonnegative constants γ, Γ such that*

$$0\leq\gamma\leq\frac{w_2}{w_1}\leq\Gamma<\infty\ \mu\text{-a.e. on }\Omega.$$

If $\Phi:I\to\mathbb{R}$ is a continuous convex function on the interval of real numbers I, $f:\Omega\to\mathbb{R}$ is μ-measurable and such that $f,\Phi\circ f\in L_{w_1}(\Omega,\mu)\cap L_{w_2}(\Omega,\mu)$, then

$$0\leq\left[\frac{\int_\Omega w_1(\Phi\circ f)\,d\mu}{\int_\Omega w_1 d\mu}-\Phi\left(\frac{\int_\Omega w_1 f\,d\mu}{\int_\Omega w_1 d\mu}\right)+1\right]^{\gamma\frac{\left(\int_\Omega w_1 d\mu\right)}{\left(\int_\Omega w_2 d\mu\right)}}-1 \tag{5.13}$$
$$\leq\frac{\int_\Omega w_2(\Phi\circ f)\,d\mu}{\int_\Omega w_2 d\mu}-\Phi\left(\frac{\int_\Omega w_2 f\,d\mu}{\int_\Omega w_2 d\mu}\right)$$
$$\leq\left[\frac{\int_\Omega w_1(\Phi\circ f)\,d\mu}{\int_\Omega w_1 d\mu}-\Phi\left(\frac{\int_\Omega w_1 f\,d\mu}{\int_\Omega w_1 d\mu}\right)+1\right]^{\Gamma\frac{\left(\int_\Omega w_1 d\mu\right)}{\left(\int_\Omega w_2 d\mu\right)}}-1.$$

Proof Since L is monotonic nondecreasing and positive homogeneous as a function of weights, we have

$$\gamma L\left(w_1;\Phi,f\right)\le L\left(w_2;\Phi,f\right)\le\Gamma L\left(w_1;\Phi,f\right),$$

namely

$$\begin{aligned}\left[K\left(w_1;\Phi,f\right)+1\right]^{\gamma\left(\int_\Omega w_1d\mu\right)}&\le\left[K\left(w_2;\Phi,f\right)+1\right]^{\left(\int_\Omega w_2d\mu\right)}\\&\le\left[K\left(w_1;\Phi,f\right)+1\right]^{\Gamma\left(\int_\Omega w_1d\mu\right)},\end{aligned}$$

which provides that

$$\begin{aligned}\left[K\left(w_1;\Phi,f\right)+1\right]^{\gamma\frac{\left(\int_\Omega w_1d\mu\right)}{\left(\int_\Omega w_2d\mu\right)}}-1&\le K\left(w_2;\Phi,f\right)\\&\le\left[K\left(w_1;\Phi,f\right)+1\right]^{\Gamma\frac{\left(\int_\Omega w_1d\mu\right)}{\left(\int_\Omega w_2d\mu\right)}}-1.\end{aligned}$$

□

Remark 20 Assume that $\mu\left(\Omega\right)<\infty$ and let $w:\Omega\to\mathbb{R}$, with $w\left(x\right)\ge0$ for μ-a.e. $x\in\Omega$, $\int_\Omega wd\mu>0$ and w is essentially bounded, i.e. $\operatorname{essinf}_{x\in\Omega}w\left(x\right)$ and $\operatorname{essup}_{x\in\Omega}w\left(x\right)$ are finite. If $\Phi:I\to\mathbb{R}$ is a continuous convex function on the interval of real numbers I, $f:\Omega\to\mathbb{R}$ is μ-measurable and such that f, $\Phi\circ f\in L_w\left(\Omega,\mu\right)\cap L\left(\Omega,\mu\right)$, then

$$\begin{aligned}0&\le\left[\frac{\int_\Omega\left(\Phi\circ f\right)d\mu}{\mu\left(\Omega\right)}-\Phi\left(\frac{\int_\Omega fd\mu}{\mu\left(\Omega\right)}\right)+1\right]^{\frac{ess\inf_{x\in\Omega}w(x)}{\frac{1}{\mu(\Omega)}\left(\int_\Omega wd\mu\right)}}-1\qquad(5.14)\\&\le\frac{\int_\Omega w\left(\Phi\circ f\right)d\mu}{\int_\Omega wd\mu}-\Phi\left(\frac{\int_\Omega wfd\mu}{\int_\Omega wd\mu}\right)\\&\le\left[\frac{\int_\Omega\left(\Phi\circ f\right)d\mu}{\mu\left(\Omega\right)}-\Phi\left(\frac{\int_\Omega fd\mu}{\mu\left(\Omega\right)}\right)+1\right]^{\frac{ess\sup_{x\in\Omega}w(x)}{\frac{1}{\mu(\Omega)}\left(\int_\Omega wd\mu\right)}}-1.\end{aligned}$$

In particular, if $w:[a,b]\to[0,\infty)$ is Lebesgue integrable, then we have the following result related to the Hermite-Hadamard inequality for the convex function $\Phi:[a,b]\to\mathbb{R}$

$$\begin{aligned}0&\le\left[\frac{1}{b-a}\int_a^b\Phi\left(t\right)dt-\Phi\left(\frac{a+b}{2}\right)+1\right]^{\frac{\operatorname{essinf}_{x\in[a,b]}w(x)}{\frac{1}{b-a}\int_a^bw(t)dt}}-1\qquad(5.15)\\&\le\frac{\int_a^bw\left(t\right)\Phi\left(t\right)dt}{\int_a^bw\left(t\right)dt}-\Phi\left(\frac{\int_\Omega w\left(t\right)tdt}{\int_a^bw\left(t\right)dt}\right)\\&\le\left[\frac{1}{b-a}\int_a^b\Phi\left(t\right)dt-\Phi\left(\frac{a+b}{2}\right)+1\right]^{\frac{\operatorname{essup}_{x\in[a,b]}w(x)}{\frac{1}{b-a}\int_a^bw(t)dt}}-1.\end{aligned}$$

5.2 Applications for the Hölder Inequality

Assume that $p > 1$. If $h : \Omega \to \mathbb{R}$ is μ-measurable, $\mu(\Omega) < \infty$, $|h|, |h|^p \in L_w(\Omega, \mu) \cap L(\Omega, \mu)$, then by (5.9) we have the bounds

$$0 \leq \frac{\operatorname{essinf}_{x\in\Omega} w(x)}{\frac{1}{\mu(\Omega)}\int_\Omega w d\mu}\left[\frac{1}{\mu(\Omega)}\int_\Omega |h|^p d\mu - \left(\frac{1}{\mu(\Omega)}\int_\Omega |h| d\mu\right)^p\right] \tag{5.16}$$
$$\leq \frac{1}{\int_\Omega w d\mu}\int_\Omega w|h|^p d\mu - \left(\frac{1}{\int_\Omega w d\mu}\int_\Omega w|h| d\mu\right)^p$$
$$\leq \frac{\operatorname{essup}_{x\in\Omega} w(x)}{\frac{1}{\mu(\Omega)}\int_\Omega w d\mu}\left[\frac{1}{\mu(\Omega)}\int_\Omega |h|^p d\mu - \left(\frac{1}{\mu(\Omega)}\int_\Omega |h| d\mu\right)^p\right].$$

Proposition 12 (Dragomir [17]) *If $f \in L_p(\Omega, \mu)$, $g \in L_q(\Omega, \mu)$ with $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$, $\mu(\Omega) < \infty$ and there exists the constants δ, $\Delta > 0$ and such that*

$$\delta \leq |g| \leq \Delta \ \mu\text{-a.e. on } \Omega, \tag{5.17}$$

then we have

$$0 \leq \frac{\delta^q}{\frac{1}{\mu(\Omega)}\int_\Omega |g|^q d\mu}\left[\frac{1}{\mu(\Omega)}\int_\Omega \frac{|f|^p}{|g|^q} d\mu - \left(\frac{1}{\mu(\Omega)}\int_\Omega \frac{|f|}{|g|^{q-1}} d\mu\right)^p\right] \tag{5.18}$$
$$\leq \frac{\int_\Omega |f|^p d\mu}{\int_\Omega |g|^q d\mu} - \left(\frac{\int_\Omega |fg| d\mu}{\int_\Omega |g|^q d\mu}\right)^p$$
$$\leq \frac{\Delta^q}{\frac{1}{\mu(\Omega)}\int_\Omega |g|^q d\mu}\left[\frac{1}{\mu(\Omega)}\int_\Omega \frac{|f|^p}{|g|^q} d\mu - \left(\frac{1}{\mu(\Omega)}\int_\Omega \frac{|f|}{|g|^{q-1}} d\mu\right)^p\right].$$

Proof The inequalities (5.18) follows from (5.16) by choosing

$$h = \frac{|f|}{|g|^{q-1}} \text{ and } w = |g|^q .$$

The details are omitted. □

Remark 21 We observe that for $p = q = 2$ we have from (5.18) the following reverse of the Cauchy-Bunyakovsky-Schwarz inequality

$$0 \leq \delta^2 \mu(\Omega)\left[\frac{1}{\mu(\Omega)}\int_\Omega \left|\frac{f}{g}\right|^2 d\mu - \left(\frac{1}{\mu(\Omega)}\int_\Omega \left|\frac{f}{g}\right| d\mu\right)^2\right]\int_\Omega |g|^2 d\mu \tag{5.19}$$
$$\leq \int_\Omega |g|^2 d\mu \int_\Omega |f|^2 d\mu - \left(\int_\Omega |fg| d\mu\right)^2$$

$$\leq \Delta^2 \mu(\Omega) \left[\frac{1}{\mu(\Omega)} \int_\Omega \left| \frac{f}{g} \right|^2 d\mu - \left(\frac{1}{\mu(\Omega)} \int_\Omega \left| \frac{f}{g} \right| d\mu \right)^2 \right] \int_\Omega |g|^2 \, d\mu,$$

provided that $f, g \in L_2(\Omega, \mu)$ and g satisfies the bounds (5.17).

Similar results can be stated by utilizing the inequality (5.13), however the details are not presented here.

5.3 Applications for f-Divergence Measures

The following result holds:

Proposition 13 (Dragomir [17]) *Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with the property that $f(1) = 0$. Assume that $p, q \in \mathcal{P}$ and there exists the constants $0 < s < 1 < S < \infty$ such that*

$$s \leq \frac{p(x)}{q(x)} \leq S \text{ for } \mu\text{-a.e. } x \in \Omega. \tag{5.20}$$

Then we have the inequalities

$$\begin{aligned} & s\left[I_{f\left(\frac{1}{\cdot}\right)}(q, p) - f\left(D_{\chi^2}(p, q) + 1\right) \right] \\ & \leq I_f(p, q) \\ & \leq S\left[I_{f\left(\frac{1}{\cdot}\right)}(q, p) - f\left(D_{\chi^2}(p, q) + 1\right) \right]. \end{aligned} \tag{5.21}$$

Proof If we use the inequality (5.6) we get

$$\begin{aligned} & s\left[\int_\Omega q f\left(\frac{q}{p}\right) d\mu - f\left(\int_\Omega \frac{q^2}{p} d\mu \right) \right] \\ & \leq \int_\Omega p f\left(\frac{q}{p}\right) d\mu \\ & \leq S\left[\int_\Omega q f\left(\frac{q}{p}\right) d\mu - f\left(\int_\Omega \frac{q^2}{p} d\mu \right) \right]. \end{aligned} \tag{5.22}$$

Since

$$\int_\Omega \frac{q^2}{p} d\mu = D_{\chi^2}(p, q) + 1$$

and

$$\int_{\Omega} qf\left(\frac{q}{p}\right) d\mu = I_{f\left(\frac{1}{\cdot}\right)}(q,p),$$

then from (5.22) we deduce the desired result (5.21). □

Consider the Kullback-Leibler divergence

$$D_{KL}(p,q) := \int_{\Omega} p(x) \ln\left[\frac{p(x)}{q(x)}\right] d\mu(x), \quad p,\ q \in \mathcal{P},$$

which is an f-divergence for the convex function $f : (0,\infty) \to \mathbb{R}$, $f(t) = -\ln t$.

If $p, q \in \mathcal{P}$ such that there exists the constants $0 < s < 1 < S < \infty$ with

$$s \le \frac{p(x)}{q(x)} \le S \text{ for } \mu\text{-a.e. } x \in \Omega. \tag{5.23}$$

then we get from (5.21) that

$$\begin{aligned} & s\left[\ln\left(D_{\chi^2}(p,q)+1\right) - D_{KL}(q,p)\right] \\ & \le D_{KL}(p,q) \\ & \le S\left[\ln\left(D_{\chi^2}(p,q)+1\right) - D_{KL}(q,p)\right]. \end{aligned} \tag{5.24}$$

Similar results for f-divergence measures can be stated by utilizing the inequality (5.13), however the details are not presented here.

6 Inequalities for Selfadjoint Operators

6.1 Preliminary Facts

The above integral inequalities can be used to obtain various reverses of Jensen's inequality for convex functions of selfadjoint operators on complex Hilbert spaces. In order to state these results, we need the following preparations.

Let A be a selfadjoint operator on the complex Hilbert space $(H, \langle .,. \rangle)$ with the spectrum $\mathrm{Sp}(A)$ included in the interval $[m, M]$ for some real numbers $m < M$ and let $\{E_\lambda\}_\lambda$ be its *spectral family*. Then for any continuous function $f : [m, M] \to \mathbb{R}$, it is well known that we have the following *spectral representation in terms of the Riemann-Stieltjes integral* (see for instance [30, p. 257]):

$$\langle f(A)x, y \rangle = \int_{m-0}^{M} f(\lambda)\, d\langle E_\lambda x, y\rangle, \tag{6.1}$$

and

$$\|f(A)x\|^2 = \int_{m-0}^{M} |f(\lambda)|^2 d\|E_\lambda x\|^2, \tag{6.2}$$

for any $x, y \in H$.

The function $g_{x,y}(\lambda) := \langle E_\lambda x, y\rangle$ is of *bounded variation* on the interval $[m, M]$ and $g_{x,y}(m-0) = 0$ while $g_{x,y}(M) = \langle x, y\rangle$ for any $x, y \in H$. It is also well known that $g_x(\lambda) := \langle E_\lambda x, x\rangle$ is *monotonic nondecreasing* and *right continuous* on $[m, M]$ for any $x \in H$.

The following result that provides an operator version for the Jensen inequality:

Theorem 10 (Mond-Pečarić [40]) *Let A be a selfadjoint operator on the Hilbert space H and assume that* $\mathrm{Sp}(A) \subseteq [m, M]$ *for some scalars m, M with $m < M$. If Φ is a convex function on $[m, M]$, then*

$$\Phi(\langle Ax, x\rangle) \leq \langle \Phi(A)x, x\rangle \tag{MP}$$

for each $x \in H$ with $\|x\| = 1$.

As a special case of Theorem 10 we have the following Hölder-McCarthy inequality:

Theorem 11 (Hölder-McCarthy [38]) *Let A be a selfadjoint positive operator on a Hilbert space H. Then for all $x \in H$ with $\|x\| = 1$,*

(i) $\langle A^r x, x\rangle \geq \langle Ax, x\rangle^r$ *for all* $r > 1$;
(ii) $\langle A^r x, x\rangle \leq \langle Ax, x\rangle^r$ *for all* $0 < r < 1$;
(iii) If A is invertible, then $\langle A^r x, x\rangle \geq \langle Ax, x\rangle^r$ *for all* $r < 0$.

The following reverse for the (MP) inequality that generalizes the scalar Lah-Ribarić inequality for convex functions is well known, see for instance [26, p. 57]:

Theorem 12 *Let A be a selfadjoint operator on the Hilbert space H and assume that* $\mathrm{Sp}(A) \subseteq [m, M]$ *for some scalars m, M with $m < M$. If Φ is a convex function on $[m, M]$, then*

$$\langle \Phi(A)x, x\rangle \leq \frac{M - \langle Ax, x\rangle}{M - m}\Phi(m) + \frac{\langle Ax, x\rangle - m}{M - m}\Phi(M) \tag{LR}$$

for each $x \in H$ with $\|x\| = 1$.

In [22] we obtained the following weighted version of (MP) and (LR).

Theorem 13 (Dragomir [22]) *Let A be a selfadjoint operator on the Hilbert space H and assume that* $\mathrm{Sp}(A) \subseteq [m, M]$ *for some scalars m, M with $m < M$. If $\Phi : [k, K] \subset \mathbb{R} \to \mathbb{R}$ is a continuous convex function on the interval $[k, K]$, $w : [m, M] \to [0, \infty)$ is continuous on $[m, M]$, $f : [m, M] \subset \mathbb{R} \to \mathbb{R}$ is a continuous function on the interval $[m, M]$ and with the property that*

$$k \leq f(t) \leq K \text{ for any } t \in [m, M], \tag{6.3}$$

then

$$\begin{aligned}
&\Phi\left(\frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle}\right) \qquad (6.4)\\
&\leq \frac{\langle w(A)(\Phi \circ f)(A) x, x\rangle}{\langle w(A) x, x\rangle}\\
&\leq \frac{\left(K-\frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle}\right) \Phi(k)+\left(\frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle}-k\right) \Phi(K)}{K-k},
\end{aligned}$$

for any $x \in H$ *with* $\langle w(A) x, x\rangle \neq 0$.

For various particular instances of (6.4) that are of interest being related to Hölder-McCarthy's inequalities mentioned above, see [22].

For classical and recent result concerning inequalities for continuous functions of selfadjoint operators, see the recent monographs, [18, 26] and [19].

6.2 Reverses for Functions of Operators

We have the following results:

Theorem 14 (Dragomir [24]) *Let* A *be a selfadjoint operator on the Hilbert space* H *such that* $\operatorname{Sp}(A) \subseteq [k, K]$ *for some scalars* k, K *with* $k<K$. *Assume that* $\Phi:[k, K] \subset \mathbb{R} \rightarrow \mathbb{R}$ *is a continuous convex function on the interval* $[k, K]$, $w:[k, K] \rightarrow[0, \infty)$ *is continuous on* $[k, K]$, $f:[k, K] \subset \mathbb{R} \rightarrow \mathbb{R}$ *is a continuous function on the interval* $[k, K]$ *and satisfies the property (6.3)*

(i) If Φ *is continuously differentiable on* (k, K), *then we have*

$$\begin{aligned}
0 &\leq \frac{\langle w(A)(\Phi \circ f)(A) x, x\rangle}{\langle w(A) x, x\rangle}-\Phi\left(\frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle}\right) \qquad (6.5)\\
&\leq \frac{\left\langle\left(\Phi^{\prime} \circ f\right)(A) f(A) w(A) x, x\right\rangle}{\langle w(A) x, x\rangle}\\
&\quad-\frac{\left\langle\left(\Phi^{\prime} \circ f\right)(A) w(A) x, x\right\rangle}{\langle w(A) x, x\rangle} \frac{\langle f(A) w(A) x, x\rangle}{\langle w(A) x, x\rangle}\\
&\leq \frac{1}{2}\left[\Phi_{-}^{\prime}(K)-\Phi_{+}^{\prime}(k)\right] \frac{\left\langle\left|f(A)-\frac{\langle f(A) w(A) x, x\rangle}{\langle w(A) x, x\rangle} 1_H\right| x, x\right\rangle}{\langle w(A) x, x\rangle}\\
&\leq \frac{1}{2}\left[\Phi_{-}^{\prime}(K)-\Phi_{+}^{\prime}(k)\right]\left[\frac{\left\langle f^2(A) w(A) x, x\right\rangle}{\langle w(A) x, x\rangle}-\left(\frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle}\right)^2\right]^{\frac{1}{2}}\\
&\leq \frac{1}{4}\left[\Phi_{-}^{\prime}(K)-\Phi_{+}^{\prime}(k)\right](K-k)
\end{aligned}$$

for any $x \in H$ *with* $\langle w(A) x, x\rangle \neq 0$.

(ii) *If we consider the function* $\Psi_{\Phi}(\cdot;k,K):(k,K)\to\mathbb{R}$ *defined by*

$$\Psi_{\Phi}(t;k,K)=\frac{\Phi(K)-\Phi(t)}{K-t}-\frac{\Phi(t)-\Phi(k)}{t-k},$$

then

$$\begin{aligned}0&\leq\frac{\langle w(A)(\Phi\circ f)(A)x,x\rangle}{\langle w(A)x,x\rangle}-\Phi\left(\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}\right)\qquad(6.6)\\&\leq\frac{\left(K-\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}\right)\left(\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}-k\right)}{K-k}\sup_{t\in(k,K)}\Psi_{\Phi}(t;k,K)\\&\leq\left(K-\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}\right)\left(\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}-k\right)\frac{\Phi'_{-}(K)-\Phi'_{+}(k)}{K-k}\\&\leq\frac{1}{4}\left[\Phi'_{-}(K)-\Phi'_{+}(k)\right](K-k)\end{aligned}$$

and

$$\begin{aligned}0&\leq\frac{\langle w(A)(\Phi\circ f)(A)x,x\rangle}{\langle w(A)x,x\rangle}-\Phi\left(\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}\right)\qquad(6.7)\\&\leq\frac{1}{4}(K-k)\Psi_{\Phi}\left(\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle};k,K\right)\\&\leq\frac{1}{4}\left[\Phi'_{-}(K)-\Phi'_{+}(k)\right](K-k)\end{aligned}$$

for any $x\in H$ *with* $\langle w(A)x,x\rangle\neq0$.

(iii) *We have the inequalities*

$$\begin{aligned}0&\leq\frac{\langle w(A)(\Phi\circ f)(A)x,x\rangle}{\langle w(A)x,x\rangle}-\Phi\left(\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}\right)\qquad(6.8)\\&\leq2\max\left\{\frac{K-\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}}{K-k},\frac{\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}-k}{K-k}\right\}\\&\quad\times\left[\frac{\Phi(k)+\Phi(K)}{2}-\Phi\left(\frac{k+K}{2}\right)\right]\end{aligned}$$

and

$$\begin{aligned}0&\leq\frac{\langle w(A)(\Phi\circ f)(A)x,x\rangle}{\langle w(A)x,x\rangle}-\Phi\left(\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}\right)\qquad(6.9)\\&\leq2\left[\frac{\Phi(k)+\Phi(K)}{2}-\Phi\left(\frac{k+K}{2}\right)\right]\end{aligned}$$

for any $x\in H$ *with* $\langle w(A)x,x\rangle\neq0$.

(iv) We also have the inequalities

$$0 \le \frac{\langle w(A)(\Phi \circ f)(A)x, x\rangle}{\langle w(A)x, x\rangle} - \Phi\left(\frac{\langle w(A) f(A)x, x\rangle}{\langle w(A)x, x\rangle}\right) \tag{6.10}$$
$$\le \frac{1}{2}\Psi_\Phi\left(\frac{\langle w(A) f(A)x, x\rangle}{\langle w(A)x, x\rangle}; k, K\right)\frac{\left\langle\left|f(A) - \frac{\langle f(A)w(A)x,x\rangle}{\langle w(A)x,x\rangle}1_H\right|x, x\right\rangle}{\langle w(A)x, x\rangle}$$
$$\le \frac{1}{2}\Psi_\Phi\left(\frac{\langle w(A) f(A)x, x\rangle}{\langle w(A)x, x\rangle}; k, K\right)$$
$$\times\left[\frac{\langle f^2(A) w(A)x, x\rangle}{\langle w(A)x, x\rangle} - \left(\frac{\langle w(A) f(A)x, x\rangle}{\langle w(A)x, x\rangle}\right)^2\right]^{\frac{1}{2}}$$
$$\le \frac{1}{4}\Psi_\Phi\left(\frac{\langle w(A) f(A)x, x\rangle}{\langle w(A)x, x\rangle}; k, K\right)(K - k)$$

for any $x \in H$ *with* $\langle w(A)x, x\rangle \neq 0$.

Proof

(i) Let $\{E_\lambda\}_\lambda$ be the spectral family of the operator A. Let $\varepsilon > 0$ and write the inequality (1.17) on the interval $[k - \varepsilon, K]$ and for the monotonic nondecreasing function $g(t) = \langle E_t x, x\rangle$, $x \in H$ with $\langle w(A)x, x\rangle \neq 0$, to get

$$0 \le \frac{\int_{k-\varepsilon}^{K}(\Phi \circ f)(t)w(t)d\langle E_t x, x\rangle}{\int_{k-\varepsilon}^{K} w(t)d\langle E_t x, x\rangle} - \Phi\left(\frac{\int_{k-\varepsilon}^{K} f(t)w(t)d\langle E_t x, x\rangle}{\int_{k-\varepsilon}^{K} w(t)d\langle E_t x, x\rangle}\right) \tag{6.11}$$
$$\le \frac{\int_{k-\varepsilon}^{K}\left(\Phi' \circ f\right)(t)f(t)w(t)d\langle E_t x, x\rangle}{\int_{k-\varepsilon}^{K} w(t)d\langle E_t x, x\rangle}$$
$$- \frac{\int_{k-\varepsilon}^{K}\left(\Phi' \circ f\right)(t)w(t)d\langle E_t x, x\rangle}{\int_{k-\varepsilon}^{K} w(t)d\langle E_t x, x\rangle}\frac{\int_{k-\varepsilon}^{K} f(t)w(t)d\langle E_t x, x\rangle}{\int_{k-\varepsilon}^{K} w(t)d\langle E_t x, x\rangle}$$
$$\le \frac{1}{2}\frac{\left[\Phi'_-(K) - \Phi'_+(k)\right]}{\int_{k-\varepsilon}^{K} w(t)d\langle E_t x, x\rangle}$$
$$\times\int_{k-\varepsilon}^{K}\left|f(t) - \frac{\int_{k-\varepsilon}^{K} f(s)w(s)d\langle E_s x, x\rangle}{\int_{k-\varepsilon}^{K} w(s)d\langle E_s x, x\rangle}\right|w(t)d\langle E_t x, x\rangle$$
$$\le \frac{1}{2}\left[\Phi'_-(K) - \Phi'_+(k)\right]$$

$$\times\left[\frac{\int_{k-\varepsilon}^{K} f^2(t)\,w(t)\,d\langle E_t x, x\rangle}{\int_{k-\varepsilon}^{K} w(s)\,d\langle E_s x, x\rangle}-\left(\frac{\int_{k-\varepsilon}^{K} f(s)\,w(s)\,d\langle E_s x, x\rangle}{\int_{k-\varepsilon}^{K} w(s)\,d\langle E_s x, x\rangle}\right)^2\right]^{\frac{1}{2}}$$
$$\le \frac{1}{4}\left[\Phi'_-(K)-\Phi'_+(k)\right](K-k).$$

Letting $\varepsilon \to 0+$ and using the spectral representation theorem summarized in (6.1) we get the required inequality (6.5).

(ii) Follows by the first part of Theorem 6 , (iii) follows by Theorem 7 while (iv) follows by the second part of Theorem 6. The details are omitted. □

We have the following generalization and reverse for the Hölder-McCarthy inequality:

Corollary 14 (Dragomir [24]) *Let A be a selfadjoint operator on the Hilbert space H such that $\mathrm{Sp}(A) \subseteq [k, K]$ for some scalars k, K with $k < K$. Assume that $w : [k, K] \to [0, \infty)$ is continuous on $[k, K]$, $f : [k, K] \subset \mathbb{R} \to \mathbb{R}$ is a continuous function on the interval $[k, K]$ and satisfies the property (6.3) with $k > 0$. Assume also that $p \in (-\infty, 0) \cup (1, \infty)$.*

(i) We have

$$0 \le \frac{\langle w(A)\,f^p(A)\,x, x\rangle}{\langle w(A)\,x, x\rangle}-\left(\frac{\langle w(A)\,f(A)\,x, x\rangle}{\langle w(A)\,x, x\rangle}\right)^p \tag{6.12}$$
$$\le p\left[\frac{\langle f^p(A)\,w(A)\,x, x\rangle}{\langle w(A)\,x, x\rangle}-\frac{\left\langle f^{p-1}(A)\,w(A)\,x, x\right\rangle}{\langle w(A)\,x, x\rangle}\frac{\langle f(A)\,w(A)\,x, x\rangle}{\langle w(A)\,x, x\rangle}\right]$$
$$\le \frac{1}{2}p\left(K^{p-1}-k^{p-1}\right)\frac{\left\langle\left|f(A)-\frac{\langle f(A)w(A)x,x\rangle}{\langle w(A)x,x\rangle}1_H\right|x, x\right\rangle}{\langle w(A)\,x, x\rangle}$$
$$\le \frac{1}{2}p\left(K^{p-1}-k^{p-1}\right)\left[\frac{\left\langle f^2(A)\,w(A)\,x, x\right\rangle}{\langle w(A)\,x, x\rangle}-\left(\frac{\langle w(A)\,f(A)\,x, x\rangle}{\langle w(A)\,x, x\rangle}\right)^2\right]^{\frac{1}{2}}$$
$$\le \frac{1}{4}p\left(K^{p-1}-k^{p-1}\right)(K-k)$$

for any $x \in H$ with $\langle w(A)\,x, x\rangle \neq 0$.

(ii) If we consider the function $\Psi_p(\cdot; k, K) : (k, K) \to \mathbb{R}$ defined by

$$\Psi_p(t; k, K) = \frac{K^p - t^p}{K-t} - \frac{t^p - k^p}{t-k},$$

then

$$\begin{aligned}0 &\le \frac{\langle w(A) f^p(A) x, x\rangle}{\langle w(A) x, x\rangle} - \left(\frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle}\right)^p \\ &\le \frac{\left(K - \frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle}\right)\left(\frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle} - k\right)}{K-k} \sup_{t\in(k,K)} \Psi_p(t; k, K) \\ &\le p\frac{K^{p-1}-k^{p-1}}{K-k}\left(K - \frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle}\right)\left(\frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle} - k\right) \\ &\le \frac{1}{4} p\left(K^{p-1}-k^{p-1}\right)(K-k)\end{aligned} \tag{6.13}$$

and

$$\begin{aligned}0 &\le \frac{\langle w(A) f^p(A) x, x\rangle}{\langle w(A) x, x\rangle} - \left(\frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle}\right)^p \\ &\le \frac{1}{4}(K-k)\Psi_p\left(\frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle}; k, K\right) \\ &\le \frac{1}{4} p\left(K^{p-1}-k^{p-1}\right)(K-k)\end{aligned} \tag{6.14}$$

for any $x \in H$ with $\langle w(A) x, x\rangle \neq 0$.

(iii) We have the inequalities

$$\begin{aligned}0 &\le \frac{\langle w(A) f^p(A) x, x\rangle}{\langle w(A) x, x\rangle} - \left(\frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle}\right)^p \\ &\le 2\max\left\{\frac{K - \frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle}}{K-k}, \frac{\frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle} - k}{K-k}\right\} \\ &\times\left[\frac{k^p+K^p}{2} - \left(\frac{k+K}{2}\right)^p\right]\end{aligned} \tag{6.15}$$

and

$$\begin{aligned}0 &\le \frac{\langle w(A) f^p(A) x, x\rangle}{\langle w(A) x, x\rangle} - \left(\frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle}\right)^p \\ &\le 2\left[\frac{k^p+K^p}{2} - \left(\frac{k+K}{2}\right)^p\right]\end{aligned} \tag{6.16}$$

for any $x \in H$ with $\langle w(A) x, x\rangle \neq 0$.

(iv) We also have the inequalities

$$\begin{aligned}0 &\le \frac{\langle w(A) f^p(A) x, x\rangle}{\langle w(A) x, x\rangle} - \left(\frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle}\right)^p \\ &\le \frac{1}{2}\Psi_p\left(\frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle}; k, K\right) \frac{\left\langle \left| f(A) - \frac{\langle f(A) w(A) x, x\rangle}{\langle w(A) x, x\rangle} 1_H \right| x, x\right\rangle}{\langle w(A) x, x\rangle} \\ &\le \frac{1}{2}\Psi_p\left(\frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle}; k, K\right) \\ &\times \left[\frac{\langle f^2(A) w(A) x, x\rangle}{\langle w(A) x, x\rangle} - \left(\frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle}\right)^2\right]^{\frac{1}{2}} \\ &\le \frac{1}{4}\Psi_p\left(\frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle}; k, K\right)(K-k)\end{aligned} \tag{6.17}$$

for any $x \in H$ with $\langle w(A) x, x\rangle \neq 0$.

If $p \in (0,1)$, then by taking $\Phi(t) = -t^p$ we can get similar inequalities. However the details are omitted.

If we take $\Phi(t) = -\ln t$, $t > 0$ in Theorem 14 then we get the following logarithmic inequalities:

Corollary 15 (Dragomir [24]) *Let A be a selfadjoint operator on the Hilbert space H such that $\operatorname{Sp}(A) \subseteq [k, K]$ for some scalars k, K with $k < K$. Assume that $w : [k, K] \to [0, \infty)$ is continuous on $[k, K]$, $f : [k, K] \subset \mathbb{R} \to \mathbb{R}$ is a continuous function on the interval $[k, K]$ and satisfies the property (6.3) with $k > 0$.*

(i) We have

$$\begin{aligned}0 &\le \ln\left(\frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle}\right) - \frac{\langle w(A) \ln f(A) x, x\rangle}{\langle w(A) x, x\rangle} \\ &\le \frac{\langle f^{-1}(A) w(A) x, x\rangle}{\langle w(A) x, x\rangle} \frac{\langle f(A) w(A) x, x\rangle}{\langle w(A) x, x\rangle} - 1 \\ &\le \frac{1}{2}\frac{K-k}{kK} \frac{\left\langle \left| f(A) - \frac{\langle f(A) w(A) x, x\rangle}{\langle w(A) x, x\rangle} 1_H \right| x, x\right\rangle}{\langle w(A) x, x\rangle} \\ &\le \frac{1}{2}\frac{K-k}{kK}\left[\frac{\langle f^2(A) w(A) x, x\rangle}{\langle w(A) x, x\rangle} - \left(\frac{\langle w(A) f(A) x, x\rangle}{\langle w(A) x, x\rangle}\right)^2\right]^{\frac{1}{2}} \\ &\le \frac{1}{4}\frac{(K-k)^2}{kK}\end{aligned} \tag{6.18}$$

for any $x \in H$ with $\langle w(A) x, x\rangle \neq 0$,

(ii) *If we consider the function* $\Psi_{-\ln}(\cdot;k,K):(k,K)\to\mathbb{R}$ *defined by*

$$\Psi_{-\ln}(t;k,K)=\frac{\ln t-\ln k}{t-k}-\frac{\ln K-\ln t}{K-t},$$

then

$$\begin{aligned}0&\le\ln\left(\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}\right)-\frac{\langle w(A)\ln f(A)x,x\rangle}{\langle w(A)x,x\rangle}\qquad(6.19)\\&\le\frac{\left(K-\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}\right)\left(\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}-k\right)}{K-k}\sup_{t\in(k,K)}\Psi_{-\ln}(t;k,K)\\&\le\frac{1}{Kk}\left(K-\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}\right)\left(\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}-k\right)\le\frac{1}{4}\frac{(K-k)^2}{kK}\end{aligned}$$

and

$$\begin{aligned}0&\le\ln\left(\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}\right)-\frac{\langle w(A)\ln f(A)x,x\rangle}{\langle w(A)x,x\rangle}\qquad(6.20)\\&\le\frac{1}{4}(K-k)\,\Psi_{-\ln}\left(\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle};k,K\right)\le\frac{1}{4}\frac{(K-k)^2}{kK}\end{aligned}$$

for any $x\in H$ *with* $\langle w(A)x,x\rangle\neq0$.

(iii) *We have the inequalities*

$$\begin{aligned}0&\le\ln\left(\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}\right)-\frac{\langle w(A)\ln f(A)x,x\rangle}{\langle w(A)x,x\rangle}\qquad(6.21)\\&\le2\max\left\{\frac{K-\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}}{K-k},\frac{\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}-k}{K-k}\right\}\ln\left(\frac{k+K}{2\sqrt{kK}}\right)\end{aligned}$$

and

$$0\le\ln\left(\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}\right)-\frac{\langle w(A)\ln f(A)x,x\rangle}{\langle w(A)x,x\rangle}\le\ln\left(\frac{k+K}{2\sqrt{kK}}\right)^2\qquad(6.22)$$

for any $x\in H$ *with* $\langle w(A)x,x\rangle\neq0$.

(iv) *We also have the inequalities*

$$0\le\ln\left(\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}\right)-\frac{\langle w(A)\ln f(A)x,x\rangle}{\langle w(A)x,x\rangle}\qquad(6.23)$$

$$\leq \frac{1}{2}\Psi_{-\ln}\left(\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle};k,K\right)\frac{\left\langle\left|f(A)-\frac{\langle f(A)w(A)x,x\rangle}{\langle w(A)x,x\rangle}1_H\right|x,x\right\rangle}{\langle w(A)x,x\rangle}$$
$$\leq \frac{1}{2}\Psi_{-\ln}\left(\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle};k,K\right)$$
$$\times\left[\frac{\left\langle f^2(A)w(A)x,x\right\rangle}{\langle w(A)x,x\rangle}-\left(\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle}\right)^2\right]^{\frac{1}{2}}$$
$$\leq \frac{1}{4}\Psi_{-\ln}\left(\frac{\langle w(A)f(A)x,x\rangle}{\langle w(A)x,x\rangle};k,K\right)(K-k)$$

for any $x \in H$ with $\langle w(A)x,x\rangle \neq 0$.

6.3 Some Examples

If we choose $w(t) = 1$ and $f(t) = t$ with $t \in [k, K] \subset [0, \infty)$ then we get from Corollary 14 that

$$0 \leq \left\langle A^p x,x\right\rangle - \langle Ax,x\rangle^p \leq p\left[\left\langle A^p x,x\right\rangle - \left\langle A^{p-1}x,x\right\rangle\langle Ax,x\rangle\right] \tag{6.24}$$
$$\leq \frac{1}{2}p\left(K^{p-1}-k^{p-1}\right)\langle|A-\langle Ax,x\rangle 1_H|x,x\rangle$$
$$\leq \frac{1}{2}p\left(K^{p-1}-k^{p-1}\right)\left[\left\langle A^2x,x\right\rangle-\langle Ax,x\rangle^2\right]^{\frac{1}{2}}$$
$$\leq \frac{1}{4}p\left(K^{p-1}-k^{p-1}\right)(K-k),$$

$$0 \leq \left\langle A^p x,x\right\rangle - \langle Ax,x\rangle^p \tag{6.25}$$
$$\leq \frac{(K-\langle Ax,x\rangle)(\langle Ax,x\rangle-k)}{K-k}\sup_{t\in(k,K)}\Psi_p(t;k,K)$$
$$\leq p\frac{K^{p-1}-k^{p-1}}{K-k}(K-\langle Ax,x\rangle)(\langle Ax,x\rangle-k)$$
$$\leq \frac{1}{4}p\left(K^{p-1}-k^{p-1}\right)(K-k),$$

$$0 \leq \left\langle A^p x,x\right\rangle - \langle Ax,x\rangle^p \leq \frac{1}{4}(K-k)\Psi_p(\langle Ax,x\rangle;k,K) \tag{6.26}$$

$$\leq \frac{1}{4} p \left(K^{p-1} - k^{p-1} \right) (K - k) ,$$

$$0 \leq \left\langle A^p x, x \right\rangle - \langle Ax, x \rangle^p \tag{6.27}$$
$$\leq 2 \max \left\{ \frac{K - \langle Ax, x \rangle}{K - k}, \frac{\langle Ax, x \rangle - k}{K - k} \right\} \left[\frac{k^p + K^p}{2} - \left(\frac{k + K}{2} \right)^p \right],$$

$$0 \leq \left\langle A^p x, x \right\rangle - \langle Ax, x \rangle^p \leq 2 \left[\frac{k^p + K^p}{2} - \left(\frac{k + K}{2} \right)^p \right] \tag{6.28}$$

and

$$0 \leq \left\langle A^p x, x \right\rangle - \langle Ax, x \rangle^p \leq \frac{1}{2} \Psi_p \left(\langle Ax, x \rangle ; k, K \right) \left\langle \left| A - \langle Ax, x \rangle 1_H \right| x, x \right\rangle \tag{6.29}$$
$$\leq \frac{1}{2} \Psi_p \left(\langle Ax, x \rangle ; k, K \right) \left[\left\langle A^2 x, x \right\rangle - \langle Ax, x \rangle^2 \right]^{\frac{1}{2}}$$
$$\leq \frac{1}{4} \Psi_p \left(\langle Ax, x \rangle ; k, K \right) (K - k)$$

for any $x \in H$, $\|x\| = 1$.

If we choose $w(t) = t^q$, $q \neq 0$ and $f(t) = t$ with $t \in [k, K] \subset [0, \infty)$ then we get from Corollary 14 that

$$0 \leq \frac{\left\langle A^{p+q} x, x \right\rangle}{\langle A^q x, x \rangle} - \left(\frac{\left\langle A^{q+1} x, x \right\rangle}{\langle A^q x, x \rangle} \right)^p \tag{6.30}$$
$$\leq p \left[\frac{\left\langle A^{p+q} x, x \right\rangle}{\langle A^q x, x \rangle} - \frac{\left\langle A^{p+q-1} x, x \right\rangle}{\langle A^q x, x \rangle} \frac{\left\langle A^{q+1} x, x \right\rangle}{\langle A^q x, x \rangle} \right]$$
$$\leq \frac{1}{2} p \left(K^{p-1} - k^{p-1} \right) \frac{\left\langle \left| A - \frac{\left\langle A^{q+1} x, x \right\rangle}{\langle A^q x, x \rangle} 1_H \right| x, x \right\rangle}{\langle A^q x, x \rangle}$$
$$\leq \frac{1}{2} p \left(K^{p-1} - k^{p-1} \right) \left[\frac{\left\langle A^{q+2} x, x \right\rangle}{\langle A^q x, x \rangle} - \left(\frac{\left\langle A^{q+1} x, x \right\rangle}{\langle A^q x, x \rangle} \right)^2 \right]^{\frac{1}{2}}$$
$$\leq \frac{1}{4} p \left(K^{p-1} - k^{p-1} \right) (K - k) ,$$

$$0 \le \frac{\langle A^{p+q}x, x\rangle}{\langle A^q x, x\rangle} - \left(\frac{\langle A^{q+1}x, x\rangle}{\langle A^q x, x\rangle}\right)^p \tag{6.31}$$
$$\le \frac{\left(K - \frac{\langle A^{q+1}x,x\rangle}{\langle A^q x,x\rangle}\right)\left(\frac{\langle A^{q+1}x,x\rangle}{\langle A^q x,x\rangle} - k\right)}{K-k} \sup_{t\in(k,K)} \Psi_p(t; k, K)$$
$$\le p\frac{K^{p-1} - k^{p-1}}{K-k}\left(K - \frac{\langle A^{q+1}x, x\rangle}{\langle A^q x, x\rangle}\right)\left(\frac{\langle A^{q+1}x, x\rangle}{\langle A^q x, x\rangle} - k\right)$$
$$\le \frac{1}{4}p\left(K^{p-1} - k^{p-1}\right)(K-k),$$

$$0 \le \frac{\langle A^{p+q}x, x\rangle}{\langle A^q x, x\rangle} - \left(\frac{\langle A^{q+1}x, x\rangle}{\langle A^q x, x\rangle}\right)^p \le \frac{1}{4}(K-k)\Psi_p\left(\frac{\langle A^{q+1}x, x\rangle}{\langle A^q x, x\rangle}; k, K\right) \tag{6.32}$$
$$\le \frac{1}{4}p\left(K^{p-1} - k^{p-1}\right)(K-k),$$

$$0 \le \frac{\langle A^{p+q}x, x\rangle}{\langle A^q x, x\rangle} - \left(\frac{\langle A^{q+1}x, x\rangle}{\langle A^q x, x\rangle}\right)^p \tag{6.33}$$
$$\le 2\max\left\{\frac{K - \frac{\langle A^{q+1}x,x\rangle}{\langle A^q x,x\rangle}}{K-k}, \frac{\frac{\langle A^{q+1}x,x\rangle}{\langle A^q x,x\rangle} - k}{K-k}\right\}\left[\frac{k^p + K^p}{2} - \left(\frac{k+K}{2}\right)^p\right],$$

$$0 \le \frac{\langle A^{p+q}x, x\rangle}{\langle A^q x, x\rangle} - \left(\frac{\langle A^{q+1}x, x\rangle}{\langle A^q x, x\rangle}\right)^p \le 2\left[\frac{k^p + K^p}{2} - \left(\frac{k+K}{2}\right)^p\right] \tag{6.34}$$

and

$$0 \le \frac{\langle A^{p+q}x, x\rangle}{\langle A^q x, x\rangle} - \left(\frac{\langle A^{q+1}x, x\rangle}{\langle A^q x, x\rangle}\right)^p \tag{6.35}$$
$$\le \frac{1}{2}\Psi_p\left(\frac{\langle A^{q+1}x, x\rangle}{\langle A^q x, x\rangle}; k, K\right)\frac{\left\langle\left|A - \frac{\langle A^{q+1}x,x\rangle}{\langle A^q x,x\rangle}1_H\right| x, x\right\rangle}{\langle A^q x, x\rangle}$$
$$\le \frac{1}{2}\Psi_p\left(\frac{\langle A^{q+1}x, x\rangle}{\langle A^q x, x\rangle}; k, K\right)\left[\frac{\langle A^{q+2}x, x\rangle}{\langle A^q x, x\rangle} - \left(\frac{\langle A^{q+1}x, x\rangle}{\langle A^q x, x\rangle}\right)^2\right]^{\frac{1}{2}}$$

$$\leq \frac{1}{4}\Psi_p\left(\frac{\left\langle A^{q+1}x, x\right\rangle}{\langle A^q x, x\rangle}; k, K\right)(K-k)$$

for any $x \in H \setminus \{0\}$.

If we choose $w(t) = 1$ and $f(t) = t$ with $t \in [k, K] \subset [0, \infty)$ then we get from Corollary 15 that

$$\begin{aligned} 0 &\leq \ln\langle Ax, x\rangle - \langle \ln Ax, x\rangle \leq \left\langle A^{-1}x, x\right\rangle \langle Ax, x\rangle - 1 \\ &\leq \frac{1}{2}\frac{K-k}{kK}\langle |A - \langle Ax, x\rangle 1_H| x, x\rangle \\ &\leq \frac{1}{2}\frac{K-k}{kK}\left[\left\langle A^2x, x\right\rangle - \langle Ax, x\rangle^2\right]^{\frac{1}{2}} \leq \frac{1}{4}\frac{(K-k)^2}{kK}, \end{aligned} \tag{6.36}$$

$$\begin{aligned} 0 &\leq \ln\langle Ax, x\rangle - \langle \ln Ax, x\rangle \\ &\leq \frac{(K - \langle Ax, x\rangle)(\langle Ax, x\rangle - k)}{K-k}\sup_{t\in(k,K)}\Psi_{-\ln}(t; k, K) \\ &\leq \frac{1}{Kk}(K - \langle Ax, x\rangle)(\langle Ax, x\rangle - k) \leq \frac{1}{4}\frac{(K-k)^2}{kK}, \end{aligned} \tag{6.37}$$

$$\begin{aligned} 0 &\leq \ln\langle Ax, x\rangle - \langle \ln Ax, x\rangle \leq \frac{1}{4}(K-k)\Psi_{-\ln}(\langle Ax, x\rangle; k, K) \\ &\leq \frac{1}{4}\frac{(K-k)^2}{kK}, \end{aligned} \tag{6.38}$$

$$\begin{aligned} 0 &\leq \ln\langle Ax, x\rangle - \langle \ln Ax, x\rangle \\ &\leq 2\max\left\{\frac{K - \langle Ax, x\rangle}{K-k}, \frac{\langle Ax, x\rangle - k}{K-k}\right\}\ln\left(\frac{k+K}{2\sqrt{kK}}\right), \end{aligned} \tag{6.39}$$

$$0 \leq \ln\langle Ax, x\rangle - \langle \ln Ax, x\rangle \leq \ln\left(\frac{k+K}{2\sqrt{kK}}\right)^2 \tag{6.40}$$

and

$$\begin{aligned} 0 &\leq \ln\langle Ax, x\rangle - \langle \ln Ax, x\rangle \\ &\leq \frac{1}{2}\Psi_{-\ln}(\langle Ax, x\rangle; k, K)\langle |f(A) - \langle Ax, x\rangle 1_H| x, x\rangle \end{aligned} \tag{6.41}$$

$$\leq \frac{1}{2}\Psi_{-\ln}\left(\langle Ax,x\rangle ;k,K\right)\left[\left\langle A^{2}x,x\right\rangle -\langle Ax,x\rangle^{2}\right]^{\frac{1}{2}}$$
$$\leq \frac{1}{4}\Psi_{-\ln}\left(\langle Ax,x\rangle ;k,K\right)(K-k)$$

for any $x\in H$ with $\|x\|=1$.

If we choose $w(t)=t^{q}$, $q\neq 0$ and $f(t)=t$ with $t\in[k,K]\subset[0,\infty)$ then we get from Corollary 15 that

$$0\leq \ln\left(\frac{\left\langle A^{q+1}x,x\right\rangle}{\langle A^{q}x,x\rangle}\right)-\frac{\langle A^{q}\ln Ax,x\rangle}{\langle A^{q}x,x\rangle} \tag{6.42}$$
$$\leq \frac{\left\langle A^{q-1}x,x\right\rangle}{\langle A^{q}x,x\rangle}\frac{\left\langle A^{q+1}x,x\right\rangle}{\langle A^{q}x,x\rangle}-1\leq \frac{1}{2}\frac{K-k}{kK}\frac{\left\langle\left|A-\frac{\langle A^{q+1}x,x\rangle}{\langle A^{q}x,x\rangle}1_{H}\right|x,x\right\rangle}{\langle A^{q}x,x\rangle}$$
$$\leq \frac{1}{2}\frac{K-k}{kK}\left[\frac{\left\langle A^{q+2}x,x\right\rangle}{\langle A^{q}x,x\rangle}-\left(\frac{\left\langle A^{q+1}x,x\right\rangle}{\langle A^{q}x,x\rangle}\right)^{2}\right]^{\frac{1}{2}}\leq \frac{1}{4}\frac{(K-k)^{2}}{kK},$$

$$0\leq \ln\left(\frac{\left\langle A^{q+1}x,x\right\rangle}{\langle A^{q}x,x\rangle}\right)-\frac{\langle A^{q}\ln Ax,x\rangle}{\langle A^{q}x,x\rangle} \tag{6.43}$$
$$\leq \frac{\left(K-\frac{\langle A^{q+1}x,x\rangle}{\langle A^{q}x,x\rangle}\right)\left(\frac{\langle A^{q+1}x,x\rangle}{\langle A^{q}x,x\rangle}-k\right)}{K-k}\sup_{t\in(k,K)}\Psi_{-\ln}(t;k,K)$$
$$\leq \frac{1}{Kk}\left(K-\frac{\left\langle A^{q+1}x,x\right\rangle}{\langle A^{q}x,x\rangle}\right)\left(\frac{\left\langle A^{q+1}x,x\right\rangle}{\langle A^{q}x,x\rangle}-k\right)\leq \frac{1}{4}\frac{(K-k)^{2}}{kK},$$

$$0\leq \ln\left(\frac{\left\langle A^{q+1}x,x\right\rangle}{\langle A^{q}x,x\rangle}\right)-\frac{\langle A^{q}\ln Ax,x\rangle}{\langle A^{q}x,x\rangle} \tag{6.44}$$
$$\leq \frac{1}{4}(K-k)\,\Psi_{-\ln}\left(\frac{\left\langle A^{q+1}x,x\right\rangle}{\langle A^{q}x,x\rangle};k,K\right)\leq \frac{1}{4}\frac{(K-k)^{2}}{kK},$$

$$0\leq \ln\left(\frac{\left\langle A^{q+1}x,x\right\rangle}{\langle A^{q}x,x\rangle}\right)-\frac{\langle A^{q}\ln Ax,x\rangle}{\langle A^{q}x,x\rangle} \tag{6.45}$$

$$\leq 2\max\left\{\frac{K-\frac{\langle A^{q+1}x,x\rangle}{\langle A^q x,x\rangle}}{K-k},\frac{\frac{\langle A^{q+1}x,x\rangle}{\langle A^q x,x\rangle}-k}{K-k}\right\}\ln\left(\frac{k+K}{2\sqrt{kK}}\right),$$

$$0\leq \ln\left(\frac{\langle A^{q+1}x,x\rangle}{\langle A^q x,x\rangle}\right)-\frac{\langle A^q\ln Ax,x\rangle}{\langle A^q x,x\rangle}\leq \ln\left(\frac{k+K}{2\sqrt{kK}}\right)^2, \tag{6.46}$$

and

$$0\leq \ln\left(\frac{\langle A^{q+1}x,x\rangle}{\langle A^q x,x\rangle}\right)-\frac{\langle A^q\ln Ax,x\rangle}{\langle A^q x,x\rangle} \tag{6.47}$$

$$\leq \frac{1}{2}\Psi_{-\ln}\left(\frac{\langle A^{q+1}x,x\rangle}{\langle A^q x,x\rangle};k,K\right)\frac{\left\langle\left|A-\frac{\langle A^{q+1}x,x\rangle}{\langle A^q x,x\rangle}1_H\right|x,x\right\rangle}{\langle A^q x,x\rangle}$$

$$\leq \frac{1}{2}\Psi_{-\ln}\left(\frac{\langle A^{q+1}x,x\rangle}{\langle A^q x,x\rangle};k,K\right)\left[\frac{\langle A^{q+2}x,x\rangle}{\langle A^q x,x\rangle}-\left(\frac{\langle A^{q+1}x,x\rangle}{\langle A^q x,x\rangle}\right)^2\right]^{\frac{1}{2}}$$

$$\leq \frac{1}{4}\Psi_{-\ln}\left(\frac{\langle A^{q+1}x,x\rangle}{\langle A^q x,x\rangle};k,K\right)(K-k)$$

for any $x\in H\setminus\{0\}$.

References

1. S.M. Ali, S.D. Silvey, A general class of coefficients of divergence of one distribution from another. J. R. Stat. Soc. Ser. B **28**, 131–142 (1966)
2. M. Beth Bassat, f-entropies, probability of error and feature selection. Inf. Control **39**, 227–242 (1978)
3. A. Bhattacharyya, On a measure of divergence between two statistical populations defined by their probability distributions. Bull. Calcutta Math. Soc. **35**, 99–109 (1943)
4. I. Burbea, C.R. Rao, On the convexity of some divergence measures based on entropy function. IEEE Trans. Inf. Theory **28**(3), 489–495 (1982)
5. P. Cerone, S.S. Dragomir, A refinement of the Grüss inequality and applications. Tamkang J. Math. **38**(1), 37–49 (2007). Preprint RGMIA Res. Rep. Coll. **5**(2), Article 14 (2002). http://rgmia.org/papers/v5n2/RGIApp.pdf
6. C.H. Chen, *Statistical Pattern Recognition* (Hoyderc Book Co., Rocelle Park, 1973)
7. X.L. Cheng, J. Sun, A note on the perturbed trapezoid inequality. J. Inequal. Pure Appl. Math. **3**(2), Article 29 (2002)
8. C.K. Chow, C.N. Lin, Approximating discrete probability distributions with dependence trees. IEEE Trans. Inf. Theory **14**(3), 462–467 (1968)
9. I. Csiszár, Information-type measures of difference of probability distributions and indirect observations. Studia Math. Hung. **2**, 299–318 (1967)

10. I. Csiszár, On topological properties of f-divergences. Studia Math. Hung. **2**, 329–339 (1967)
11. I. Csiszár, J. Körner, *Information Theory: Coding Theorem for Discrete Memoryless Systems* (Academic, New York, 1981)
12. S.S. Dragomir, A converse result for Jensen's discrete inequality via Grüss' inequality and applications in information theory. An. Univ. Oradea Fasc. Mat. **7**, 178–189 (1999/2000)
13. S.S. Dragomir, On a reverse of Jessen's inequality for isotonic linear functionals. J. Inequal. Pure Appl. Math. **2**(3), Article 36 (2001)
14. S.S. Dragomir, A Grüss type inequality for isotonic linear functionals and applications. Demonstratio Math. **36**(3), 551–562 (2003). Preprint RGMIA Res. Rep. Coll. **5**(Supplement), Article 12 (2002). http://rgmia.org/papers/v5e/GTIILFApp.pdf
15. S.S. Dragomir, A converse inequality for the Csiszár Φ-divergence. Tamsui Oxf. J. Math. Sci. **20**(1), 35–53 (2004). Preprint in S.S. Dragomir (ed.), *Inequalities for Csiszár f-Divergence in Information Theory*. RGMIA Monographs (Victoria University, 2000). http://rgmia.org/papers/Csiszar/Csiszar.pdf
16. S.S. Dragomir, Bounds for the normalized Jensen functional. Bull. Aust. Math. Soc. **74**(3), 471–476 (2006)
17. S.S. Dragomir, Superadditivity of the Jensen integral inequality with applications. Miskolc Math. Notes **13**(2), 303–316 (2012). Preprint RGMIA Res. Rep. Coll. **14**, Article 75 (2011). http://rgmia.org/papers/v14/v14a75.pdf
18. S.S. Dragomir, *Operator Inequalities of the Jensen, Čebyšev and Grüss Type*. Springer Briefs in Mathematics (Springer, New York, 2012), xii+121 pp. ISBN: 978-1-4614-1520-6
19. S.S. Dragomir, *Operator Inequalities of Ostrowski and Trapezoidal Type*. Springer Briefs in Mathematics (Springer, New York, 2012), x+112 pp. ISBN: 978-1-4614-1778-1
20. S.S. Dragomir, Some reverses of the Jensen inequality with applications. Bull. Aust. Math. Soc. **87**(2), 177–194 (2013). Preprint RGMIA Res. Rep. Coll. **14**, Article 72 (2011). http://rgmia.org/papers/v14/v14a72.pdf
21. S.S. Dragomir, Reverses of the Jensen inequality in terms of first derivative and applications. Acta Math. Vietnam. **38**(3), 429–446 (2013). Preprint RGMIA Res. Rep. Coll. **14**, Article 71 (2011). http://rgmia.org/papers/v14/v14a71.pdf
22. S.S. Dragomir, Jensen type weighted inequalities for functions of selfadjoint and unitary operators. Ital. J. Pure Appl. Math. **32**, 247–264 (2014)
23. S.S. Dragomir, A refinement and a divided difference reverse of Jensen's inequality with applications. Rev. Colomb. Mat. **50**(1), 17–39 (2016). Preprint RGMIA Res. Rep. Coll. **14**, Article 74 (2011). http://rgmia.org/papers/v14/v14a74.pdf
24. S.S. Dragomir, Weighted reverse inequalities of Jensen type for functions of selfadjoint operators. Transylv. J. Math. Mech. **8**(1), 29–44 (2016). Preprint RGMIA Res. Rep. Coll. **18**, Article 110 (2015). http://rgmia.org/papers/v18/v18a110.pdf
25. S.S. Dragomir, N.M. Ionescu, Some converse of Jensen's inequality and applications. Rev. Anal. Numér. Théor. Approx. **23**(1), 71–78 (1994)
26. T. Furuta, J. Mićić Hot, J. Pečarić, Y. Seo, *Mond-Pečarić Method in Operator Inequalities. Inequalities for Bounded Selfadjoint Operators on a Hilbert Space* (Element, Zagreb, 2005)
27. D.V. Gokhale, S. Kullback, *Information in Contingency Tables* (Marcel Decker, New York, 1978)
28. J.H. Havrda, F. Charvat, Quantification method classification process: concept of structural α-entropy. Kybernetika **3**, 30–35 (1967)
29. E. Hellinger, Neue Bergrüirdung du Theorie quadratisher Formerus von uneudlichvieleu Veränderlicher. J. für Reine Augeur. Math. **36**, 210–271 (1909)
30. G. Helmberg, *Introduction to Spectral Theory in Hilbert Space* (Wiley, New York, 1969)
31. H. Jeffreys, An invariant form for the prior probability in estimating problems. Proc. R. Soc. Lond. A **186**, 453–461 (1946)
32. T.T. Kadota, L.A. Shepp, On the best finite set of linear observables for discriminating two Gaussian signals. IEEE Trans. Inf. Theory **13**, 288–294 (1967)
33. T. Kailath, The divergence and Bhattacharyya distance measures in signal selection. IEEE Trans. Commun. Technol. **COM-15**, 52–60 (1967)

34. J.N. Kapur, A comparative assessment of various measures of directed divergence. Adv. Manag. Stud. **3**, 1–16 (1984)
35. D. Kazakos, T. Cotsidas, A decision theory approach to the approximation of discrete probability densities. IEEE Trans. Pattern Anal. Mach. Intell. **1**, 61–67 (1980)
36. S. Kullback, R.A. Leibler, On information and sufficiency. Ann. Math. Stat. **22**, 79–86 (1951)
37. J. Lin, Divergence measures based on the Shannon entropy. IEEE Trans. Inf. Theory **37**(1), 145–151 (1991)
38. C.A. McCarthy, c_p. Isr. J. Math. **5**, 249–271 (1967)
39. M. Mei, The theory of genetic distance and evaluation of human races. Jpn. J. Hum. Genet. **23**, 341–369 (1978)
40. B. Mond, J. Pečarić, Convex inequalities in Hilbert space. Houst. J. Math. **19**, 405–420 (1993)
41. C.P. Niculescu, An extension of Chebyshev's inequality and its connection with Jensen's inequality. J. Inequal. Appl. **6**(4), 451–462 (2001)
42. E.C. Pielou, *Ecological Diversity* (Wiley, New York, 1975)
43. C.R. Rao, Diversity and dissimilarity coefficients: a unified approach. Theor. Popul. Biol. **21**, 24–43 (1982)
44. A. Rényi, On measures of entropy and information, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1 (University of California Press, Berkeley, 1961), pp. 547–561
45. A.W. Roberts, D.E. Varberg, *Convex Functions* (Academic, New York, 1973)
46. A. Sen, *On Economic Inequality* (Oxford University Press, London, 1973)
47. B.D. Sharma, D.P. Mittal, New non-additive measures of relative information. J. Comb. Inf. Syst. Sci. **2**(4), 122–132 (1977)
48. H. Shioya, T. Da-Te, A generalisation of Lin divergence and the derivative of a new information divergence. Electron. Commun. Jpn. **78**(7), 37–40 (1995)
49. S. Simić, On a global upper bound for Jensen's inequality. J. Math. Anal. Appl. **343**, 414–419 (2008)
50. I.J. Taneja, Generalised information measures and their applications. http://www.mtm.ufsc.br/~taneja/bhtml/bhtml.html
51. H. Theil, *Economics and Information Theory* (North-Holland, Amsterdam, 1967)
52. H. Theil, *Statistical Decomposition Analysis* (North-Holland, Amsterdam, 1972)
53. F. Topsoe, Some inequalities for information divergence and related measures of discrimination. Preprint RGMIA Res. Rep. Coll. **2**(1), 85–98 (1999)
54. I. Vajda, *Theory of Statistical Inference and Information* (Kluwer Academic Publishers, Dordrecht, 1989)

Ordering Structures and Their Applications

Gabriele Eichfelder and Maria Pilecka

1 Introduction

Order theory is one of the basic subjects in mathematics. Every time we need to compare two elements of a given space with each other, we use special ordering structures. Already in the space $\mathbb{R}^2$ relations between two elements are not so intuitive any more as it is for two elements of a real line. Defining which element is smaller or greater than the other one leads to a binary relation which may be a pre-order, partial order or total order depending on its properties. Ordering structures are closely related to cones in the considered space. There are direct connections between properties of a binary relation and a corresponding cone. The basic definitions and results on this topic are depicted in Sects. 2 and 3.

In vector optimization, we need not only to compare two elements of a space with each other but above all to find the best element of a set of candidates. Such best elements can be defined in many different ways, but all of these definitions are based on the given ordering structures in the so called "objective space". In set optimization even more general problems compared to vector optimization are considered. Here, the aim is to find a best set in the set of candidate sets, which is a subset of the power set of a given space. This leads to ordering structures which are weaker than the ordering structures in vector optimization in the following sense: if we consider a partially ordered space and generalize the binary relations to set

G. Eichfelder (✉)
Institute for Mathematics, Technische Universität Ilmenau, Ilmenau, Germany
e-mail: gabriele.eichfelder@tu-ilmenau.de

M. Pilecka
Institut für Numerische Mathematik und Optimierung, Technische Universität Bergakademie Freiberg, Freiberg, Germany
e-mail: maria.pilecka@math.tu-freiberg.de

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization
and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_9

relations, we may obtain only a pre-order or even only a reflexive or transitive binary relation on the power set of this space. Examples of such relations are given in Sect. 5.

Due to the connections between binary relations and cones, one may presume that some special cones imply specific properties of the ordering structures. We discuss this topic for two classes of cones: polyhedral cones (Sect. 3.3) and Bishop-Phelps (BP) cones (Sect. 3.2). The BP cones possess a very rich structure and allow for instance to define a scalarization which can be used to solve vector optimization problems.

In the optimization theory, besides vector or set optimization another scientific topic—cone programming, where some restrictions are defined based on a considered cone, make use of order theory. This is true especially for the copositive or semidefinite optimization problems where the variables are matrices which are completely positive or positive semidefinite, respectively, see [7, 23] and references therein.

In many applications of vector optimization, e.g. in intensity-modulated radiation therapy, modelling a binary relation in a usual way, i.e. assuming that the preferred or dominated directions form a cone which is independent of an element of the given space, seems not to be suitable. For such problems, the so called variable ordering structure is considered. In this case, an ordering relation may depend on the considered elements of the space, see Sect. 4. Due to this fact such a relation may not possess some important properties as for instance transitivity in general. However, it is still possible to derive conditions and algorithms helping to find the best elements of the given set [19, 20]. The idea of a binary relation depending on an element of the space is also used in decision theory [72], where the so called domination sets are introduced, see Sect. 3.4.

2 Pre- and Partial Orders

In this section, we give the most basic definitions from linear algebra which we need to define orders in a linear space. We start by introducing one of the most fundamental notions in order theory—a binary relation.

Definition 1 Assume that S is a set and $\leq$ is a subset of $S \times S$. Then $\leq$ is called a *binary relation* on S. If we have $(a, b) \in \leq$, we use the notation $a \leq b$.

Some properties which may characterize binary relations are given in the following definition.

Definition 2 Let S be a nonempty set with a binary relation $\leq$. Let $a,\ b,\ c \in S$ be arbitrarily chosen. The binary relation $\leq$ is said to be

(i) *reflexive* if $a \leq a$.
(ii) *transitive* if $a \leq b$ and $b \leq c$ imply $a \leq c$.
(iii) *symmetric* if $a \leq b$ implies $b \leq a$.

(iv) *antisymmetric* if $a \leq b$ and $b \leq a$ imply $a = b$.
(v) *complete (total)* if $a \leq b$ or $b \leq a$.

If we consider any set S, then the subset relation is a reflexive, transitive and antisymmetric binary relation defined on the power set of S.

Depending on the satisfied properties, we distinguish the following binary relations.

Definition 3 The binary relation $\leq$ on the set S is said to be

(i) a *pre-order* if it is reflexive and transitive.
(ii) a *partial order* if it is reflexive, transitive and antisymmetric or in other words, if it is a pre-order that is antisymmetric.
(iii) a *total order* if it is reflexive, transitive, antisymmetric, and complete, and hence, it is a partial order which is complete.
(iv) an *equivalence relation* if it is reflexive, transitive and symmetric.

When the relation $\leq$ is a pre-order/a partial/a total order, we say that S is a *pre-ordered/partially/totally ordered set.*

Hence, the subset relation is a partial order on the power set of a set S. It is not a total order as it is not complete: we can easily find two sets $A, B \subseteq S$ satisfying $A \not\subseteq B$ and $B \not\subseteq A$ in general. This reveals an important property of both pre-ordered and partial ordered sets. Namely, two arbitrary elements of these sets cannot be compared in terms of the binary relation in general. For a comprehensive survey on the relevant properties of binary relations and the notions of a partial, pre- and total order we refer to [27, Chapter 2]. For the definitions introduced above see for instance the books [48, 66].

Example 1 Let us consider the set $S = \mathbb{R}^2$. For real numbers $x, y \in \mathbb{R}$, $x \leq y$ denotes in the following $y - x \in \mathbb{R}_+$ as usual.

- $\leq_1 := \{(x, y) \in \mathbb{R}^2 \times \mathbb{R}^2 \mid x_1 \leq y_1\}$ is a pre-order since it is reflexive, transitive, but neither antisymmetric nor symmetric. It is also complete since we can compare any two elements of $\mathbb{R}^2$ using this relation with each other.
- $\leq_2 := \{(x, y) \in \mathbb{R}^2 \times \mathbb{R}^2 \mid x_1 \leq y_1 \wedge x_2 \leq y_2\}$ is a partial order since it is reflexive, transitive, and antisymmetric but not complete and not symmetric. This partial order is also called the natural order since it intuitively generalizes the usual total order $\leq$ on $\mathbb{R}$. It is also called componentwise order.
- $\leq_{lex} := \{(x, y) \in \mathbb{R}^2 \times \mathbb{R}^2 \mid x_1 < y_1 \vee (x_1 = y_1 \wedge x_2 \leq y_2)\}$ is a total order which is not symmetric. It is called the lexicographic order.
- $\sim_1 := \{(x, y) \in \mathbb{R}^2 \times \mathbb{R}^2 \mid x \leq_1 y \wedge y \leq_1 x\}$ is an equivalence relation which is obviously neither antisymmetric nor complete.

Example 2 Let S be the set of all functions $f : \mathbb{R}^2 \to \mathbb{R}$. Then

$$\leq_F := \{(f, g) \in S \times S \mid \forall x \in \mathbb{R}^2 : \ f(x) \leq g(x)\}$$

is a partial order which is not symmetric and not complete.

For the relations $\leq_*$ in Example 1 where $* \in \{1, 2, lex\}$, we can easily define so called strict binary relations by $<_* := \{(x, y) \in \mathbb{R}^2 \times \mathbb{R}^2 \mid x \leq_* y \wedge x \neq y\}$. These relations are obviously not reflexive.

The above definitions do not present all possible properties of binary relations considered in the literature. To give one example, in [62, Def. 5.55] and [63, Def. 1], Mordukhovich defines a general nonreflexive preference relation $\prec$ on a subset S of a topological space Y by a binary relation which is *locally satiated* around $\bar{z} \in S$, i.e. for an arbitrary element $z \in S$ in a neighbourhood of $\bar{z}$, it is required that $z \in \operatorname{cl}\{u \in Y \mid u \prec z\}$ (where cl denotes the closure), and *almost transitive*, that means that for $z, v, w \in Y$, if $z \prec w$ and $v \in \operatorname{cl}\{u \in Y \mid u \prec z\}$ holds, then we have $v \prec w$. A vast majority of common strict binary relations (e.g. $<_1$ and $<_2$) are covered by this definition. However, the relation $<_{lex}$ is not a preference relation in this sense. Let us think of the elements $(0, 0), (0, 1), (0, 2) \in \mathbb{R}^2$. Then we have $(0, 0) <_{lex} (0, 1)$ and $(0, 2) \in \operatorname{cl}\left\{u \in \mathbb{R}^2 \mid u <_{lex} (0, 0)\right\} = \left\{u \in \mathbb{R}^2 \mid u_1 \leq 0\right\}$. However, $(0, 2) \not<_{lex} (0, 1)$ holds which shows that this relation is not almost transitive.

In decision theory, a binary relation is used in order to model mathematically the preferences of a decision maker over the alternatives. Such binary relation is often assumed to be complete. For further information on this application of order theory, we refer to [34]. Other ideas related to decision theory will be described in Sect. 3.4.

3 Ordering Structures in Linear Spaces

In case the set S in the definitions from the previous section is a real linear space, i.e. a vector space, then there is a helpful tool for working with pre-orders: they have a representation as a convex cone. We will introduce the relevant definitions in this section as well as their relation to the concepts of a pre- and a partial order. Moreover, we will examine some special binary relations which are introduced by Bishop-Phelps cones or by polyhedral cones. Finally, we present the concept of ordering sets. Throughout this section let Y be an arbitrary real linear space.

3.1 Pre-orders, Partial Orders and Cones

In this subsection, we introduce a collection of standard notions and basic results related to partially ordered real linear spaces which can be found in many introductory books on vector optimization as in the books by Göpfert and Nehse [32], Jahn [48] and Luc [60] as well as in books on analysis and applications, see for instance [44].

Definition 4 For $S = Y$, we say that the pre-order is

(i) *compatible with addition* if

$$\forall x, y, w, z \in Y : \; x \leq y \wedge w \leq z \;\Rightarrow\; x + w \leq y + z.$$

(ii) *compatible with multiplication* with a nonnegative real number if

$$\forall x, y \in Y: \;\; x \leq y, \; \alpha \in \mathbb{R}_+ \;\Rightarrow\; \alpha x \leq \alpha y.$$

If a pre-order satisfies the axioms (i) and (ii) from the definition above, then we say that it is *compatible with the linear structure of the space*. A real linear space equipped with a partial order which is compatible with the linear structure of the space is called a *partially ordered linear space*.

Example 3 The space $\mathbb{R}^m$ equipped with the componentwise partial order, defined for all $x, y \in \mathbb{R}^m$ by

$$x \leq y \quad \Leftrightarrow \quad x_i \leq y_i \;\text{ for all } i = 1, \ldots, m,$$

is a partially ordered linear space.

Recall that a set S is called *convex* if $\lambda \in [0, 1], \; x, y \in S$ imply

$$\lambda x + (1 - \lambda)\, y \in S.$$

We write $\mathrm{conv}(S)$ for the convex hull of a set S.

In order to discuss the relations between pre- and partial orders and convex cones, we introduce the following notations. We use here the definitions as given in [48].

Definition 5

(i) A nonempty set $K \subseteq Y$ is called a *cone* if it holds

$$y \in K, \; \lambda \in \mathbb{R}_+ \;\Rightarrow\; \lambda\, y \in K.$$

(ii) A set $S \subseteq Y$ is called *pointed* if

$$S \cap (-S) = \{0_Y\}.$$

(iii) A nonempty convex subset B of a convex cone $K \neq \{0_Y\}$ is called a *base* for K if each $y \in K \setminus \{0_Y\}$ has a unique representation of the form

$$y = \lambda\, b \;\text{ for some }\; \lambda \in \mathbb{R}_+ \setminus \{0\} \;\text{ and some }\; b \in B.$$

(iv) Let $S \subseteq Y$ be a nonempty set. The cone

$$\mathrm{cone}(S) := \{\lambda\, y \in Y \mid \lambda \in \mathbb{R}_+, \; y \in S\}$$

is called the *cone generated by S*.

Please note that also a slightly different definition is used in the literature for the concept of a cone: for instance in [14] a set $K \subseteq Y$ is called a cone if $y \in K, \; \lambda \in$

$\mathbb{R}_+ \setminus \{0\}$ implies $\lambda\, y \in K$, i.e. in that definition the zero does not have to be an element of a cone. In the definition which we use here, the zero is always included in a cone.

Using the above definition, we obviously deduce that for a base B of a convex cone K, it holds $\mathrm{cone}(B) = K$. A cone K with $K \neq \{0_Y\}$ and $K \neq Y$ is said to be a *nontrivial cone*. It is easy to show that any convex cone with a base is pointed. We have the following characterization of convex cones.

Lemma 1 *A cone $K \subseteq Y$ is convex if and only if $K + K \subseteq K$.*

Example 4 The sets $K_1 := \mathbb{R}^2_+$ and

$$K_2 := \{x \in \mathbb{R}^2 \mid 2x_1 + x_2 \leq 0,\ -0.5x_1 - x_2 \leq 0\}$$

are convex cones, while the set $K_1 \cup K_2$ is a cone which is not convex. All cones K_1, K_2, and $K_1 \cup K_2$ are pointed.

Now we are ready to give the relation between partial orders and convex cones:

Theorem 1

(i) If $\leq$ is a pre-order on Y which is compatible with the linear structure of the space, then the set

$$K := \{y \in Y \mid 0_Y \leq y\}$$

is a convex cone. If, in addition, $\leq$ is antisymmetric, then K is pointed.

(ii) If $K \subseteq Y$ is a convex cone, then the binary relation

$$\leq_K := \{(y, z) \in Y \times Y \mid z - y \in K\}$$

is a pre-order on Y which is compatible with the linear structure of the space. If, in addition, K is pointed, then $\leq_K$ is antisymmetric.

A convex cone which characterizes a partial order on a real linear space is called an *ordering cone*.

For the componentwise partial order in $\mathbb{R}^m$, cf. Example 3, the associated ordering cone is

$$\mathbb{R}^m_+ := \left\{y \in \mathbb{R}^m \mid y_i \geq 0 \text{ for all } i = 1, \ldots, m\right\}.$$

Some important cones are depicted in the following example, where we illustrate natural ordering cones in different spaces.

Example 5

(i) Let $\mathbb{S}_n$ be the space of symmetric matrices possessing n columns and rows. Then the set

$$\mathbb{S}^+_n := \left\{y \in \mathbb{S}_n \mid \forall\, x \in \mathbb{R}^n : x^\top y x \geq 0\right\}$$

of all positive semidefinite matrices is a convex cone in this space. The set

$$\left\{ y \in \mathbb{S}_n \mid \forall\, x \in \mathbb{R}^n_+ : x^\top y x \geq 0 \right\}$$

of all copositive matrices is another convex cone in this space.

(ii) Let Ω be any compact Hausdorff space. By $C(\Omega)$ we denote the real linear space of real-valued functions which are continuous on Ω. The natural ordering cone in this space is

$$C(\Omega)^+ := \{ y \in C(\Omega) \mid y(\omega) \geq 0 \text{ for all } \omega \in \Omega \} .$$

This cone is closed w.r.t. the supremum norm and it has a nonempty topological interior, see [48, Ex. 1.49].

(iii) Given a domain $\Omega \subseteq \mathbb{R}^n$, consider the real linear space of all (equivalence classes of) p-th power Lebesgue-integrable real-valued functions on Ω denoted by $L^p(\Omega)$ with $1 \leq p < \infty$. Then the natural ordering cone is

$$L^p(\Omega)^+ := \left\{ y \in L^p(\Omega) \mid y(\omega) \geq 0 \text{ f.a.a. } \omega \in \Omega \right\} .$$

The important property of this cone is the fact that the topological interior of this cone is empty, cf. [48, Ex. 1.51], which makes such assumption very restrictive while considering infinite dimensional spaces.

Engau studied in [27, 28] also properties which should be satisfied by a so-called constant preference structure (in opposition to a local or variable ordering structure, see Sect. 4) in the Euclidean space $Y = \mathbb{R}^m$. He stated as basic assumption that the binary relation should be compatible with scalar multiplication and with addition, similar as in Definition 4, and showed that it can thus be represented by a convex cone K, see our Theorem 1. In addition to that, he assumed *monotonicity*, i.e. for some $y \in \mathbb{R}^m$ and e^i the ith unit vector, it should hold that

$$y - e^i \leq y \text{ for all } i = 1, \ldots, m. \tag{1}$$

This implies for the ordering cone K the assumption $\mathbb{R}^m_+ \subseteq K$.

For many examinations in vector optimization, elements of the dual cone play an important role. Let Y be a partially ordered linear space with the ordering cone K. The dual cone of K is a subset of the dual space and introduces a pre-order in the dual space which is again compatible with the linear structure of the space. We write Y^* for the algebraic dual space of Y, i.e. for the space of all linear functions $l\colon Y \to \mathbb{R}$.

Definition 6 Let $K \subseteq Y$ be a convex cone.

(i) The cone

$$K^* := \{ l \in Y^* \mid l(y) \geq 0 \text{ for all } y \in K \}$$

is the *dual cone* of the cone K.

(ii) The set

$$K^{\#} := \{l \in Y^* \mid l(y) > 0 \text{ for all } y \in K \setminus \{0_Y\}\}$$

is the *quasi interior* of the dual cone K^*.

If we have $K = K^*$, we call K self-dual. Such a property characterizes some important cones such as $\mathbb{R}^n_+$ and $\mathbb{S}^+_n$.

Example 6 The dual cone of the cone of copositive matrices, see Example 5, w.r.t. the inner product

$$\langle y, z \rangle = \operatorname{trace}(yz) \;\; \forall\, y, z \in \mathbb{S}_n,$$

is the cone of completely positive matrices defined by

$$\left\{ y \in \mathbb{S}_n \mid y = zz^\top \text{ for some } z \in \mathbb{R}^{n \times m}_+,\ m \in \mathbb{N} \right\},$$

see for instance [12].

Note that the quasi interior of the dual cone is a superset of the algebraic interior of this cone under not too strong assumptions, see [48, Lem. 1.25]. Additionally, there is an interesting characterization of the base of a cone $K \subseteq Y$ and the quasi interior of the corresponding dual cone, which we use in the next subsection.

Lemma 2 ([48, Lem. 1.28]) *Let $K \subseteq Y$ be a nontrivial convex cone. Then for every $l \in K^{\#}$ the set*

$$B := \{y \in K \mid l(y) = 1\}$$

is a base of K.

3.2 Bishop-Phelps Cones

In this subsection, we study in more detail a special class of convex cones, the so called Bishop-Phelps (BP) cones. Bishop-Phelps cones have been introduced by Bishop and Phelps in 1962 in [6] and are characterized by a rich and useful mathematical structure. For instance, they allow the formulation of special scalarization functionals in vector optimization. Results on the usage of Bishop-Phelps cones in optimization and on their properties can be found in [37, 47]. We base our presentations mainly on these two papers.

As the definition of Bishop-Phelps cones requires a norm, we assume throughout this subsection that $(Y, \|\cdot\|)$ is a real normed space. We denote the topological dual space, i.e. the space of all continuous linear functionals $f\colon Y \to \mathbb{R}$, by Y^*. Here, $\|\cdot\|_*$ denotes the induced norm in Y^*, where

$$\|\phi\|_* := \sup_{y \neq 0_Y} \frac{|\phi(y)|}{\|y\|} \quad \text{for all} \quad \phi \in Y^*.$$

A Bishop-Phelps cone is defined by an element ϕ from the dual space Y^* as follows:

Definition 7 For an arbitrary continuous linear functional $\phi \in Y^*$, the cone

$$C(\phi) := \{y \in Y \mid \|y\| \leq \phi(y)\} \tag{2}$$

is called *Bishop-Phelps cone* (BP cone).

A cone $K \subseteq Y$ for which a functional $\phi \in Y^*$ and a norm $\|\cdot\|$ equivalent to the norm of the space exist such that K can be written as in (2) is called *representable as a BP cone*.

According to [47], the cone

$$C_p := \{y \in \mathbb{R}^n \mid \|(y_1, \ldots, y_{n-1})\|_p \leq y_n\} \subseteq \mathbb{R}^n$$

with $\|\cdot\|_p$ an l_p norm with $p \geq 1$ or $p = \infty$ is representable as a BP cone. It holds $C_p = C(\sqrt[p]{2}e_n)$ for $p \in [1, \infty)$ and $e_n := (0, \ldots, 0, 1)^\top$, and $C_\infty = C(e_n)$. The cone C_2 is the well-known Lorentz cone (see Example 11). Thus the Lorentz cone

$$C_2 = \{y \in \mathbb{R}^3 \mid \|(y_1, y_2)\|_2 \leq y_3\}$$

in $Y = \mathbb{R}^3$ has the representation

$$C_2 = \left\{y \in \mathbb{R}^3 \mid \|y\|_2 \leq \sqrt{2}(0, 0, 1)y\right\}.$$

Note that the original concept of a BP cone is slightly different from the one introduced in Definition 7. Originally, for an arbitrary $\vartheta \in Y^*$ with $\|\vartheta\|_* = 1$ and some scalar $t \in (0, 1)$, the cone

$$\{y \in Y \mid t\,\|y\| \leq \vartheta(y)\} \tag{3}$$

is considered. It is easy to see that any cone satisfying (3) is also a BP cone in the sense of Definition 7 with $\phi = \vartheta/t$, and any BP cone in the sense of Definition 7 with $\|\phi\|_* > 1$ satisfies (3) with $\vartheta = \phi/\|\phi\|_*$ and $t = 1/\|\phi\|_*$. Hence, Definition 7, which we use in the following, generalizes the notion given in (3) also to the case when $t \geq 1$ is satisfied.

Example 7 Let $Y = \mathbb{R}^2$ and assume that the space is equipped with the Manhattan norm. Then for instance for $(\phi_1, \phi_2) = (1, 1)$, we have $C(\phi_1, \phi_2) = \mathbb{R}^2_+$. Assume

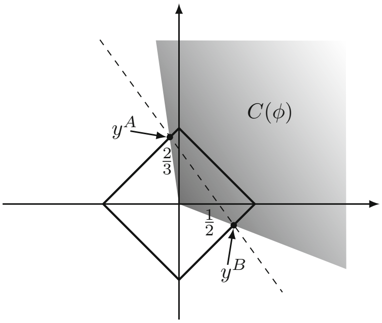


Fig. 1 BP cone $C(\phi_1, \phi_2)$ of Example 7 for $\phi_1 = 2$ and $\phi_2 = 3/2$, as well as the unit ball w.r.t. the Manhattan norm and (in dashed line) the set $\{(y_1, y_2) \in \mathbb{R}^2 \mid (\phi_1, \phi_2)^\top (y_1, y_2) = 1\}$, cf. [22]

$\phi_1, \phi_2 \geq 1$, then $\mathbb{R}^2_+ \subseteq C(\phi_1, \phi_2)$, $(0, 1/\phi_2) \in C(\phi_1, \phi_2)$, $(1/\phi_1, 0) \in C(\phi_1, \phi_2)$ and

$$C(\phi_1, \phi_2) = \text{cone conv } \left(\{y^A, y^B\}\right)$$

with

$$y^A := \left(\frac{1-\phi_2}{\phi_1+\phi_2}, \frac{1+\phi_1}{\phi_1+\phi_2}\right)^\top \text{ and } y^B := \left(\frac{1+\phi_2}{\phi_1+\phi_2}, \frac{1-\phi_1}{\phi_1+\phi_2}\right)^\top ,$$

see Fig. 1.

The most important properties of BP cones are collected in the following lemma.

Lemma 3 *Let $\phi \in Y^*$ be given.*

(i) *$C(\phi)$ is a closed, pointed and convex cone.*
(ii) *If $\|\phi\|_* > 1$, then $C(\phi)$ is nontrivial; if $\|\phi\|_* < 1$ then $C(\phi) = \{0_Y\}$.*
(iii) *If $\|\phi\|_* = 1$, then $C(\phi) = \{y \in Y \mid \|y\| = \phi(y)\}$; if, additionally, Y is a reflexive Banach space, then $C(\phi)$ is nontrivial.*
(iv) *$\{y \in Y \mid \|y\| < \phi(y)\} \subseteq int(C(\phi))$, where $int(A)$ denotes the topological interior of a set A.*
If $\|\phi\|_* > 1$, then the interior of $C(\phi)$ is nonempty and

$$int(C(\phi)) = \{y \in Y \mid \|y\| < \phi(y)\}.$$

(v) *$\phi \in C(\phi)^\#$.*
(vi) *If the set $\{y \in C(\phi) \mid \phi(y) = 1\}$ is nonempty, then it is a closed and bounded base for the cone $C(\phi)$.*
(vii) *$C(\phi)^* = cl(cone(B(\phi, 1)))$ with $B(\phi, 1) := \{y^* \in Y^* \mid \|y^* - \phi\|_* \leq 1\}$.*

As one can see, BP cones have a rich structure. They are always pointed convex cones and thus introduce by Theorem 1 a partial order on Y which is compatible with the linear structure of the space. Moreover, the interior of such cones can also be described easily provided $\|\phi\|_* > 1$. Hence, it is important to know which cones do belong to the class of BP cones. This result goes back to Petschke [64].

Theorem 2 *A nontrivial cone $K \subseteq Y$ is representable as a BP cone if and only if K is a convex cone with a closed and bounded base. In the Euclidean space $Y = \mathbb{R}^n$ a convex cone $K \subseteq Y$ is representable as a BP cone if and only if K is closed and pointed.*

Note that there exist important classes of cones which do not have bounded bases. For instance, the bases of the natural ordering cones in the spaces l^p and L^p for $1 < p < \infty$ are not bounded, see [5, 11]. Additional examples for BP cones can be found in Sect. 4.3.

The following example illustrates how the special structure of the ordering cone can be used for scalarization results in vector optimization.

Example 8 Let $S \subseteq Y$ be a nonempty set and let Y be partially ordered by the convex cone K with

$$K := C(\phi) = \{y \in Y \mid \|y\| \leq \phi(y)\}$$

for some $\phi \in Y^*$ with $\|\phi\|_* > 1$. By Lemma 3 (iv) we have $\text{int}(C(\phi)) = \{y \in Y \mid \|y\| < \phi(y)\} \neq \emptyset$. An element $\bar{y} \in S$ is denoted to be a weakly efficient element of S in case it holds

$$(\{\bar{y}\} - \text{int}(K)) \cap S = \emptyset.$$

One can define a functional $\xi_{\bar{y}} \colon Y \to \mathbb{R}$ by

$$\xi_{\bar{y}}(y) = \phi(y - \bar{y}) + \|y - \bar{y}\|,$$

cf. [22]. Then it holds $y \in \{\bar{y}\} - \text{int}(K)$ with $K = C(\phi)$ if and only if $\|\bar{y} - y\| < \phi(\bar{y} - y)$, i.e. if and only if $\xi_{\bar{y}}(y) < 0$. Hence, $\bar{y}$ is a weakly efficient element of S if and only if it holds for all $y \in S$

$$\xi_{\bar{y}}(y) \geq 0 = \xi_{\bar{y}}(\bar{y}).$$

Bishop-Phelps cones are related to augmented dual cones which extend the usual definition of a dual cone. In the following definition we use the notion of quasi interior of the dual cone, see Definition 6.

Definition 8 ([55]) Let $K \subseteq Y$ be a closed pointed convex cone.

(i) The set

$$K^{a*} := \left\{(\phi, \alpha) \in K^{\#} \times \mathbb{R}_+ \mid \phi(y) - \alpha\|y\| \geq 0 \;\text{ for all } y \in K\right\}$$

is called *augmented dual cone*.

(ii) Let $\operatorname{int}(K) \neq \emptyset$. The set

$$K^{a\circ} := \Big\{(\phi, \alpha) \in K^{\#} \times \mathbb{R}_+ \mid \phi(y) - \alpha\|y\| > 0 \text{ for all } y \in \operatorname{int}(K)\Big\}$$

is called *weak augmented dual cone*.

(iii) The set

$$K^{a\#} := \Big\{(\phi, \alpha) \in K^{\#} \times \mathbb{R}_+ \mid \phi(y) - \alpha\|y\| > 0 \text{ for all } y \in K \setminus \{0_Y\}\Big\}$$

is called *augmented quasi interior of the dual cone*.

Note that the weak augmented dual cone, despite of its name, is not a cone in the sense of Definition 5 as it does not include the zero. It holds $K^{a\#} \subseteq K^{a\circ} \subseteq K^{a*}$ for any closed pointed convex cone K. For instance for $K = \mathbb{R}^n_+$ and $Y = \mathbb{R}^n$, we have

$$K^{a*} = \{(\phi, \alpha) \in \operatorname{int}(\mathbb{R}^n_+) \times \mathbb{R}_+ \mid \phi_i \geq \alpha,\ i = 1, \ldots, n\},$$

see [55, Ex. 4.7].

For the definition of the augmented dual cones, it is crucial that the quasi interior of the dual cone $K^{\#}$ is nonempty. For instance the Krein-Rutman theorem, below as cited in [48, Thm 3.38] (see also [44]), gives conditions ensuring that:

Theorem 3 (Krein-Rutman Theorem) *In a real separable normed space* $(Y, \|\cdot\|)$ *with a closed and pointed convex cone* $K \subseteq Y$ *the quasi interior* $K^{\#}$ *of the topological dual cone is nonempty.*

Thus, in the finite dimensional Euclidean space $Y = \mathbb{R}^n$ for any closed pointed convex cone, the quasi interior of the dual cone is nonempty and according to [75], it equals the interior of the dual cone, see also [39, p. 199]. The pointedness of K is essential as for any convex cone K, the condition $K^{\#} \neq \emptyset$ already implies the pointedness of K [48, Lem. 1.27]. For BP cones $C(\phi)$, the quasi interior of the dual cone is nonempty according to Lemma 3 (v).

We obtain the following relation of BP cones and elements of the augmented dual cones [20, Lem. 1.21]:

Lemma 4 *Let* $\phi \in Y^*$ *define a BP cone* $C(\phi) = \{y \in Y \mid \|y\| \leq \phi(y)\}$. *Then*

$$(\phi, \alpha) \in (C(\phi))^{a*} \ \textit{for all}\ \ \alpha \in [0, 1]$$

and $(\phi, 0) \in (C(\phi))^{a\#}$.

If $\|\phi\|_* > 1$, *then*

$$(\phi, \alpha) \in (C(\phi))^{a\circ} \ \textit{for all}\ \ \alpha \in [0, 1].$$

BP cones are included in the class of supernormal cones (see the definition below) and for closed pointed supernormal cones, the augmented dual cones are known to be nontrivial, cf. [20]. We start by recalling the definition of supernormal cones:

Definition 9 ([33]) Let $K \subseteq Y$ be a nontrivial convex cone. K is said to be *supernormal* (or nuclear or has the angle property) if there exists $\phi \in Y^*$ such that

$$K \subseteq \{y \in Y \mid \|y\| \leq \phi(y)\} = C(\phi). \tag{4}$$

For instance, in $\mathbb{R}^n$ every pointed convex cone is supernormal and every BP cone is a supernormal cone [46, p. 635].

Lemma 5 ([20]) *Let $K \subseteq Y$ be a nontrivial closed pointed convex cone. Then the following statements are equivalent:*

(i) There exists $(\phi, \alpha) \in K^{a*}$ with $\alpha \neq 0$.
(ii) K is supernormal.

There is the following relation between supernormal cones and BP cones:

Lemma 6 ([46]) *Let $K \subseteq Y$ be a nontrivial closed pointed convex cone. Then the following is equivalent:*

(i) K is representable as a BP cone.
(ii) K is supernormal.

Hence, the augmented dual cone of some nontrivial closed pointed convex cone contains elements (ϕ, α) with $\alpha \neq 0$ if and only if the cone is representable as a BP cone. The elements of the augmented dual cones can also be used for scalarization results in vector optimization, see [55]:

Example 9 Let $S \subseteq Y$ be a nonempty set and let Y be partially ordered by the closed pointed convex cone K with $\text{int}(K) \neq \emptyset$. Let $(\phi, \alpha) \in K^{a\circ}$ and assume that it holds

$$\bar{y} \in \text{argmin}\{\phi(y) + \alpha \|y\| \mid y \in S\}.$$

Then $\bar{y}$ is a weakly efficient element of S, i.e. $(\{\bar{y}\} - \text{int}(K)) \cap S = \emptyset$, cf. Example 8. To see this, assume that there exists $y \in S$ with $\bar{y} - y \in \text{int}(K)$. By the definition of the weak augmented dual cone, we get

$$\phi(\bar{y} - y) - \alpha\|\bar{y} - y\| > 0$$

and thus by the triangle inequality

$$\phi(\bar{y}) - \phi(y) > \alpha\|y\| - \alpha\|\bar{y}\|$$

or

$$\phi(\bar{y}) + \alpha\|\bar{y}\| > \phi(y) + \alpha\|y\|,$$

which is a contradiction to the minimality of $\bar{y}$.

3.3 Polyhedral Cones

Another special class of cones possessing many useful properties are polyhedral cones. We present some of these properties in this subsection.

Definition 10 Let Y be a real locally convex linear space. A cone $K \subseteq Y$ is *polyhedral* if there exist $y^{*i} \in Y^*$, $i = 1, \dots, p$, such that

$$K = \left\{ y \in Y \mid \langle y^{*i}, y \rangle \geq 0,\ i = 1, \dots, p \right\}.$$

In other words a polyhedral cone can be represented as the intersection of a finite number of closed half spaces or it is a solution set of a homogeneous system of inequalities.

Such cones are convex and closed, and hence, they induce a pre-order on Y. If we consider a linear operator $L : Y \to \mathbb{R}^p$ given by $L := \left(y^{*1}, y^{*2}, \dots, y^{*p} \right)$ such that $L(y) = \left(y^{*1}(y), y^{*2}(y), \dots, y^{*p}(y) \right)^\top$, then the polyhedral cone

$$K = \{ y \in Y \mid 0_{\mathbb{R}^p_+} \leq_{\mathbb{R}^p_+} L(y) \}$$

is pointed if and only if L is injective. In this case K induces a partial order on Y.

Assume that K is the cone from Definition 10. The dual cone to K is

$$K^* = \left\{ \sum_{i=1}^{p} \lambda_i y^{*i} \in Y^* \mid \lambda_i \geq 0, i = 1, \dots, p \right\}.$$

Moreover, for a polyhedral cone K, we have $K = K^{**}$.

Let us now consider finite dimensional spaces, i.e. we set $Y = \mathbb{R}^n$. Then we can write a polyhedral cone $K \subseteq \mathbb{R}^n$ as

$$K = \{ x \in \mathbb{R}^n \mid 0_{\mathbb{R}^p} \leq_{\mathbb{R}^p_+} \overline{K} x \} \tag{5}$$

for some matrix $\overline{K} \in \mathbb{R}^{p \times n}$. If $\overline{K}$ has rank n then the cone is pointed. We get for arbitrary points $a, b \in \mathbb{R}^n$

$$a \leq_K b \ \Leftrightarrow\ b - a \in K \ \Leftrightarrow\ \overline{K}(b - a) \in \mathbb{R}^p_+ \ \Leftrightarrow\ \overline{K} a \leq_{\mathbb{R}^p_+} \overline{K} b.$$

This observation may be also used for solving vector optimization problems, see [15]. We illustrate this issue in the following example [15, Ex. 1.19].

Example 10 Let the Euclidean space $\mathbb{R}^3$ be partially ordered by

$$K := \{ x \in \mathbb{R}^3 \mid 0_{\mathbb{R}^4} \leq_{\mathbb{R}^4_+} \begin{pmatrix} 1 & 0 & 1 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \\ 0 & 1 & 1 \end{pmatrix} x \}.$$

The cone

$$K = \{x \in \mathbb{R}^3 \mid -x_3 \leq x_1 \leq x_3,\ -x_3 \leq x_2 \leq x_3,\ x_3 \geq 0\}$$

is a pyramid with apex in the origin. The aim may now be to find the best element of a set $\Omega \subseteq \mathbb{R}^3$ w.r.t. the relation induced by K. Equivalently, we may search for the best element of the set $S = \{(x_1 + x_3, -x_1 + x_3, -x_2 + x_3, x_2 + x_3) \in \mathbb{R}^4 \mid x \in \Omega\}$ w.r.t. the natural order in $\mathbb{R}^4$.

Let us use the notation from the example and discussion above. It is important to note that due to the structure of the natural ordering cone it may be simpler to determine the best element of the set S compared to searching for the best x within the set Ω w.r.t. K. However, such transformation of the considered problem may also have a drawback. It may happen that the dimension of the space $\mathbb{R}^p$ is much greater than the dimension of the space $\mathbb{R}^n$, what makes the new problem in $\mathbb{R}^p$ more complex.

We have the following fundamental result of Weyl which relates polyhedral and finitely generated cones. For a recent proof, see [53].

Lemma 7 *A cone $K \subseteq \mathbb{R}^n$ is polyhedral if and only if K is finitely generated, i.e. there are $y^i \in \mathbb{R}^n$, $i = 1, \ldots, k$, $k \in \mathbb{N}$ such that we have*

$$K = \left\{ y \in \mathbb{R}^n \mid y = \sum_{i=1}^{k} \lambda_i y^i, \lambda_i \geq 0, i = 1, \ldots, k \right\}.$$

Next we give an example of a cone which is not finitely generated.

Example 11 Let $Y = \mathbb{R}^{n+1}$ and let $\|\cdot\|_2$ denote the Euclidean norm in $\mathbb{R}^n$. Then

$$C := \left\{ (x, t) \in \mathbb{R}^{n+1} \mid \|x\|_2 \leq t \right\}$$

is the Lorentz (second-order/ice-cream) cone, which is not finitely generated.

Using the formula from the previous lemma, we may deduce also another characterization of a polyhedral cone K. For such a cone there is a matrix $P \in \mathbb{R}^{n \times k}$ with

$$K = \{y \in \mathbb{R}^n \mid y = Px,\ x \in \mathbb{R}^k_+\}. \tag{6}$$

If the matrix P has rank k then the cone K is pointed. For a cone K as in (6) we have for arbitrary points $a, b \in \mathbb{R}^k$

$$a \leq_{\mathbb{R}^k_+} b \ \Leftrightarrow\ b - a \in \mathbb{R}^k_+ \ \Leftrightarrow\ P(b - a) \in K \ \Leftrightarrow\ Pa \leq_K Pb,$$

where the second equivalence only holds in case the rank of P is k.

The following example illustrates the usefulness of the structure of finitely generated cones for the notion of set-semidefinite matrices, cf. [23].

Example 12 Let $Y = \mathbb{R}^n$ and $K \subseteq \mathbb{R}^n$ be some convex cone. The K-semidefinite cone is defined by

$$C^K := \{A \in \mathbb{S}_n \mid y^\top Ay \geq 0 \text{ for all } y \in K\}.$$

For $K = \mathbb{R}^n$ we obtain the cone of positive semidefinite matrices and for $K = \mathbb{R}^n_+$ the cone of copositive matrices, see Example 5. If B is a base of K, then it holds

$$C^K = \{A \in \mathbb{S}_n \mid y^\top Ay \geq 0 \text{ for all } y \in B\}.$$

If K is a polyhedral cone with a representation as in (6), then

$$C^K = \{A \in \mathbb{S}_n \mid P^\top AP \text{ is copositive}\}.$$

For testing a matrix on copositivity much more numerical methods have been developed than for testing on general set-semidefiniteness, see for instance [7–9].

A detailed study of properties of polyhedral cones in finite dimensions can be found in [4, 31]. See also [73] for examples of preference modelling using special polyhedral cones.

3.4 Ordering Sets

Instead of defining a binary relation with the aid of a cone (see Theorem 1), it is possible to use a more general set. In literature we can find different concepts of such generalizations. In this subsection we present among others the idea of the improvement and the domination sets.

Given a general ordering set $\Theta \subseteq Y$ for any $y^1, y^2 \in Y$, a binary relation $\leq_\Theta$ is defined by:

$$y^1 \leq_\Theta y^2 \quad\Leftrightarrow\quad y^2 - y^1 \in \Theta.$$

Such idea was studied in [67] for strongly star-shaped conic sets. This class of sets contains not only all convex cones but also nonconvex sets. To give one simple example: a finite union of closed convex cones $K_i,\ i \in I := \{1, \ldots, l\},\ l \in \mathbb{N}$, i.e.

$$K := \bigcup_{i \in I} K_i,$$

satisfying the condition $\bigcap_{i \in I} \operatorname{int}(K_i) \neq \emptyset$, is a strongly star-shaped conic set. In this case, each cone K_i induces a pre-order $\leq_{K_i}$ on Y. Moreover, we can interpret the relation $\leq_K$, generated by the cone K, in the following way: $y^1 \leq_K y^2$ if and only if there exists $i \in I$ such that $y^1 \leq_{K_i} y^2$. Hence, $\leq_K$ is obviously not a pre-order if K is not convex. See [67] for the exact definition of a star-shaped conic set as well as properties of such binary relations.

Binary relations which are no pre-orders are motivated by applications from mathematical economics, see [61] and references therein as cited in [67]. A different idea was introduced in [10], where the so-called improvement set is defined in finite dimensional spaces equipped with the partial order given by the natural ordering cone. This concept was generalized in [35] for infinite dimensional spaces and general cones as follows.

Definition 11 Let Y be equipped with a pre-order induced by a nontrivial cone $K \subseteq Y$. A nonempty set $E \subseteq Y$ is said to be an *improvement set* with respect to K if $0_Y \notin E$ and $E = E + K$ holds.

For each set $\emptyset \neq A \subseteq Y$ satisfying $A \cap (-K \setminus \{0_Y\}) = \emptyset$, the set $A + (K \setminus \{0_Y\})$ is an improvement set. Another simple example of an improvement set provides the interior of K if it is nonempty.

Improvement sets allow to generalize the definitions of the different best points of a given set depending on the choice of the improvement set E.

Definition 12 Let Ω be a nonempty set and $f : \Omega \to Y$ a vector valued map. Then $\bar{x} \in \Omega$ is denoted an *E-optimal solution* of the vector optimization problem

$$\min_{x \in \Omega} f(x),$$

if

$$(\{f(\bar{x})\} - E) \cap f(\Omega) = \emptyset.$$

For other binary relations based on ordering sets, see [62, Section 5.3], where also generalized optimality conditions for vector optimization problems are given.

The last concept which we like to describe in this subsection is domination set revealing a connection between decision theory and vector optimization based on [72]. Consider a decision maker (DM) who wants to choose an element x from the set of feasible decisions $\Omega \subseteq X$. The choice is based on values of a function $f : \Omega \to Y$ giving outcomes of the decisions. The best decision depends on the preferences of the DM within the set $f(\Omega)$. Assume now that the preference relation $\preceq$ illustrates the preferences of the DM in the following way: for $a, b \in Y$, $a \preceq b$ means that a is better than b. Then the best decisions $\bar{x}$ are those where $f(\bar{x})$ is contained in the set

$$\operatorname{Min}(f(\Omega)) := \{y \in f(\Omega) \mid \forall \tilde{y} \in f(\Omega) : \tilde{y} \preceq y \Rightarrow y \preceq \tilde{y}\}.$$

A decision of the DM may require a few steps characterized by different preference relations in each step. At the beginning of this decision process the set $\operatorname{Min}(f(\Omega))$ may contain more than one element. However, in the last step only one decision should be chosen. We proceed now with the definition of a domination set.

Definition 13 Let $\preceq$ be a binary relation on Y. Consider $D(y) := \{d \in Y \mid y+d \preceq y\}$ for each $y \in Y$. If there is $D \subseteq Y$ satisfying $D(y) = D$ for all $y \in Y$, then D is the *domination set* of $\preceq$.

From this definition, we obtain the following properties of such a preference relation.

Proposition 1 ([72]) *Let $\preceq$ be a binary relation on Y and let $D \subseteq Y$. D is a domination set of $\preceq$ if and only if*

$$\forall y^1, y^2 \in Y : \; y^2 \preceq y^1 \;\Leftrightarrow\; y^2 \in \{y^1\} + D.$$

There exists a domination set of $\preceq$ if and only if

$$\forall y^1, y^2, y \in Y : y^1 \preceq y^2 \Rightarrow (y^1 + y) \preceq (y^2 + y). \tag{7}$$

If D is a domination set of $\preceq$, it follows:

(a) *$\preceq$ is reflexive $\Leftrightarrow 0_Y \in D$.*
(b) *$\preceq$ is asymmetric (i.e. $y^1 \preceq y^2 \;\Rightarrow\; y^2 \not\preceq y^1$) $\Leftrightarrow D \cap (-D) = \emptyset$.*
(c) *$\preceq$ is antisymmetric $\Leftrightarrow D \cap (-D) = \{0_Y\}$.*
(d) *$\preceq$ is transitive $\Leftrightarrow D + D \subseteq D$.*
(e) *$\preceq$ fulfills the condition*

$$\forall y^1, y^2 \in Y \; \forall \lambda \in \mathbb{R}_+ \setminus \{0\} : y^1 \preceq y^2 \Rightarrow (\lambda y^1) \preceq (\lambda y^2), \tag{8}$$

if and only if $D \cup \{0_Y\}$ is a cone.
(f) *$\preceq$ is a transitive relation which satisfies condition (8) if and only if $D \cup \{0_Y\}$ is a convex cone.*
(g) *$\preceq$ is a partial order which fulfills condition (8) if and only if D is a pointed convex cone.*

Note that if a binary relation is a pre-order, the condition (7) is equivalent to the compatibility of this relation with addition, see Definition 4, which indeed may not be satisfied for general binary relations. If we consider a preference relation for eating cake during a coffee break, then three pieces of cake may be preferred to one. However, five pieces may not be preferred to three. Similarly, six pieces may not be preferred to two and hence, the condition (8) (compatibility with multiplication) may also not be fulfilled.

Using the domination sets, we may also define the best elements of the considered set in a following way.

Definition 14 Let $S, D \subseteq Y$. An element $\bar{y} \in S$ is called an *efficient element* of S w.r.t. D if

$$(\{\bar{y}\} - D) \cap S \subseteq \{\bar{y}\}.$$

The set of all efficient elements of S w.r.t. D is denoted by $\mathrm{Eff}(S, D)$.

Now we are ready to give a relation between the best decisions of a DM (with function values within the set Min($f(\Omega)$)) and the efficient elements of the set $f(\Omega)$, which are widely used in vector optimization, see [72].

Proposition 2 *Suppose that $\preceq$ is a preference relation on Y with the domination set D. Let $\Omega \subseteq X$ and $f : \Omega \to Y$ be given. Then*

$$\mathit{Min}\,(f(\Omega)) = \mathit{Eff}\,(f(\Omega), D \setminus (-D))$$

holds. Moreover, if $\preceq$ is asymmetric or antisymmetric, then we have

$$\mathit{Min}\,(f(\Omega)) = \mathit{Eff}\,(f(\Omega), D)\,.$$

In [72] also properties of the efficient set and the weak efficient set defined using a general domination set D are discussed based on results from [71] and [70]. Moreover, a scalarization concept for characterizing the (weakly) efficient elements by functionals with uniform sublevel sets is presented there as well.

Note that for $y \in Y$ the set $D(y)$ from Definition 13 is closely connected to the image of the ordering map at y introduced in the forthcoming section.

4 Variable Ordering Structures

Next to partially ordered linear spaces, i.e. linear spaces with a pointed convex cone which introduces a partial order, also other ordering concepts play an important role in several applications. We collect in this section some basic definitions, properties and results on variable ordering structures. More details as well as a literature survey on this topic can be found in the book [20]. As before, in this section, let Y be a real linear space.

4.1 Introduction to Variable Ordering Structures

In the last years, multiobjective optimization problems with a variable ordering structure have gained interest motivated by several applications in such different fields as economics or medical image registration, cf. [2, 17, 68, 69, 73]. The basic idea is that instead of one fixed ordering cone for the whole image space, an individual ordering cone is attached to each element of the space. This corresponds to the interpretation that the preferences for the decision making in the objective space and also the corresponding binary relation depend on the considered point in the objective space.

Recall that in multiobjective optimization one optimizes several objective functions at the same time. Thus one has to compare elements in a linear space (for instance in $\mathbb{R}^m$ in case one minimizes m objective functions at the same time). That the importance of criteria may change during the decision-making process and that it may depend on current objective values was already recognized by Karaskal and Michalowski in [54]. Wiecek discussed this issue and gave some examples of such a process in [73]. In [28], Engau examined the role of variable ordering structures in preference modeling. He gave the following example which motivates that an ordering cone and thus a pre-order might not always be appropriate to model a decision making problem. We have discussed this issue also already in Sect. 3.4.

Example 13 ([28]) Let $Y = \mathbb{R}^2$, and let the set

$$S = \{y \in \mathbb{R}^2 \mid y_1 + y_2 \geq 1,\ y_1 \geq 0,\ y_2 \geq 0\}$$

be given. Assume that $K \subseteq \mathbb{R}^2$ is a convex cone with $\mathbb{R}^2_+ \subseteq K$. The latter corresponds to the requirement of monotonicity, see (1). We define $k^1 := (-1, 1)$ and $k^2 = (1, -1)$ and search for the efficient elements of S, i.e. for those elements $\bar{y} \in S$ for which

$$(\{\bar{y}\} - K) \cap S \subseteq \{\bar{y}\},$$

cf. Definition 14.

In case $k^1 \in K$ and $k^2 \notin K$ hold, the only efficient element is $z^1 := (1, 0)$. In case $k^1 \notin K$ but $k^2 \in K$, then the only efficient element is $z^2 := (0, 1)$. If $k^1 \in K$ and $k^2 \in K$, then there is no efficient element at all. If we have $k^1 \notin K$ and $k^2 \notin K$, then all elements of the line segment $L := \{y \in \mathbb{R}^2 \mid y_1 + y_2 = 1, y_1 \in [0, 1]\}$ are efficient. In particular, it is not possible to define an ordering cone that excludes the two extreme points z^1 and z^2 while maintaining a set of efficient elements in the middle of the line segment L.

In case of a variable ordering structure and by defining optimal elements of a set by using the binary relation $\leq_1$ or $\leq_2$ as defined below, the drawback of Example 13 can be overcome, see Example 15.

A variable ordering structure is mathematically defined by a set-valued map which associates to each element y of the linear space Y an individual cone of preferred or of dominated directions $\mathscr{D}(y) \subseteq Y$. Based on this cone-valued map and depending on the interpretation as set of preferred or of dominated/deteriorating directions, one obtains two binary relations by

$$y \leq_1 z \ :\Leftrightarrow\ z - y \in \mathscr{D}(y) \tag{9}$$

and by

$$y \leq_2 z \ :\Leftrightarrow\ z - y \in \mathscr{D}(z). \tag{10}$$

In the literature one can find a large number of publications on theoretical results for vector optimization problems with a variable ordering structure. Also more general concepts are studied where $\mathscr{D}$ is assumed to be an arbitrary set-valued map. Some of these studies make use of the assumption that the zero is included in the boundary of each image set $\mathscr{D}(y)$ and that a direction $d \in Y$ exists with $\{\lambda d \in Y \mid \lambda > 0\} \subseteq \mathscr{D}(y)$ for all $y \in Y$. See also page 293 for a short discussion on this topic.

We start by the basic definition which clarifies what we mean by a variable ordering structure in this section.

Definition 15 Let $\mathscr{D}\colon Y \to 2^Y$ be a set-valued map with $\mathscr{D}(y)$ a nonempty convex cone for all $y \in Y$. If elements in the space Y are compared using the binary relation (9) or (10), then the cone-valued map $\mathscr{D}$ is called an *ordering map* and it is said that $\mathscr{D}$ defines a *variable ordering (structure)* on Y.

The following example illustrates the binary relations (9) and (10).

Example 14 Let $Y = \mathbb{R}^2$ be equipped with a variable ordering structure defined by the ordering map $\mathscr{D}\colon \mathbb{R}^2 \to 2^{\mathbb{R}^2}$ with

$$\mathscr{D}(y) = \begin{cases} \mathbb{R}^2_+ & \text{for all } \; y \in \mathbb{R}^2 \setminus \{(0,0)\} \\ \text{cone conv}\,(\{(1,1),(1,0)\}) & \text{if } \; y = (0,0). \end{cases}$$

Then for all $y, z \in \mathbb{R}^2$ with $y \neq 0_{\mathbb{R}^2}$

$$y \leq_1 z \quad \Leftrightarrow \quad y_i \leq z_i \;\text{ for }\; i = 1, 2.$$

For $y = 0_{\mathbb{R}^2}$ and arbitrary $z \in \mathbb{R}^2$, we obtain

$$\begin{aligned} y = 0_{\mathbb{R}^2} \leq_1 z \;&\Leftrightarrow\; z \in \text{cone conv}\,(\{(1,1),(1,0)\}) \\ &\Leftrightarrow\; z_1 \geq z_2 \geq 0. \end{aligned}$$

Hence,

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \leq_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 1 \end{pmatrix} \leq_1 \begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad \text{but} \quad \begin{pmatrix} 0 \\ 0 \end{pmatrix} \not\leq_1 \begin{pmatrix} 1 \\ 2 \end{pmatrix},$$

which shows that $\leq_1$ is not transitive.

For the second relation, relation $\leq_2$, we get for all $y, z \in \mathbb{R}^2$ with $z \neq 0_{\mathbb{R}^2}$

$$y \leq_2 z \quad \Leftrightarrow \quad y_i \leq z_i \;\text{ for }\; i = 1, 2.$$

For $z = 0_{\mathbb{R}^2}$ and arbitrary $y \in \mathbb{R}^2$, we obtain

$$\begin{aligned} y \leq_2 z = 0_{\mathbb{R}^2} \;&\Leftrightarrow\; -y \in \text{cone conv}\,(\{(1,1),(1,0)\}) \\ &\Leftrightarrow\; y_1 \leq y_2 \leq 0. \end{aligned}$$

Therefore, also for this relation, we can find an example showing that $\leq_2$ is not transitive as well.

Of course, if $\mathscr{D}(y) = K$ for all $y \in Y$ for some convex cone K, both binary relations (9) and (10) define the same pre-order on Y. However, in general the relations do not coincide with each other and do not define a pre-order as can be seen in the example above.

Let us assume that for a given relation $\leq_*$, if we have $a \leq_* b$, then a is preferred to b. The binary relation defined in (9) represents the idea of domination: all elements of the set

$$\{y\} + \mathscr{D}(y) \setminus \{0_Y\} = \{z \in Y \setminus \{y\} \mid y \leq_1 z\}$$

are considered to be worse (less preferred) than the element y and are thus dominated by y. The set $\mathscr{D}(y)$ collects the set of directions in which the elements are worse compared to y, i.e. the deteriorating directions:

$$\mathscr{D}(y) = \{d \in Y \mid y + d \text{ is worse than } y\} \cup \{0_Y\}.$$

The binary relation (10) corresponds to the concept of preference, as all elements of the set

$$\{y\} - \mathscr{D}(y) \setminus \{0_Y\} = \{z \in Y \setminus \{y\} \mid z \leq_2 y\}$$

are considered to be better or more preferred than the element y. Therefore, it is more natural to define in a first step a cone-valued map $\mathscr{P}\colon Y \to 2^Y$ with $\mathscr{P}(y)$ a convex cone for all $y \in Y$ and

$$\mathscr{P}(y) = \{d \in Y \mid y + d \text{ is preferred to } y\} \cup \{0_Y\}.$$

Then

$$y \leq_{\mathscr{P}} z \quad \Leftrightarrow \quad y \in \{z\} + \mathscr{P}(z)$$

for all $y, z \in Y$. By defining $\mathscr{D}(y) := -\mathscr{P}(y)$ for all $y \in Y$ we get a unified notation and the binary relation $\leq_2$ by

$$y \leq_2 z \quad \Leftrightarrow \quad y \in \{z\} - \mathscr{D}(z)$$

for all $y, z \in Y$.

Note that the underlying concepts are fundamentally different and that in general

$$\{d \in Y \mid y + d \text{ is worse than } y\} \neq -\{d \in Y \mid y + d \text{ is preferred to } y\}.$$

Finally, we give an example of a variable ordering structure for which the drawback of Example 13 can be overcome. In order to do this, we need to introduce the concept of optimal elements for sets based on variable ordering structures:

Remark 1 By using any of the binary relations $\leq_1$ or $\leq_2$, we may follow the well known definition from vector optimization and define an optimal element of a nonempty set $S \subseteq Y$: we say that $\bar{y} \in S$ is an optimal element of S if there is no $y \in S \setminus \{\bar{y}\}$ with $y \leq \bar{y}$. In case of $\leq = \leq_1$ we speak of a *nondominated element* of S w.r.t. the ordering map $\mathscr{D}$. In case of $\leq=\leq_2$ of a *minimal element* of S w.r.t. the ordering map $\mathscr{D}$. Thus, given $\mathscr{D}$, an element $\bar{y} \in S$ is said to be nondominated element of S w.r.t. $\mathscr{D}$, if there is no $y \in S$ with $y \neq \bar{y}$ and $\bar{y} \in \{y\} + \mathscr{D}(y)$. And $\bar{y} \in S$ is said to be a minimal element of S w.r.t. $\mathscr{D}$ if

$$(\{\bar{y}\} - \mathscr{D}(\bar{y})) \cap S = \{\bar{y}\}.$$

Example 15 Let $Y = \mathbb{R}^2$ and S as in Example 13. Also, we use k^1, k^2, z^1, and z^2 as defined there. Now we define an ordering map on $Y = \mathbb{R}^2$ in the following way:

$$\mathscr{D}(y) = \begin{cases} \mathbb{R}^2_+ & \text{if } y \notin \mathbb{R}^2_+, \\ \operatorname{cone}\operatorname{conv}(\mathbb{R}^2_+ \cup \{k^2\}) & \text{if } y \in \operatorname{cone}(\{\lambda z^2 + (1-\lambda)z^1\}) \\ & \text{for some } \lambda \in [0, 0.3], \\ \operatorname{cone}\operatorname{conv}(\mathbb{R}^2_+ \cup \{k^1\}) & \text{if } y \in \operatorname{cone}(\{\lambda z^2 + (1-\lambda)z^1\}) \\ & \text{for some } \lambda \in [0.7, 1], \\ \operatorname{cone}\operatorname{conv}(\{(-\lambda, 1)\} \cup \{(1, -\lambda)\}) & \text{if } y \in \operatorname{cone}(\{\lambda z^2 + (1-\lambda)z^1\}) \\ & \text{for some } \lambda \in (0.3, 0.7). \end{cases}$$

Then only the elements $\{\lambda z^2 + (1-\lambda)z^1 \in S \mid \lambda \in [0.3, 0.7]\}$ are nondominated elements of S.

A different ordering map for the set in Example 15 is proposed in Example 5 in [28] which leads to the following set of minimal elements of S:

$$\left\{ y \in S \mid y_1 + y_2 = 1,\ 1 - \frac{1}{2}\sqrt{2} < y_1 < \frac{1}{2}\sqrt{2} \right\}.$$

That ordering map has images which are so called ideal-symmetric convex cones and which are special classes of Bishop-Phelps cones. Such special ordering maps are discussed in Sect. 4.3.

Variable ordering structures play also an important role for the theory of consumer demand in economics:

Example 16 In the traditional theory of consumer demand, see [51, 52] by John and the references therein, one assumes that a consumer's choice is derived from maximizing the utility. It is differentiated between the local and the global preferences, where the local preferences in the space $Y = \mathbb{R}^m$ are expressed by the following: Let $y \in \mathbb{R}^m$ be given. A direction $d \in \mathbb{R}^m$ is preferred if

$$w(y)^\top d < 0,$$

non-preferred if $w(y)^\top d > 0$, and indifferent if $w(y)^\top d = 0$ with $w\colon \mathbb{R}^m \to \mathbb{R}^m$ some function. The concepts of [52] using the notation above reads as: to any element $y \in \mathbb{R}^m$ the cone of non-preferred elements, including the indifferent elements, is given by

$$\mathscr{D}(y) = \{d \in \mathbb{R}^m \mid w(y)^\top d \geq 0\},$$

which is a convex but not pointed cone. In fact, $\mathscr{D}(y)$ is a halfspace.

According to Allen [1] and Georgescu-Roegen [29, 30] a point $\bar{y} \in \mathbb{R}^m$ is defined to be an equilibrium position, if no direction away from $\bar{y}$ to any other alternative y is preferred, i.e. if

$$w(\bar{y})^\top (y - \bar{y}) \geq 0.$$

This corresponds to $y \in \{\bar{y}\} + \mathscr{D}(\bar{y})$ for all feasible y, i.e., using the notation from [16], $\bar{y}$ has to be a so-called strongly minimal element.

Additionally, to guarantee the equilibrium to be stable, the so-called principle of persisting nonpreferences, it is required [30, 52] that for any $y \in \mathbb{R}^m$ it holds that

$$w(y)^\top d \geq 0 \;\text{ implies }\; w(y+d)^\top d \geq 0\,.$$

Equivalently, this can be written as

$$d \in \mathscr{D}(y+d) \;\text{ for all }\; d \in \mathscr{D}(y).$$

This property corresponds to the map w being pseudomonotone. Additionally, w is assumed to be continuous and is then called a local preference representation on Y.

4.2 Basic Properties of Variable Ordering Structures

First we study the assumptions which guarantee that a variable ordering structure is a pre- or a partial order.

Lemma 8 ([19, Lem. 2.1])

(i) The relations defined in (9) and (10) are reflexive.

(ii) The binary relation $\leq_1$ defined in (9) is transitive if

$$\mathscr{D}(y+d) \subseteq \mathscr{D}(y) \text{ for all } y \in Y \text{ and for all } d \in \mathscr{D}(y). \tag{11}$$

If $\mathscr{D}(y)$ is algebraically closed for all $y \in Y$, then (11) also is necessary for the transitivity of $\leq_1$.

(iii) The binary relation $\leq_2$ defined in (10) is transitive if

$$\mathscr{D}(y-d) \subseteq \mathscr{D}(y) \text{ for all } y \in Y \text{ and for all } d \in \mathscr{D}(y). \tag{12}$$

If $\mathscr{D}(y)$ is algebraically closed for all $y \in Y$, then (12) also is necessary for the transitivity of $\leq_2$.

(iv) *The property given in Definition 4(i) (compatibility with addition) is satisfied by any of the two relations $\leq_1$ or $\leq_2$ if and only if $\mathscr{D}$ is a constant map.*

(v) *The property given in Definition 4(ii) (compatibility with nonnegative scalar multiplication) is satisfied by any of the two relations $\leq_1$ or $\leq_2$ if and only if*

$$\mathscr{D}(y) \subseteq \mathscr{D}(\alpha\, y) \ \textit{for all} \ \ y \in Y \ \ \textit{and for all} \ \ \alpha > 0. \tag{13}$$

(vi) *The relations defined in (9) and (10) are antisymmetric if $\mathscr{D}(Y) := \bigcup_{y \in Y} \mathscr{D}(y)$ is pointed.*

Proof

(i) The relations are both reflexive as the sets $\mathscr{D}(y)$ are assumed to be cones and thus $0_Y \in \mathscr{D}(y)$ for all $y \in Y$.

(ii) We first show that the condition (11) is sufficient. Let $x, y, z \in Y$ be arbitrarily given. As $x \leq_1 y$ and $y \leq_1 z$ correspond to $y - x \in \mathscr{D}(x)$ and $z - y \in \mathscr{D}(y)$, (11) implies $\mathscr{D}(y) \subseteq \mathscr{D}(x)$ and we get $z - x = (z - y) + (y - x) \in \mathscr{D}(y) + \mathscr{D}(x) \subseteq \mathscr{D}(x)$ and hence $x \leq_1 z$.

Next, we show that condition (11) is also necessary if $\mathscr{D}(y)$ is algebraically closed for all $y \in Y$. For that we assume $\leq_1$ to be transitive, but (11) does not hold. Then there exists some $x \in Y$ and some $d \in \mathscr{D}(x)$ as well as some

$$k \in \mathscr{D}(x + d) \setminus \{0_Y\} \ \text{ with } \ k \notin \mathscr{D}(x). \tag{14}$$

For all $s > 0$ we obtain $sk \in \mathscr{D}(x + d) \setminus \{0_Y\}$ and $sk \notin \mathscr{D}(x)$. We set

$$y := x + d \ \text{ and } \ z_s := y + sk = x + d + sk \ \text{ for all } \ s > 0.$$

Then $y - x = d \in \mathscr{D}(x)$ and $z_s - y = sk \in \mathscr{D}(x + d) = \mathscr{D}(y)$ for all $s > 0$. Because $\leq_1$ is transitive, it holds $z_s - x = d + sk \in \mathscr{D}(x)$ for all $s > 0$, i.e. $\frac{1}{s}d + k \in \mathscr{D}(x)$ for $s > 0$ implying, because $\mathscr{D}(x)$ is algebraically closed, $k \in \mathscr{D}(x)$ in contradiction to (14).

(iii) We first show that the condition (12) is sufficient. Let $x, y, z \in Y$ be arbitrarily given. As $x \leq_2 y$ and $y \leq_2 z$ correspond to $y - x \in \mathscr{D}(y)$ and $z - y \in \mathscr{D}(z)$, (12) implies $\mathscr{D}(y) \subseteq \mathscr{D}(z)$ and we get $z - x = (z - y) + (y - x) \in \mathscr{D}(z) + \mathscr{D}(y) \subseteq \mathscr{D}(z)$ and hence $x \leq_1 z$.

Next, we show that condition (12) is also necessary if $\mathscr{D}(y)$ is algebraically closed for all $y \in Y$. For that we assume $\leq_2$ is transitive, but (12) does not hold. Then there exists some $z \in Y$ and some $d \in \mathscr{D}(z)$ as well as some $k \in Y \setminus \{0_Y\}$ with

$$sk \in \mathscr{D}(z - d) \setminus \{0_Y\} \ \text{ and } \ sk \notin \mathscr{D}(z) \ \text{ for all } s > 0. \tag{15}$$

We set $y := z - d$ and $x_s := y - sk = z - d - sk$ for $s > 0$ and obtain $y - x_s = sk \in \mathscr{D}(z - d) = \mathscr{D}(y)$ and $z - y = d \in \mathscr{D}(z)$ for all $s > 0$. Because $\leq_2$ is transitive, it holds $z - x_s = d + sk \in \mathscr{D}(z)$ for all $s > 0$ implying $k \in \mathscr{D}(z)$ in contradiction to (15).

(iv) The property given in Definition 4(i) corresponds for both relations to the property $\mathscr{D}(y) + \mathscr{D}(z) \subseteq \mathscr{D}(y + z)$ for any $y, z \in Y$, i.e. to the subadditivity of the cone-valued map $\mathscr{D}$. Similar as in the proof of Lemma 2.21 in [18], we show that this implies that $\mathscr{D}$ is a constant map.

Let $y \in Y$ be arbitrarily chosen. By the subadditivity we have

$$\mathscr{D}(y) + \mathscr{D}(-y) \subseteq \mathscr{D}(0_Y)$$

and as $0_Y \in \mathscr{D}(-y)$ also

$$\mathscr{D}(y) \subseteq \mathscr{D}(0_Y) \subseteq \mathscr{D}(Y)\ \forall\ y \in Y$$

and hence

$$\mathscr{D}(0_Y) = \mathscr{D}(Y).$$

The subadditivity also implies $\mathscr{D}(y)+\mathscr{D}(0_Y) \subseteq \mathscr{D}(y)$ for all $y \in Y$, and hence

$$\mathscr{D}(Y) = \mathscr{D}(0_Y) \subseteq \mathscr{D}(y) \subseteq \mathscr{D}(Y)$$

which leads to the assertion.

(v) As $\mathscr{D}(y)$ is a cone for all $y \in Y$ it holds $\mathscr{D}(y) = \alpha\mathscr{D}(y)$ for all $\alpha > 0$ and thus the property given in Definition 4(ii) corresponds for both relations to the property $\mathscr{D}(y) \subseteq \mathscr{D}(\alpha\, y)$ for all $y \in Y$ and all $\alpha > 0$.

(vi) $y \leq_1 z$ and $z \leq_1 y$ are equivalent to $z \in \{y\} + \mathscr{D}(y)$ and $z \in \{y\} - \mathscr{D}(z)$, thus $z - y \in \mathscr{D}(Y) \cap (-\mathscr{D}(Y))$, i.e. $z = y$. Analogously for $\leq_2$.

Hence, we speak of a variable ordering (structure) or just variable order given by the ordering map $\mathscr{D}$, even though the binary relations $\leq_1$ and $\leq_2$ are neither transitive nor compatible with the linear structure of the space in general. By that we emphasize that the partial order defined by a convex cone K is replaced by a relation which is defined by the cone-valued map $\mathscr{D}$.

Next we give examples of ordering maps which satisfy some of the assumptions of Lemma 8.

Example 17 Let $Y = \mathbb{R}^2$ and consider the ordering map $\mathscr{D}\colon \mathbb{R}^2 \to 2^{\mathbb{R}^2}$ with

$$\mathscr{D}(y_1, y_2) := \begin{cases} \left\{ \begin{pmatrix} r\cos\varphi \\ r\sin\varphi \end{pmatrix} \in \mathbb{R}^2 \;\middle|\; r \geq 0,\ \varphi \in [0, \frac{\pi}{8}] \right\} & \text{if } y_1 \geq \frac{\pi}{2}, \\ \left\{ \begin{pmatrix} r\cos\varphi \\ r\sin\varphi \end{pmatrix} \in \mathbb{R}^2 \;\middle|\; r \geq 0,\ \varphi \in [0, \frac{\pi}{2} + \frac{\pi}{8} - y_1] \right\} & \text{if } y_1 \in (\frac{\pi}{8}, \frac{\pi}{2}), \\ \mathbb{R}^2_+ & \text{if } y_1 \leq \frac{\pi}{8}. \end{cases}$$

For an illustration of some of the cones $\mathscr{D}(y)$ see Fig. 2.

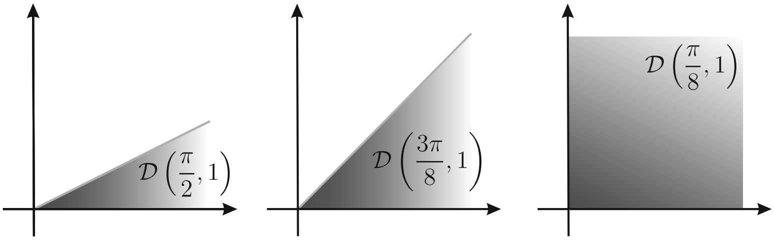

$\mathcal{D}\left(\frac{\pi}{2}, 1\right)$
$\mathcal{D}\left(\frac{3\pi}{8}, 1\right)$
$\mathcal{D}\left(\frac{\pi}{8}, 1\right)$

Fig. 2 Some of the cones $\mathscr{D}(y)$ of Example 17, cf. [20]

It holds $\mathscr{D}(y) \subseteq \mathbb{R}^2_+$ and $\mathscr{D}(y)$ is a closed pointed convex cone for all $y \in \mathbb{R}^2$. The cone $\mathscr{D}(y)$ depends on y_1 only, and for $z_1 \geq y_1$ for some $y, z \in \mathbb{R}^2$, it holds $\mathscr{D}(z) \subseteq \mathscr{D}(y)$. As for any $y \in \mathbb{R}^2$ and any $d \in \mathscr{D}(y)$ we have $d_1 \geq 0$ and thus $y_1 + d_1 \geq y_1$, we conclude that (11) is satisfied and hence, $\leq_1$ defined by $\mathscr{D}$ is transitive.

Note that condition (11) can be written as

$$\mathscr{D}(y + d) + \mathscr{D}(y) \subseteq \mathscr{D}(y) \text{ for all } y \in Y \text{ and all } d \in \mathscr{D}(y),$$

as $\mathscr{D}(y)$ are convex cones for all $y \in Y$.

An ordering map $\mathscr{D}$ which defines a transitive binary relation $\leq_2$ is given in the next example:

Example 18 Let $Y = \mathbb{R}^2$ and consider the ordering map $\mathscr{D}\colon \mathbb{R}^2 \to 2^{\mathbb{R}^2}$ with

$$\mathscr{D}(y_1, y_2) := \begin{cases} \left\{ \begin{pmatrix} r\cos\varphi \\ r\sin\varphi \end{pmatrix} \in \mathbb{R}^2 \,\middle|\, r \geq 0,\ \ \varphi \in [0, \frac{\pi}{8}] \right\} & \text{if } y_1 \leq \frac{\pi}{8}, \\ \left\{ \begin{pmatrix} r\cos\varphi \\ r\sin\varphi \end{pmatrix} \in \mathbb{R}^2 \,\middle|\, r \geq 0,\ \ \varphi \in [0, y_1] \right\} & \text{if } y_1 \in (\frac{\pi}{8}, \frac{\pi}{2}), \\ \mathbb{R}^2_+ & \text{if } y_1 \geq \frac{\pi}{2}. \end{cases}$$

Again, $\mathscr{D}(y) \subseteq \mathbb{R}^2_+$ and $\mathscr{D}(y)$ is a closed pointed convex cone for all $y \in \mathbb{R}^2$. The cone $\mathscr{D}(y)$ depends on y_1 only, and for $z_1 \leq y_1$ for some $y, z \in \mathbb{R}^2$, it holds $\mathscr{D}(z) \subseteq \mathscr{D}(y)$. As for any $y \in \mathbb{R}^2$ and any $d \in \mathscr{D}(y)$ we have $d_1 \geq 0$ and thus $y_1 - d_1 \leq y_1$, we conclude that (12) is satisfied.

Also, as $y_1 \leq y_1 + d_1$, we obtain that

$$\mathscr{D}(y) \subseteq \mathscr{D}(y + d) \ \text{ for all } y \in Y \text{ and for all } d \in \mathscr{D}(y). \tag{16}$$

Property (16) is similar to the f-inclusive condition defined by Huang et al. in [43]. Example 18 is a modification of an example presented there.

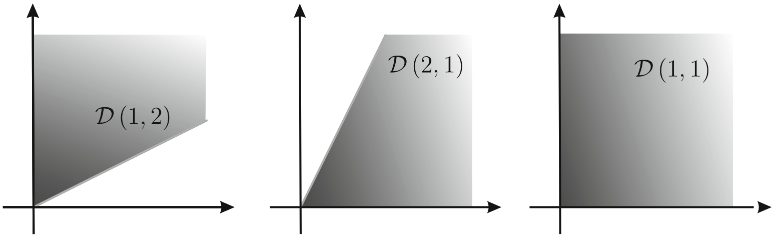


Fig. 3 Some of the cones $\mathscr{D}(y)$ of Example 19, cf. [18]

Next we recall an example from [20] for an ordering map $\mathscr{D}$ which satisfies condition (13). Hence the corresponding relations $\leq_1$ and $\leq_2$ are compatible with nonnegative scalar multiplication for such a map.

Example 19 Let $Y = \mathbb{R}^2$ and consider the ordering map $\mathscr{D}\colon \mathbb{R}^2 \to 2^{\mathbb{R}^2}$ with

$$\mathscr{D}(y) = \begin{cases} \left\{ \begin{pmatrix} r\cos\varphi \\ r\sin\varphi \end{pmatrix} \in \mathbb{R}^2 \;\middle|\; r \geq 0,\ \varphi \in \left[\bar{\varphi}_y - \frac{\pi}{4}, \bar{\varphi}_y + \frac{\pi}{4}\right] \cap \left[0, \frac{\pi}{2}\right] \right\} & \text{if } y \neq 0_{\mathbb{R}^2}, \\ \mathbb{R}^2_+ & \text{if } y = 0_{\mathbb{R}^2}. \end{cases}$$

where $\bar{\varphi}_y \in [0, \pi/2)$ is defined by

$$y = (r_y \cos(l\bar{\varphi}_y), r_y \sin(l\bar{\varphi}_y)) \text{ for some } l \in \mathbb{N} \text{ and some } r_y \in \mathbb{R},\ r_y > 0.$$

For an illustration of some of these cones see Fig. 3. Then $\mathscr{D}(y) = \mathscr{D}(y/\|y\|_2)$ for all $y \in \mathbb{R}^2 \setminus \{0_{\mathbb{R}^2}\}$ and thus $\mathscr{D}(y) = \mathscr{D}(\alpha y)$ for all $\alpha > 0$ and all $y \in Y$.

Practical requirements for variable ordering structures have already been studied by Engau in [28]. He discussed basic properties of such ordering structures denoted as local preferences, which we recall in the following.

Definition 16 Let S be an arbitrary nonempty set with a binary relation $\leq$. Let $y, d, d^1, d^2 \in S$ and $\lambda > 0$ be arbitrarily chosen. The binary relation $\leq$ is said to be

(i) *multiplicative*, if $y - d \leq y$ implies $y - \lambda\, d \leq y$.
(ii) *additive*, if $y - d^1 \leq y$ and $y - d^2 \leq y$ imply $y - (d^1 + d^2) \leq y$.
(iii) Let $Y = \mathbb{R}^m$ be the Euclidean space. Let the *ideal point* z of S exist, i.e. $z \in \mathbb{R}^m$ satisfies $z_i := \inf\{y_i \in \mathbb{R} \mid y \in S\}$, and denote $\bar{y} := y - z$. The binary relation $\leq$ is said to be *ideal-symmetric* on S, if for $(d^1)^\top \bar{y} = (d^2)^\top \bar{y}$ with $\|d^1\|_2 = \|d^2\|_2$ it holds:

$$y - d^1 \leq y \;\Leftrightarrow\; y - d^2 \leq y.$$

Note that the ideal point of a set $S \subseteq \mathbb{R}^m$ is a lower bound of this set w.r.t. $\mathbb{R}^m_+$, i.e. it holds $S \subseteq \{z\} + \mathbb{R}^m_+$.

Recall that for $\leq = \leq_2$, we can write

$$\mathscr{D}(y) = \{d \in Y \mid y - d \leq_2 y\}.$$

By [28, Proposition 3] multiplicativity and additivity imply that in this case $\mathscr{D}(y)$ has to be a convex cone. Together with the assumption of monotonicity (1) one obtains also that $\mathbb{R}^m_+ \subseteq \mathscr{D}(y)$.

For $\leq = \leq_1$, we have

$$\mathscr{D}(y) = \{d \in Y \mid y \leq_1 y + d\}.$$

A similar conclusion requires the following modification of Definition 16: in (i) it hast to be that $y \leq y + d$ implies $y \leq y + \lambda\, d$ and in (ii) it has to be that $y \leq y + d^1$ and $y \leq y + d^2$ imply $y \leq y + (d^1 + d^2)$.

The assumption of ideal-symmetry motivates to use ordering maps $\mathscr{D}$ on $\mathbb{R}^m$ where the images $\mathscr{D}(y) \subseteq \mathbb{R}^m$ are symmetric w.r.t. some $s \in \mathbb{R}^m \setminus \{0_{\mathbb{R}^m}\}$, i.e. $d \in \mathscr{D}(y)$ implies $d' \in \mathscr{D}(y)$ for all $d' \in \mathbb{R}^m$ with $d^\top s = (d')^\top s$, $\|d\|_2 = \|d'\|_2$.

At the end of this section, we give other basic assumptions on the ordering map which are often used in literature.

Bao et al. [3] do not assume that the images of the ordering map $C : Y \to 2^Y$ are cones. However, $C(y)$ is closed and $0_Y \in \operatorname{bd}(C(y))$ holds for all $y \in Y$. Moreover, there is a vector $d \in Y$ which satisfies $C(y) + \lambda d \subseteq C(y)$ for all $y \in Y$ and $\lambda \geq 0$. The last condition means that there is a common (unbounded) direction which is contained in all ordering sets in the considered space. In [13], also an ordering map $\mathscr{C} \colon X \to 2^Y$ was considered, with X some linear space. This was done in the context of a set optimization problem

$$\min_{x \in \Omega} F(x)$$

with feasible set $\Omega \subseteq X$ and set-valued objective map $F \colon X \to 2^Y$. Under the assumption that $\mathscr{C}(x)$ is a closed, convex, nontrivial and pointed cone for each $x \in X$, optimality conditions are developed there. There are some theoretical advantages if the ordering map acts between the same spaces as the objective map. However, situations like $\mathscr{C}(x^1) \neq \mathscr{C}(x^2)$ but $F(x^1) = F(x^2)$ might occur for some $x^1, x^2 \in X$.

4.3 Ordering Maps with BP Cones

We focus in the following on variable ordering structures which are defined by an ordering map $\mathscr{D}$ where the images $\mathscr{D}(y)$ are (representable as) Bishop-Phelps cones, see Sect. 3.2. For that reason, we assume within this subsection that $(Y, \|\cdot\|)$ is a real normed space. Moreover, we assume that to each $y \in Y$ we can associate a norm

$\|\cdot\|_y$ which is equivalent to, but eventually different from, the norm of the space Y, as well as an element $l_y \in Y^*$ such that we can write

$$\mathscr{D}(y) = C(l_y) = \{u \in Y \mid \|u\|_y \leq l_y(u)\} \text{ for all } y \in Y.$$

We will make use of the map $\ell\colon Y \to Y^*$ which is defined by $\ell(y) := l_y$ and thus we can also write

$$\mathscr{D}(y) = C(\ell(y)) = \{u \in Y \mid \|u\|_y \leq \ell(y)(u)\} \text{ for all } y \in Y.$$

We start with an example of such an ordering map:

Example 20 Let Y be the Euclidean space $\mathbb{R}^2$, $\|\cdot\|_y := \|\cdot\|_2$ for all $y \in \mathbb{R}^2$, and define $\ell\colon \mathbb{R}^2 \to \mathbb{R}^2$ by

$$\ell(y_1, y_2) := \left(\frac{3 + \sin y_1}{2}, \frac{3 + \cos y_2}{2} \right)^\top \in [1, 2] \times [1, 2]. \tag{17}$$

Then $\mathbb{R}^2_+ \subseteq C(\ell(y))$ for all $y \in \mathbb{R}^2$. The cones $C(\ell(y))$ can be visualized as follows: The two extreme rays of the pointed convex cone $C(\ell(y))$ are given by two rays starting in the origin being defined by the intersection points of the unit circle and the line connecting the points

$$\left(\frac{1}{\ell_1(y)}, 0\right) \text{ and } \left(0, \frac{1}{\ell_2(y)}\right),$$

see Fig. 4. For instance, $C(\ell(3\pi/2, \pi)) = \mathbb{R}^2_+$.

Recall that it is not a strong restriction to assume that the images of the ordering map are representable as BP cones. In most literature, the convex cones appearing in vector optimization problems are closed and pointed. According to Theorem 2, in finite dimensions such cones are all representable as BP cones. However, already in $\mathbb{R}^m$ one might need different equivalent norms to represent different nontrivial closed pointed convex cones as BP cones. In $\mathbb{R}^2$ it is enough to choose just one norm

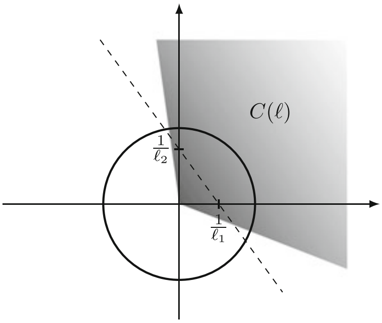


Fig. 4 BP cone $C(\ell(y))$ of Example 20 for $\ell_1 = \ell_1(y)$ and $\ell_2 = \ell_2(y)$, as well as the unit ball w.r.t. the Euclidean norm and (in dashed line) the line connecting the points $(1/\ell_1, 0)$ and $(0, 1/\ell_2)$, cf. [22]

but already in $\mathbb{R}^3$ one has to use different norms to model for instance a polyhedral cone and the Lorentz cone. In applications however, there might be an ordering map with different cones $\mathscr{D}(y)$ but presumably they will all be of the same type, for instance all polyhedral, and can all be modelled with the same norm.

Examples for such ordering maps are provided by Engau in [27, 28]. The properties which are assumed there for an ordering map imply that the images $\mathscr{D}(y)$ are BP cones w.r.t. the Euclidean norm.

Example 21 Let the cone-valued map $\mathscr{D}\colon A \to 2^{\mathbb{R}^m}$ be defined on some bounded set $A \subseteq \mathbb{R}^m$ by

$$\mathscr{D}(y) := \left\{ d \in \mathbb{R}^m \mid d^\top (y-p) \geq \gamma \cdot \|d\|_2 \cdot [y-p]_{\min} \right\} \quad \text{for all} \quad y \in A$$

where $\gamma \in (0,1]$, $p_i := \inf_{y \in A} y_i$ for $i = 1, \dots, m$, and

$$[y-p]_{\min} := \min_{i=1,\dots,m} y_i - p_i,$$

compare [28, 42].

Here, $\gamma \in (0,1]$ is a scalar which controls the angle of the elements of the cone with the vector $y - p$. These cones $\mathscr{D}(y)$ are Bishop-Phelps cones for each $y \in A$. This can be seen by setting

$$\ell(y) := \frac{1}{\gamma\ [y-p]_{\min}}(y-p)$$

and thus

$$\mathscr{D}(y) = C(\ell(y)) := \{u \in Y \mid \|u\|_2 \leq \ell(y)^\top u\}. \tag{18}$$

Moreover, it holds $\mathbb{R}^m_+ \subseteq \mathscr{D}(y)$ for all $y \in A$, because for any $d \in \mathbb{R}^m_+ \setminus \{0_{\mathbb{R}^m}\}$ one gets

$$\frac{d^\top (y-p)}{\gamma \|d\|_2 [y-p]_{\min}} \geq \frac{d^\top (y-p)}{\gamma \|d\|_1 [y-p]_{\min}} \geq \frac{\|d\|_1 [y-p]_{\min}}{\gamma \|d\|_1 [y-p]_{\min}} = \frac{1}{\gamma} \geq 1,$$

i.e. $d \in \mathscr{D}(y)$.

In particular, when the norm $\|\cdot\|_y$ in the definition of the BP cones $\mathscr{D}(y)$ is assumed to equal the norm $\|\cdot\|$ of the space Y and is thus equal for all $y \in Y$ as in the previous example, these cones reduce to the BP cones

$$\mathscr{D}(y) = C(\ell(y)) = \{u \in Y \mid \|u\| \leq \ell(y)(u)\}. \tag{19}$$

Example 20 also provides such an ordering map. It shows that even if the norm $\|\cdot\|$ does not depend on y a wide range of different cones is covered by the images $\mathscr{D}(y)$ in (19).

As in Example 8, the representation of the cones $\mathscr{D}(y)$ as BP cones can be used for scalarization results. Extensive examinations can be found in [22]. We give here an example for a characterization of the nondominated elements of a set as defined in Remark 1: $\bar{y} \in S \subseteq Y$ is said to be a nondominated element of the set S w.r.t. the ordering map $\mathscr{D}$, if there is no $y \in S$ with $y \neq \bar{y}$ and $\bar{y} \in \{y\} + \mathscr{D}(y)$.

Example 22 Let $S \subseteq Y$ be a nonempty set and let Y be equipped with a variable ordering which is defined by the ordering map $\mathscr{D}$ where the cones $\mathscr{D}(y)$ are defined as in (19) with $\|\ell(y)\|_* \geq 1$.

One can define a functional $\xi_{\bar{y}}\colon Y \to \mathbb{R}$ by

$$\xi_{\bar{y}}(y) = \ell(y)(y - \bar{y}) + \|y - \bar{y}\|,$$

cf. [22]. Then it holds $\bar{y} \in \{y\} + \mathscr{D}(y)$ for some $y \in Y$ if and only if

$$\|\bar{y} - y\| \leq \ell(y)(\bar{y} - y),$$

i.e. if and only if $\xi_{\bar{y}}(y) \leq 0 = \xi_{\bar{y}}(\bar{y})$, as $\ell(y)$ and $\ell(\bar{y})$ are linear maps. Hence, we have that $\bar{y} \in S$ is a nondominated element of the set S w.r.t. $\mathscr{D}$ if and only if

$$0 < \xi_{\bar{y}}(y) \;\text{ for all } y \in S \setminus \{\bar{y}\}, \tag{20}$$

i.e. if and only if $\bar{y}$ is the unique minimal solution of

$$\min_{y \in S} \xi_{\bar{y}}(y).$$

Now assume that $Y = \mathbb{R}^2$ is the Euclidean space and that $\ell\colon \mathbb{R}^2 \to \mathbb{R}^2$ which defines $\mathscr{D}(y) = C(\ell(y))$ is given as in (17), i.e.

$$\ell(y_1, y_2) := \left(\frac{3 + \sin y_1}{2}\,,\; \frac{3 + \cos y_2}{2}\right)^\top \in [1,2] \times [1,2],$$

and with $\|\cdot\|_y := \|\cdot\|_2$ for all $y \in \mathbb{R}^2$. Further, let

$$S := \{(y_1, y_2) \in \mathbb{R}^2 \mid y_1 \geq 0,\; y_2 \geq 0,\; y_2 \geq \pi - y_1\}.$$

We show that $\bar{y} = (0, \pi)$ is a nondominated element of S w.r.t. the ordering map $\mathscr{D}$ by verifying that (20) holds.

Obviously, for all $y \in S \setminus \{\bar{y}\}$ with $y_2 \geq \pi$, we have

$$\xi_{\bar{y}}(y) = \frac{3 + \sin y_1}{2}(y_1 - 0) + \frac{3 + \cos y_2}{2}(y_2 - \pi) + \|y - (0, \pi)^\top\|_2 > 0.$$

For $y \in S \setminus \{\bar{y}\}$ with $y_2 < \pi$, we have $0 > y_2 - \pi \geq -y_1$ and thus

$$\begin{aligned}\xi_{\bar{y}}(y) &\geq \frac{3+\sin y_1}{2} y_1 + \frac{3+\cos y_2}{2}(-y_1) + \sqrt{y_1^2 + (y_2-\pi)^2} \\ &> \underbrace{\left(\frac{\sin y_1 - \cos y_2}{2}\right)}_{\in[-1,1]} y_1 + y_1 \geq 0.\end{aligned}$$

Another possible scalarization which can be used when a representation as BP cones is available was given in [21]. For some $\bar{y} \in Y$ the functional

$$y \mapsto \ell(y)(y-\bar{y}) \qquad \forall\, y \in Y$$

is also a nonlinear scalarization functional which gives for instance with

$$\ell(y)(y-\bar{y}) > 0 \qquad \forall\, y \in S \setminus \{\bar{y}\}$$

a sufficient condition for $\bar{y} \in S$ to be a nondominated element of S w.r.t. $\mathscr{D}$, cf. [21, Thm 4.2].

We end this section with an example of an ordering map with BP cones as values in the infinite dimensional Hilbert space $L_2([0,1])$, cf. [20]:

Example 23 Let $Y = L_2([0,1])$ denote the real linear space of all (equivalence classes of) quadratic Lebesgue-integrable functions $f\colon [0,1] \to \mathbb{R}$ with inner product

$$\langle f, g\rangle := \int\limits_0^1 f(x)g(x)\mathrm{d}x \qquad \forall\, f, g \in L_2([0,1]).$$

Then Y is a Hilbert space and we can set $Y = Y^*$. Let a map $\ell\colon Y \to Y^*$ be defined by

$$\ell(f) = f + e \qquad \forall\, f \in L_2([0,1])$$

with $e \in L_2([0,1])$, $e(x) := 1$ for all $x \in [0,1]$. Then a cone-valued map $\mathscr{D}\colon Y \to 2^Y$ is defined by

$$\begin{aligned}\mathscr{D}(f) &:= C(\ell(f)) \\ &= \left\{g \in L_2([0,1]) \mid \langle \ell(f), g\rangle \geq \sqrt{\langle g, g\rangle}\right\} \\ &= \left\{g \in L_2([0,1]) \,\middle|\, \int\limits_0^1 (f(x)g(x) + g(x))\mathrm{d}x \geq \sqrt{\int\limits_0^1 (g(x))^2 \mathrm{d}x}\right\}\end{aligned}$$

for all $f \in L_2([0,1])$.

5 Set Relations

In this section, we introduce binary relations defined on the power set of a partially ordered linear space Y, i.e. on the set of all nonempty subsets of the space Y. Such relations generalize the relations considered before and are defined using the ordering cone. At first, we introduce three relations which are widely used in literature nowadays, cf. [48, 58, 74]. In the following, let the linear space Y be partially ordered by the pointed convex cone $K \subseteq Y$.

Definition 17 Let A and B be nonempty subsets of Y. Then we define:

(i) the *l-less order relation* by

$$A \preccurlyeq_l B \ \Leftrightarrow \ B \subseteq A + K,$$

(ii) the *u-less order relation* by

$$A \preccurlyeq_u B \ \Leftrightarrow \ A \subseteq B - K,$$

(iii) the *set less order relation* by

$$A \preccurlyeq_s B \ \Leftrightarrow \ B \subseteq A + K \text{ and } A \subseteq B - K.$$

The above relations are pre-orders on the power set of Y—they are reflexive and transitive. However, none of these relations is antisymmetric in general. Hence, for arbitrary sets $A, B \subseteq Y$, from the relations $A \preccurlyeq_. B$ and $B \preccurlyeq_. A$, it does not follow $A = B$. However, these two inequalities combined together define an equivalence relation w.r.t. the considered pre-order $\preccurlyeq_.$, cf. [50],

$$A \sim_{\preccurlyeq_.} B \quad \Leftrightarrow \quad A \preccurlyeq_. B \text{ and } B \preccurlyeq_. A.$$

Therefore, for any pre-order from Definition 17, there is an equivalence class

$$[A]_{\preccurlyeq_.} := \{B \subseteq Y \mid A \sim_{\preccurlyeq_.} B\}.$$

All three relations from Definition 17 play an important role in set-valued optimization using the so called set approach which means that in contrast to the vector approach, where elements of the sets are considered, the whole sets are compared with each other. Based on the above relations, solution notions can be defined which generalize the notions used in vector optimization. In the literature, it is possible to find papers regarding the existence of solutions [40, 50] or optimality conditions [50, 59, 65]. For a comprehensive collection of results on set-valued optimization, we refer the reader to the book [56] and the survey paper [38]. An important application of set-valued optimization is for instance robust optimization [26, 45] and mathematical finance [38, 56].

In the next example (adapted from [24, Ex. 3.2]) we discuss a special situation where there is no need to consider the set less order relation since it can be replaced by the l-less order relation.

Example 24 Let K have a nonempty algebraic interior ($\mathrm{cor}(K) \neq \emptyset$). Then K is reproducing, i.e. $K - K = Y$. Let $\mathcal{M} \subseteq 2^Y$ be a family of nonempty sets with the property $A + K = A$ for all $A \in \mathcal{M}$. Then, for arbitrary sets $A, B \in \mathcal{M}$, it follows that

$$A \subseteq Y = B + Y = B + K - K = B - K,$$

i.e. we have $A \preccurlyeq_u B$. Therefore, for arbitrary $A, B \in \mathcal{M}$, $A \preccurlyeq_s B$ holds if $A \preccurlyeq_l B$ is satisfied, and it suffices to study the l-less order relation.

Analogous argumentation can also be used to motivate the usefulness of the u-less order relation. The u-less order relation plays also an important role in robust multiobjective optimization. For instance optimality in case of decision uncertainty can be defined by using set-valued optimization and the u-less order relation, see [26].

Scalarization is another very important topic in set-valued optimization, see [36, 41, 57]. Recently, some results on vectorization in set optimization combined with two set relations—the set less and the minmax less order relation—have been derived by Jahn [49].

There are a lot of other possibilities how to compare sets with each other.

Definition 18 Let A and B be nonempty subsets of the real linear space Y. Then we define:

(i) the certainly less order relation by

$$A \preccurlyeq_c B \;\Leftrightarrow\; \forall a \in A\; \forall b \in B : \; a \leq_K b.$$

(ii) the possibly less order relation by

$$A \preccurlyeq_p B \;\Leftrightarrow\; \exists a \in A\; \exists b \in B : \; a \leq_K b.$$

The above relations are no pre-orders since the relation $\preccurlyeq_c$ is not reflexive and for $\preccurlyeq_p$, the transitivity fails in general. In [50] the definition of the relation $\preccurlyeq_c$ is slightly modified in order to obtain reflexivity. For an illustration of the relations from Definitions 17 and 18 where $Y = \mathbb{R}^2$ and $K = \mathbb{R}^2_+$, see Fig. 5.

Between the relations from Definitions 17 and 18 for two nonempty sets A and B, we have the following implications:

$$\begin{tikzcd}
 & A \preccurlyeq_l B \arrow[dr, Rightarrow] & \\
A \preccurlyeq_c B \arrow[ur, Rightarrow] \arrow[dr, Rightarrow] & & A \preccurlyeq_p B. \\
 & A \preccurlyeq_u B \arrow[ur, Rightarrow] &
\end{tikzcd}$$

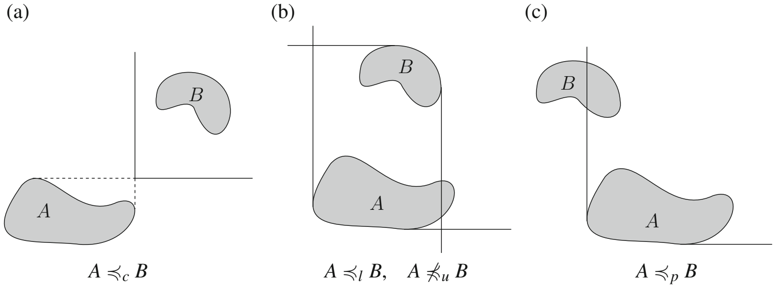
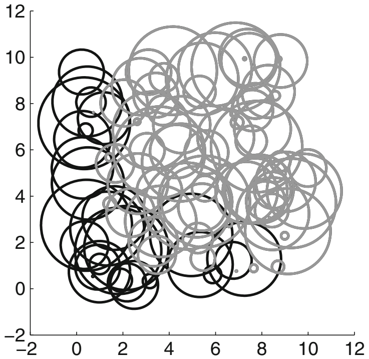


Fig. 5 Illustration of: (**a**) the certainly less order relation, (**b**) the l-less and u-less order relation, and (**c**) the possibly less order relation

Fig. 6 Sets of Example 25. The gray circles are the boundaries of those disks which are larger than some other disk w.r.t. the certainly less order relation. Compare [24, Fig. 3]



It is well-known that the certainly less order relation is a very strong concept, and that in practical applications it might happen that no sets can be compared using this set relation. Therefore, defining a solution w.r.t. this relation seems not to be reasonable at first glance. However, if there is a large number of sets, for instance in a low-dimensional space such as $\mathbb{R}^2$, then there might be many sets which can be compared. In this case, the certainly less order relation (or characterization results for this relation) can be used to pre-select some sets. This is illustrated in the following example which is taken from [24].

Example 25 Let the space $\mathbb{R}^2$ be partially ordered by the natural ordering, i.e. $Y = \mathbb{R}^2$ and $K = \mathbb{R}^2_+$. Let $\mathscr{M}$ be a family of n nonempty closed disks with random radius in the open interval $]0, 2[$ and random center points $(x_i, y_i) \in]0, 10[\times]0, 10[$ with $i = 1, \dots, n$. (The implementation uses the Matlab function rand, which constructs pseudorandom values drawn from the standard uniform distribution on the open interval $]0, 1[$.) Figure 6 shows such a family of sets $\mathscr{M}$ with $n = 100$ disks. We determined those disks, which are larger w.r.t. the certainly less order relation than any other set and marked them in the figure in gray—these have been $k = 72$ disks. So only the remaining 28 disks have to be considered for other set relations as the l-less or the set less order relation, in case one wants to find the "minimal" sets. In a simulation, for families $\mathscr{M}$ of random disks with $n = 5000$ disks, repeated 100 times, the average result is that 79% of the disks are larger than any other of the n sets w.r.t. the certainly less order relation, and thus, these sets can be deleted in a pre-selection.

There are even more concepts for set relations which are based, for instance, on comparing the sets of minimal or maximal elements of the considered sets as defined in [50, Sec. 3.2]. Moreover, all relations from Definitions 17 and 18 are generalized to the case where the considered linear space is equipped with a variable ordering structure in [24, 25].

References

1. R.G.D. Allen, The foundations of a mathematical theory of exchange. Economica **12**, 197–226 (1932)
2. D. Baatar, M.M. Wiecek, Advancing equitability in multiobjective programming. Comput. Math. Appl. **52**, 225–234 (2006)
3. T.Q. Bao, G. Eichfelder, B. Soleimani, C. Tammer, Ekeland's variational principle for vector optimization with variable ordering structure. J. Convex Anal. **24**(2), 393–415 (2017)
4. G.P. Barker, Theory of cones. Linear Algebra Appl. **39**, 263–291 (1981)
5. E.M. Bednarczuk, Bishop-Phelps cones and convexity: applications to stability of vector optimization problems. INRIA Rapport de Recherche, vol. 2806 (1996)
6. E. Bishop, R.R. Phelps, The support functionals of a convex set. Proc. Symp. Pure Math. **7**, 27–35 (1962)
7. I. Bomze, G. Eichfelder, Copositivity detection by difference-of-convex decomposition and ω-subdivision. Math. Program. Ser. A **138**, 365–400 (2013)
8. C. Brás, G. Eichfelder, J. Júdice, Copositivity tests based on the linear complementarity problem. Comput. Optim. Appl. **63**(2), 461–493 (2016)
9. S. Bundfuss, M. Dür, Algorithmic copositivity detection by simplicial partition. Linear Algebra Appl. **428**, 1511–1523 (2008)
10. M. Chicco, F. Mignanego, L. Pusillo, S. Tijs, Vector optimization problems via improvement sets. J. Optim. Theory Appl. **150**, 516–529 (2011)
11. J.P. Dauer, R.J. Gallagher, Positive proper efficient points and related cone results in vector optimization theory. SIAM J. Control Optim. **28**, 158–172 (1990)
12. M. Dür, Copositive programming - a survey, Chapter in *Recent Advances in Optimization and its Applications in Engineering*, ed. by M. Diehl et al. (Springer, Heidelberg, 2010), pp. 3–20
13. M. Durea, R. Strugariu, C. Tammer, On set-valued optimization problems with variable ordering structure. J. Global Optim. **61**(4), 745–767 (2015)
14. M. Ehrgott, *Multicriteria Optimization* (Springer, Heidelberg, 2005)
15. G. Eichfelder, *Adaptive Scalarization Methods in Multiobjective Optimization* (Springer, Heidelberg, 2008)
16. G. Eichfelder, Optimal elements in vector optimization with a variable ordering structure. J. Optim. Theory Appl. **151**(2), 217–240 (2011)
17. G. Eichfelder, Variable ordering structures in vector optimization, in *Recent Developments in Vector Optimization*, Chap. 4, ed. by Q.H. Ansari, J.-C. Yao (Springer, Heidelberg, 2012), pp. 95–126
18. G. Eichfelder, Cone-valued maps in optimization. Appl. Anal. **91**(10), 1831–1846 (2012)
19. G. Eichfelder, Numerical procedures in multiobjective optimization with variable ordering structures. J. Optim. Theory Appl. **162**(2), 489–514 (2014)

20. G. Eichfelder, *Variable Ordering Structures in Vector Optimization* (Springer, Heidelberg, 2014)
21. G. Eichfelder, T. Gerlach, Characterization of properly optimal elements with variable ordering structures. Optimization **65**(3), 571–588 (2016)
22. G. Eichfelder, T.X.D. Ha, Optimality conditions for vector optimization problems with variable ordering structures. Optimization **62**(5), 597–627 (2013)
23. G. Eichfelder, J. Jahn, Set-semidefinite optimization. J. Convex Anal. **15**(4), 767–801 (2008)
24. G. Eichfelder, M. Pilecka, Set approach for set optimization with variable ordering structures Part I: set relations and relationship to vector approach. J. Optim. Theory Appl. **171**, 931–946 (2016)
25. G. Eichfelder, M. Pilecka, Set approach for set optimization with variable ordering structures Part II: scalarization approaches. J. Optim. Theory Appl. **171**, 947–963 (2016)
26. G. Eichfelder, C. Krüger, A. Schöbel, Decision uncertainty in multiobjective optimization. J. Global Optim. **69**(2), 485–510 (2017)
27. A. Engau, Domination and decomposition in multiobjective programming. Dissertation, University of Clemson (2007)
28. A. Engau, Variable preference modeling with ideal-symmetric convex cones. J. Global Optim. **42**, 295–311 (2008)
29. N. Georgescu, The pure theory of consumer's behaviour. Q. J. Econ. **50**, 545–593 (1936)
30. N. Georgescu, Choice and revealed preference. South. Econ. J. **21**, 119–130 (1954)
31. M. Gerstenhaber, Theory of convex polyhedral cones, in *Activities Analysis of Production and Allocation*, Chap. 18, ed. by T.C. Koopmans (Wiley, New York, 1951), pp. 298–316
32. A. Göpfert, R. Nehse, *Vektoroptimierung: Theorie, Verfahren und Anwendungen* (Teubner, Leipzig, 1990)
33. A. Göpfert, H. Riahi, C. Tammer, C. Zălinescu, *Variational Methods in Partially Ordered Spaces* (Springer, New York, 2003)
34. M. Grabisch, *Set Functions, Games and Capacities in Decision Making* (Springer, Cham, 2016)
35. C. Gutiérrez, B. Jiménez, V. Novo, Improvement sets and vector optimization. Eur. J. Oper. Res. **223**(2), 304–311 (2012)
36. C. Gutiérrez, B. Jiménez, E. Miglierina, E. Molho, Scalarization in set optimization with solid and nonsolid ordering cones. J. Global Optim. **61**(3), 525–552 (2015)
37. T.X.D. Ha, J. Jahn, Properties of Bishop-Phelps cones. J. Nonlinear and Convex Anal. **18**(3), 415–429 (2017)
38. A.H. Hamel, F. Heyde, A. Löhne, B. Rudloff, C. Schrage, Set optimization – a rather short introduction, in *Set Optimization and Applications - The State of the Art*, Chap. 3, ed. by A.H. Hamel et al. (Springer, Berlin, Heidelberg, 2015), pp. 65–141
39. S. Helbig, Approximation of the efficient point set by perturbation of the ordering cone. Z. Oper. Res. **35**(3), 197–220 (1991)
40. E. Hernández, L. Rodríguez-Marín, Existence theorems for set optimization problems. Nonlinear Anal. **67**(6), 1726–1736 (2007)
41. E. Hernández, L. Rodríguez-Marín, Nonconvex scalarization in set optimization with set-valued maps. J. Math. Anal. Appl. **325**(1), 1–18 (2007)
42. C. Hirsch, P.K. Shukla, H. Schmeck, Variable preference modeling using multi-objective evolutionary algorithms, in *Evolutionary Multi-Criterion Optimization - 6th International Conference*, ed. by R.H.C. Takahashi, K. Deb, E.F. Wanner, S. Greco (Springer, Heidelberg, 2011), pp. 91–105
43. N.J. Huang, X.Q. Yang, W.K. Chan, Vector complementarity problems with a variable ordering relation. Eur. J. Oper. Res. **176**, 15–26 (2007)
44. D.H. Hyers, G. Isac, T.M. Rassias, *Topics in Nonlinear Analysis & Applications* (World Scientific, Singapore, 1997)

45. J. Ide, E. Köbis, D. Kuroiwa, A. Schöbel, C. Tammer, The relationship between multi-objective robustness concepts and set-valued optimization. Fixed Point Theory Appl. **83** (2014). https://doi.org/10.1186/1687-1812-2014-83
46. G. Isac, A.O. Bahya, Full nuclear cones associated to a normal cone. Application to Pareto efficiency. Appl. Math. Lett. **15**, 633–639 (2002)
47. J. Jahn, Bishop-Phelps cones in optimization. Int. J. Optim. Theory Methods Appl. **1**, 123–139 (2009)
48. J. Jahn, *Vector Optimization - Theory, Applications, and Extensions*, 2nd edn. (Springer, Heidelberg, 2011)
49. J. Jahn, Vectorization in set optimization. J. Optim. Theory Appl. **167**(3), 783–795 (2015)
50. J. Jahn, T.X.D. Ha, New order relations in set optimization. J. Optim. Theory Appl. **148**, 209–236 (2011)
51. R. John, The concave nontransitive consumer. J. Global Optim. **20**, 297–308 (2001)
52. R. John, Local and global consumer preferences, in *Generalized Convexity and Related Topics*, ed. by I. Konnov, D.T. Luc, A. Rubinov (Springer, Berlin, 2006), pp. 315–326
53. V. Kaibel, Another proof of the fact that polyhedral cones are finitely generated (2009). arXiv:0912.2927v1
54. E.K. Karaskal, W. Michalowski, Incorporating wealth information into a multiple criteria decision making model. Eur. J. Oper. Res. **150**, 204–219 (2003)
55. R. Kasimbeyli, A nonlinear cone separation theorem and scalarization in nonconvex vector optimization. SIAM J. Optim. **20**, 1591–1619 (2010)
56. A.A. Khan, C. Tammer, C. Zălinescu, *Set-valued Optimization: An Introduction with Applications* (Springer, Berlin, 2015)
57. E. Köbis, M.A. Köbis, Treatment of set order relations by means of a nonlinear scalarization functional: a full characterization. Optimization **65**(10), 1805–1827 (2016)
58. D. Kuroiwa, On set-valued optimization. Nonlinear Anal. **47**(2), 1395–1400 (2001)
59. D. Kuroiwa, On derivatives of set-valued maps and optimality conditions for set optimization. J. Nonlinear Convex Anal. **10**(1), 41–50 (2009)
60. D.T. Luc, *Theory of Vector Optimization* (Springer, Berlin, 1989)
61. V.L. Makarov, M.J. Levin, A.M. Rubinov, *Mathematical Economic Theory: Pure and Mixed Types of Economic Mechanisms* (North-Holland, Amsterdam, 1995)
62. B.S. Mordukhovich, *Variational Analysis and Generalized differentiation, II: Applications* (Springer, Berlin, 2006)
63. B.S. Mordukhovich, Multiobjective optimization problems with equilibrium constraints. Math. Program. Ser. B **117**, 331–354 (2009)
64. M. Petschke, On a theorem of Arrow, Barankin, and Blackwell SIAM J. Control Optim. **28**, 395–401 (1990)
65. L. Rodríguez-Marín, M. Sama, (Λ, C)-contingent derivatives of set-valued maps. J. Math. Anal. Appl. **335**(2), 974–989 (2007)
66. B.S.W. Schröder, *Ordered Sets: An Introduction* (Birkhäuser, Boston, 2001)
67. A.M. Rubinov, R.N. Gasimov, Scalarization and nonlinear scalar duality for vector optimization with preferences that are not necessarily a pre-order relation. J. Global Optim. **29**, 455–477 (2004)
68. M. Wacker, Multikriterielle Optimierung bei der Registrierung medizinischer Daten. Diploma thesis, University of Erlangen-Nürnberg (2008)
69. M. Wacker, F. Deinzer, Automatic robust medical image registration using a new democratic vector optimization approach with multiple measures, in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2009*, ed. by G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble, C. Taylor (Springer, Berlin, 2009), pp. 590–597
70. P. Weidner, Vergleichende Darstellung von Optimalitätsbegriffen und Dualitätsansätzen in der Vektoroptimierung, Diploma Thesis, Martin-Luther-Universität Halle-Wittenberg (1983)
71. P. Weidner, Charakterisierung von Mengen effizienter Elemente in linearen Räumen auf der Grundlage allgemeiner Bezugsmengen. PhD Thesis, Martin-Luther-Universität Halle-Wittenberg (1985)

72. P. Weidner, Functions with uniform sublevel sets and scalarization in linear spaces (2016). arXiv:1608.04091
73. M.M. Wiecek, Advances in cone-based preference modeling for decision making with multiple criteria. Decis. Mak. Manuf. Serv. **1**, 153–173 (2007)
74. R.C. Young, The algebra of many-valued quantities. Math. Ann. **104**(1), 260–290 (1931)
75. P.L. Yu, Cone convexity, cone extreme points, and nondominated solutions in decision problems with multiobjectives. J. Optim. Theory Appl. **14**, 319–377 (1974)

An Overview on Singular Nonlinear Elliptic Boundary Value Problems

Francesca Faraci and George Smyrlis

1 Introduction

The study of singular nonlinear problems started with the pioneering work of Fulks and Maybee [6] as a mathematical model for describing the heat conduction in an electric medium.

More precisely, if $\Omega \subset \mathbb{R}^3$ is a bounded region of the space occupied by an electrical conductor and $u(x, t)$ denotes the temperature at the point $x \in \Omega$ and time t, then, u satisfies the equation

$$cu_t - k\Delta u = \frac{E^2(x,t)}{f(u)},$$

where $E(x, t)$ describes the local voltage drop, $f(u)$ is the electrical resistivity (which depends on the temperature) and c and k are the specific heat and the thermal conductivity of the conductor respectively. The function f is positive, increasing and tends to zero as the argument tends to zero. The model example is precisely $f(u) = u^\gamma$ for some positive real number γ. The authors proved existence and uniqueness results by using fixed point theory, and showed also that the solution $u(x, t)$ of the parabolic equation tends uniformly to the unique solution of the corresponding elliptic problem as $t \to +\infty$. The study of stationary solutions of such equations leads then to investigate semilinear elliptic equations of the type

$$-\Delta u = h(x)u^{-\gamma}$$

F. Faraci
Department of Mathematics and Informatics, University of Catania, Catania, Italy
e-mail: ffaraci@dmi.unict.it

G. Smyrlis (✉)
Department of Mathematics, National Technical University of Athens, Athens, Greece
e-mail: gsmyrlis@math.ntua.gr

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_10

on a bounded domain $\Omega \subset \mathbb{R}^N$ ($N > 2$), for suitable positive functions $h(x)$ and positive γ.

In general, if $g : \Omega \times]0, +\infty[\to \mathbb{R}$ is a real function, the semilinear elliptic boundary value problem

$$\begin{cases} -\Delta u = g(x, u), & \text{in } \Omega \\ u > 0, & \text{in } \Omega \\ u = 0, & \text{on } \partial\Omega \end{cases} \tag{1}$$

is said to be *singular* if g exhibits a singularity at zero, i.e.

$$\lim_{t \to 0^+} g(x, t) = \infty \quad \text{for almost all} \quad x \in \Omega.$$

We say that $u \in W_0^{1,2}(\Omega)$ is a *weak solution* of (1) if $u > 0$ almost everywhere in Ω and

$$g(x, u)\varphi \in L^1(\Omega), \quad \int_\Omega \nabla u \nabla \varphi \, dx = \int_\Omega g(x, u)\varphi \, dx$$

for all $\varphi \in W_0^{1,2}(\Omega)$. A *classical solution* of (1) is a function $u \in C^2(\Omega) \cap C(\overline{\Omega})$ such that $u > 0$ in Ω, and

$$-\Delta u(x) = g(x, u(x)), \qquad \text{for all } x \in \Omega.$$

We recall also that a *generalized* solution of (1) is a function $u \in W_{loc}^{2,q}(\Omega) \cap C(\overline{\Omega})$ (q being suitably selected in $]1, +\infty[$) such that $u > 0$ in Ω, and

$$-\Delta u(x) = g(x, u(x)), \qquad \text{for almost all } x \in \Omega.$$

If $f : \Omega \times [0, +\infty[\to \mathbb{R}$ is a Carathéodory function and $g(x, t) = t^{-\gamma} + f(x, t)$ for every $x \in \Omega$ and $t > 0$, then problem (1) reads as

$$\begin{cases} -\Delta u = u^{-\gamma} + f(x, u), & \text{in } \Omega \\ u > 0, & \text{in } \Omega \\ u = 0, & \text{on } \partial\Omega. \end{cases} \tag{2}$$

If $0 < \gamma < 1$ then, it is possible to associate to problem (2) an energy functional on the Sobolev space $W_0^{1,2}(\Omega)$ given by

$$\mathscr{E}(u) = \frac{1}{2}\int_\Omega |\nabla u(x)|^2 dx - \frac{1}{1-\gamma}\int_\Omega u(x)^{1-\gamma} dx - \int_\Omega \int_0^{u(x)} f(x, t) dt \, dx.$$

Although not continuously Gâteaux differentiable, $\mathscr{E}$ is continuous, and variational methods (in the broad sense) are still applicable in order to produce weak solutions.

When $\gamma \geq 1$, such kind of problems have been less investigated. Notice in fact that the above functional is not defined on the whole space $W_0^{1,2}(\Omega)$. However, the existence of one or two solutions (classical or weak) can be still obtained by using upper-lower arguments, suitable truncation methods or techniques from non-smooth analysis.

Singular elliptic problems received a considerable attention after the seminal paper of Crandall et al. [2] where the existence of a classical solution for singular problems driven by general elliptic operators of second order was established.

For sake of simplicity, we will state the main theorems of [2] for boundary value problems of the form (1), where Ω is a bounded domain in $\mathbb{R}^N$ $(N > 2)$ with boundary $\partial\Omega$ of class C^3 and $g \in C(\overline{\Omega}\times]0,+\infty[)$ satisfies the condition

$$\lim_{t\to 0^+} g(x,t) = \infty \quad \text{uniformly for} \quad x \in \overline{\Omega}. \tag{3}$$

One of the main results in [2] is the following

Theorem 1.1 *([2, Theorem 1.1]) In addition to (3), assume that*

$$g \in C^1(\overline{\Omega}\times]0,+\infty[)$$

and

$$g(x,t) \ \textit{is non-increasing in} \ t \in]0,+\infty[, \quad \textit{for} \ x \in \overline{\Omega}.$$

Then (1) possesses a unique classical solution $u \in C^2(\Omega) \cap C(\overline{\Omega})$.

Thus, in the particular case when $g(x,t) = t^{-\gamma}$, the above result provides existence and uniqueness of a classical solution for (1) for every $\gamma > 0$.

In the proof of Theorem 1.1, the authors used the upper-lower solution method to solve, for every $\varepsilon > 0$ the approximate problems

$$\begin{cases} -\Delta u_\varepsilon = g(x, \varepsilon + u_\varepsilon), & \text{in } \Omega \\ u_\varepsilon = 0, & \text{on } \partial\Omega, \end{cases}$$

and then they showed the uniform convergence of $\{u_\varepsilon\}$ as $\varepsilon \to 0^+$ to a solution u of (1). The proof of the convergence exploits Sobolev embeddings theorems which ensure both that $\{u_\varepsilon\}$ is compact in $C^{1,\alpha}_{\mathrm{loc}}(\Omega)$ for some $\alpha \in (0,1)$ and $g(x,u_\varepsilon(x)) \to g(x,u(x))$ uniformly in C^1. It is worth noticing that Theorem 1.1 (existence part) still holds, if monotonicity of g is replaced by the assumption that g is bounded from above.

For the case where g is merely continuous, the authors obtained *generalized* positive solutions of (1). Namely, we have the following

Theorem 1.2 *([2, Theorem 1.21]) In addition to (3), assume that*

$$g \in C(\overline{\Omega}\times]0, +\infty[).$$

Then (1) *possesses a generalized solution* $u \in W^{2,q}_{loc}(\Omega) \cap C(\overline{\Omega})$ *for some* $q > N$.

When $g(x,t)$ does not depend on x, stronger global regularity properties than continuity of solutions can be obtained (see [2]).

By using the upper-lower solution method combined with fixed point theory, Coclite and Palmieri [1] established a bifurcation-type result for the parametric problem

$$\begin{cases} -\Delta u = u^{-\gamma} + \lambda u^{r-1}, & \text{in } \Omega \\ u > 0, & \text{in } \Omega \\ u = 0, & \text{on } \partial\Omega \end{cases} \tag{4}$$

where $2 < r < 2^*$ (being 2^* the critical Sobolev exponent), $\gamma > 0$ and $\lambda > 0$ is a real parameter.

Namely, they proved the following

Theorem 1.3 *([1, Corollary 4]) There exists a positive real number* λ^* *such that the problem (4) has at least one classical solution belonging to* $C^2(\Omega) \cap C(\overline{\Omega})$ *for* $0 \le \lambda < \lambda^*$ *and no solutions for* $\lambda > \lambda^*$.

When $f(x,t) = \beta(x)t^{-\gamma}$, i.e. the singularity is multiplied by a suitable positive weight β, the result of [2] has been extended in the work of Lazer and McKenna [14] who considered the problem

$$\begin{cases} -\Delta u = \beta(x)u^{-\gamma}, & \text{in } \Omega \\ u > 0, & \text{in } \Omega \\ u = 0, & \text{on } \partial\Omega. \end{cases} \tag{5}$$

Assuming that the function β is positive and sufficiently smooth on Ω and $\gamma > 0$ the authors ensured, again with the aid of upper-lower solution techniques, the existence and uniqueness of a classical solution to (5). Moreover in [14], it is also proved that the unique positive solution of (5) lies in $W^{1,2}_0(\Omega)$ (thus, it is also a weak solution) iff $\gamma < 3$.

A different approach is given in [10] where the problem is reduced to the equivalent integral equation

$$u(x) = \int_\Omega G(x,y)\beta(y)u(y)^{-\gamma}dy.$$

Precise estimates of the Green function G and of its gradient near $\partial\Omega$ allow to apply Schauder's fixed point theorem in an appropriate setting.

Extending the above work, Lair and Shaker [13] considered the equation

$$-\Delta u=\beta(x)g(u),\quad \text{in } \mathbb{R}^N \quad (N>2) \tag{6}$$

and the boundary value problem

$$\begin{cases}-\Delta u=\beta(x)g(u), & \text{in } \Omega\\ u>0 & \text{in } \Omega\\ u=0, & \text{on } \partial\Omega,\end{cases} \tag{7}$$

where g is a positive non-increasing continuously differentiable function on $]0,+\infty[$ and $\Omega \subseteq \mathbb{R}^N$ $(N>2)$ is a bounded domain with sufficiently smooth boundary $\partial\Omega$.

It is shown that (6) has a unique positive classical solution in $\mathbb{R}^N$ that decays to zero, if β is a nontrivial nonnegative continuous function on $\mathbb{R}^N$ satisfying

$$\int_0^\infty t\left(\max_{|x|=t}\beta(x)\right)dt < \infty.$$

In the proof, the authors use the upper-lower solution method combined with extended maximum principle and comparison principle results for singular problems to obtain an increasing sequence v_k , $k=1,2,\ldots$ of positive classical solutions of the approximate equations

$$-\Delta v(x)=\beta(x)g\left(v(x)+\frac{1}{k}\right),\quad k=1,2,\ldots$$

with $v_k(x)\to 0$, as $|x|\to\infty$.

Then, by using the bootstrap argument, they prove that the pointwise limit of the sequence $\{v_k\}$ is the desired solution of (6).

Regarding problem (7), it is shown that there exists a unique positive weak solution $u\in W_0^{1,2}(\Omega)$, provided that

$$\beta\in L^2(\Omega)_+\setminus\{0\},\qquad \int_0^\varepsilon g(t)dt < \infty,\quad \text{for some } \varepsilon>0.$$

The proof is made by minimizing the functional $u\to \frac{1}{2}||\nabla u||_2^2$ on a certain weakly closed subset of $W_0^{1,2}(\Omega)$ and by using the Lagrange multiplier rule.

For the perturbed problem

$$\begin{cases}-\Delta u=\beta(x)u^{-\gamma}+\lambda u^{r-1}, & \text{in } \Omega\\ u>0, & \text{in } \Omega\\ u=0, & \text{on } \partial\Omega,\end{cases} \tag{8}$$

where $0 < \gamma < 1$, $\beta : \Omega \to \mathbb{R}$ is positive and smooth enough and $\lambda > 0$ is real parameter, the existence and uniqueness of a classical solution was proved by Shi and Yao in [19] when $1 < r < 2$. In the presence of a superlinear nonlinearity, a multiplicity theorem was proved by Sun et al. [20]. Inspired by the work of Lair and Shaker [13], Sun, Wu and Long studied problem (8) assuming that $2 < r < 2^*$, $\beta \in L^2(\Omega)_+ \setminus \{0\}$ satisfies

$$||\beta||_2 \leq q \left(\frac{S}{|\Omega|^a} \right)^{\frac{r-1+\gamma}{r-2}}, \tag{9}$$

where $|\Omega|$ is the Lebesgue measure of Ω and

$$q = \frac{1}{|\Omega|^{\frac{r-2+2\gamma}{2r}}} \cdot \frac{r-2}{r-1+\gamma} \cdot \left(\frac{1+\gamma}{r-1+\gamma} \right)^{\frac{1+\gamma}{r-2}},$$

$$a = 2 \cdot \frac{2^* - r}{2^* \cdot r}, \qquad S = \inf \left\{ \frac{||\nabla u||_2^2}{||u||_{2^*}^2} : u \in W_0^{1,2}(\Omega) \setminus \{0\} \right\}.$$

By using Ekeland' s variational principle, the authors established in [20] the following multiple existence result:

Theorem 1.4 *([20, Theorem 1]) If (9) holds, then, there exists a positive real number λ^* such that for $\lambda \in]0, \lambda^*[$, the problem (8) possesses at least two positive weak solutions $u_1,\ u_2 \in W_0^{1,2}(\Omega)$.*

We also mention the work of Zhang [21], where, with the aid of critical point theory on certain convex closed sets of $C_0^1(\overline{\Omega})$ the existence of at least two positive weak solutions for singular problems of the form

$$\begin{cases} -\Delta u = \beta(x) u^{-\gamma} + \lambda f(u), & \text{in } \Omega, \\ u > 0, & \text{in } \Omega \\ u = 0, & \text{on } \partial\Omega, \end{cases}$$

was established under certain hypotheses on the nonsingular term f and when $0 < \gamma < 1$. Here $\beta \in L^2(\Omega)_+ \setminus \{0\}$ satisfying

$$\beta \cdot \varphi^{-\gamma} \in L^s(\Omega), \quad \frac{N}{2} < s < \frac{2N}{N-2}, \quad N > 2,$$

where φ is a positive smooth principal eigenfunction of the Dirichlet operator $(-\Delta, W_0^{1,2}(\Omega))$.

Regarding semi-linear singular problems, we refer also to the work of Hirano et al. [11] who considered problem (4) under the assumption that γ is an arbitrary positive number. The main result produced in [11] is the following bifurcation-type result:

Theorem 1.5 *([11, Theorem 1]) There exists a positive real number* λ^* *such that for* $\lambda \in]0, \lambda^*[$, *the problem (4) possesses at least two positive weak solutions* $u_1, u_2 \in C^\infty(\Omega) \cap L^\infty(\Omega)$*; for* $\lambda = \lambda^*$, *it possesses at least one positive weak solution* $u \in C^\infty(\Omega) \cap L^\infty(\Omega)$*; for* $\lambda > \lambda^*$, *it has no positive weak solutions.*

The proof of Theorem 1.5 is based on nonsmooth analysis, seeking solutions of (4) as critical points of the corresponding energy functional in some suitable nonsmooth sense. The main tool is a linking theorem for functionals which are C^1-perturbations of convex functions.

Recently, several authors have paid attention to singular equations driven by p-Laplacian $(p > 1)$. In this connection, we mention the papers of Perera and Silva [15] and Giacomoni et al. [9].

Perera and Silva [15] studied the parametric singular problem

$$\begin{cases} -\Delta_p u = \beta(x)u^{-\gamma} + \lambda f(x,u), & \text{in } \Omega \\ u > 0, & \text{in } \Omega \\ u = 0, & \text{on } \partial\Omega \end{cases} \tag{10}$$

where $\gamma > 0$, $f(x,s)$ is $(p-1)$-superlinear with respect to s near ∞ and $\beta(\cdot)$ is a nontrivial nonnegative measurable function. In [15], f is allowed to change sign (this fact extends previous similar results) and it is bounded from below by integrable functions on bounded intervals of the variable s. Moreover, γ does not necessarily lie in $(0, 1)$ and this is a source of certain difficulties. For this reason, in [15] it is also assumed that

(H) there exist $q > N$ and $\varphi_0 \in C_0^1(\overline{\Omega})$ s.t. $\varphi_0 > 0$ on Ω, $\beta\varphi_0^{-\gamma} \in L^q(\Omega)$.

Under certain hypotheses on f it is shown that (10) has two positive weak solutions for small $\lambda > 0$. The approach is variational and it is based on the upper-lower solution method and on classical critical point theory, namely, Mountain Pass theorem, applied on certain truncations of the corresponding energy functional.

Giacomoni et al. [9] focused their attention on the singular boundary value problem $(0 < \gamma < 1)$

$$\begin{cases} -\Delta_p u = \lambda u^{-\gamma} + u^{r-1}, & \text{in } \Omega, \\ u > 0, & \text{in } \Omega \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{11}$$

where $p < r \le p^*$ and p^* is the critical Sobolev exponent.

In [9], the authors employed variational methods to prove a bifurcation-type result for the positive solutions of (11). In particular, they produced two positive smooth solutions of (11) for $\lambda > 0$ small. For this purpose, first they established two new results of separate interest: a strong comparison principle for singular problems and a regularity result for solutions to problem (11).

The literature is not so rich when searching for three or more solutions. As far as we know the only contribution in this direction is the paper by Zhao et al. [22], where the existence of three weak solutions is proved via an application of an abstract "three critical points" theorem.

In [22], the authors dealt with the problem

$$\begin{cases} -\Delta_p u = \lambda\beta(x)u^{-\gamma} + \lambda f(x,u), & \text{in } \Omega \\ u > 0, & \text{in } \Omega \\ u = 0, & \text{on } \partial\Omega \end{cases} \tag{12}$$

where $\Omega \subset \mathbb{R}^N$, $p > N$, Δ_p is the p-Laplacian operator, $\gamma > 0$, $f : \Omega \times [0,+\infty[\to \mathbb{R}$ is a Carathéodory function, β is a nonnegative function with a suitable summability (see assumption (H) above). The main theorem in [22] reads as follows:

Theorem 1.6 *([22, Theorem 2.6]) Let* $\gamma > 0$, $f : \Omega \times [0,+\infty[\to \mathbb{R}$ *be a Carathéodory function and* $a : \Omega \to \mathbb{R}$ *a non-negative function satisfying the following assumptions:*

(i) *there exists* $\bar{u} \in C_0^1(\overline{\Omega})_+$ *and* $q > N$ *such that* $\beta\bar{u}^{-\gamma} \in L^q(\Omega)$*;*
(ii) *there exist* $\delta, c_1 > 0$ *such that* $f(x,t) \geq c_1\beta(x)$ *for all* $t \in [0,\delta]$*, a.e.* $x \in \Omega$*;*
(iii) $\displaystyle\lim_{t\to+\infty} \frac{f(x,t)}{|t|^{p-1}} = 0$*;*
(iv) *there exists an open ball* $B_N = B_N(x_0, R_0) \subset \Omega$ *such that*

$$\int_{B_N}\int_{c(R_0^N\omega_N)^{1/p}}^{u_1} [\beta(x)t^{-\gamma} + f(x,t)]dtdx >$$

$$\int_{\Omega}\int_{u_2}^{c(R_0^N\omega_N)^{1/p}} [\beta(x)t^{-\gamma} + f(x,t)]dtdx,$$

for suitable constant $c > 0$ *and functions* u_1, u_2*, where* ω_N *is the volume of the ball* $B_N(0,1)$*.*

Then, there exist an open interval $\Sigma \subset]0,+\infty[$ *and a constant* $\rho > 0$ *such that for every* $\lambda \in \Sigma$*, the problem* (12) *has at least three distinct positive solutions in* $W_0^{1,p}(\Omega)$ *with their norms less than* ρ*.*

The proof of the above result relies on an application of an abstract three critical points theorem by Ricceri. It is crucial in this sense that $p > N$ so that the compactness of the embedding of $W_0^{1,p}(\Omega)$ in $C(\overline{\Omega})$ can be exploited. On the other hand, the parameter γ can be any positive number.

For an overview on singular elliptic problems we refer to [7] where existence and uniqueness properties, but also qualitative properties, including bifurcation, asymptotic analysis, blow-up of solutions are discussed.

The purpose of the present contribution is to give some new multiplicity results for singular problems of the type (2). In Sect. 3 we will present first some results from [5] and [3] on the existence of three weak solutions under a sublinear at infinity behaviour of the nonlinear term f. The existence of k solutions or even infinitely many solutions will be discussed in Sect. 4 on the basis of the recent contribution [4].

2 Preliminaries

In this section, for the convenience of the reader, we briefly recall some of the definitions and of the mathematical tools that we will use in this work.

Throughout the sequel Ω is a bounded domain in $\mathbb{R}^N$ $(N > 2)$ with smooth boundary $\partial\Omega$, $0 < \gamma < 1$, $f : [0, +\infty[\to \mathbb{R}$ is a continuous function, such that $f(0) = 0$. Without loss of generality we may assume that $f(t) = 0$ for every $t < 0$. We assume that f exhibits a subcritical behaviour, i.e. there exist constants $c > 0$ and $q < 2^*$ (being 2^* the Sobolev critical exponent) such that

$$f(t) \leq c(1 + t^{q-1}), \quad \text{for all } t \geq 0.$$

Denote by $F : \mathbb{R} \to \mathbb{R}$ the primitive of f, i.e. $F(t) = \int_0^t f(s)ds$.

Let us recall that, for $\lambda > 0$, a *weak solution* of

$$\begin{cases} -\Delta u = \lambda u^{-\gamma} + f(u), & \text{in } \Omega \\ u > 0, & \text{in } \Omega \\ u = 0, & \text{on } \partial\Omega \end{cases} \tag{$\mathscr{P}_\lambda$}$$

is a function $u \in W_0^{1,2}(\Omega)$ such that $u > 0$ almost everywhere in Ω and

$$u^{-\gamma}\varphi \in L^1(\Omega), \qquad \int_\Omega \nabla u \nabla \varphi \, dx = \int_\Omega (\lambda u^{-\gamma} + f(u))\varphi \, dx$$

for all $\varphi \in W_0^{1,2}(\Omega)$.

We can associate to problem $(\mathscr{P}_\lambda)$ the following energy functional

$$\mathscr{E}(u) = \frac{1}{2}\int_\Omega |\nabla u(x)|^2 dx - \frac{\lambda}{1-\gamma}\int_\Omega u(x)^{1-\gamma} dx - \int_\Omega \int_0^{u(x)} f(t)dt \, dx$$

which is well defined on the Sobolev space $W_0^{1,2}(\Omega)$. Although not continuously Gâteaux differentiable, $\mathscr{E}$ is continuous and under suitable assumptions on f it has a global minimum. It is worth noticing (see Proposition 3.1) that local minima of $\mathscr{E}$ are weak solutions of $(\mathscr{P}_\lambda)$ in the sense given above.

It is well known that in the ordered Banach space $C_0^1(\overline{\Omega})$ the positive cone

$$C_+ = \{u \in C_0^1(\overline{\Omega}) : u(x) \geq 0 \ \forall\, x \in \Omega\}$$

has a non empty interior given by

$$\mathrm{int}C_+ = \{u \in C_+ : u(x) > 0 \ \ \forall\, x \in \Omega,\ \ \frac{\partial u}{\partial n}(x) < 0 \ \ \forall\, x \in \partial\Omega\}$$

(n being the outward unit normal to $\partial\Omega$). Moreover, on the Sobolev space $W_0^{1,2}(\Omega)$, we consider the norm

$$\|u\| = \left(\int_\Omega |\nabla u(x)|^2 dx\right)^{1/2}.$$

Denote by λ_1 the first eigenvalue of the Laplace operator $(-\Delta, W_0^{1,2}(\Omega))$ and by $\varphi_1 \in \mathrm{int}C_+$ the first positive normalized eigenfunction.

We recall that the problem

$$\begin{cases} -\Delta u = \lambda u^{-\gamma}, & \text{in } \Omega \\ u > 0, & \text{in } \Omega \\ u = 0, & \text{on } \partial\Omega \end{cases}$$

admits a unique classical solution, i.e. a function $u_\lambda \in C^2(\Omega) \cap C(\overline{\Omega})$ for any positive γ. Actually, there exists $\varepsilon_\lambda > 0$ with $u_\lambda \geq \varepsilon_\lambda \varphi_1$ in Ω (see [8, Lemma A.4]) and $u_\lambda \in \mathrm{int}C_+$ (see [8, Lemmas A.6, A.7, B.1]).

But if $\gamma > 1$, $u_\lambda \notin C^1(\overline{\Omega})$ and $u_\lambda \in W_0^{1,2}(\Omega)$ if and only if $\gamma < 3$. Thus, for singular problem a classical solution might not be a weak solution.

As usual, for $u \in W_0^{1,2}(\Omega)$, $u_+ = \max\{u, 0\}$ and $u_- = \max\{-u, 0\}$ belong to $W_0^{1,2}(\Omega)$.

Let X be a Banach space, X^* be its dual and $\mathscr{F} \in C^1(X)$. We say that $\mathscr{F}$ satisfies the *Palais-Smale condition*, if the following is true: "Every sequence $\{u_n\}_{n\geq 1} \subseteq X$ such that

$$\{\mathscr{F}(u_n)\}_{n\geq 1} \text{ is bounded} \quad \text{and} \quad \mathscr{F}'(u_n) \to 0 \ \text{ in } \ X^* \ \text{ as } \ n \to \infty,$$

admits a strongly convergent subsequence".

3 Three Solutions

As far as we know the only contributions concerning the existence of three solutions for singular problem is the paper by Zhao et al. [22], where the application of an abstract "three critical points" theorem provides, under technical assumptions, the existence of multiple solutions to a double parameter problem.

Our contribution, in such framework, provides three solutions under more natural assumptions, namely the superlinearity of f at zero and its sublinearity at infinity. Also, our result holds in the higher dimensional case.

The proof is based on a very careful application of an abstract result of Ricceri [17] ensuring the existence of two local minimizers which turn out to be weak solutions of the problem according to the very general definition given here. The existence of the third solution is obtained by applying the well known Mountain Pass Theorem of Pucci and Serrin [16] to an appropriate truncation of the energy functional.

For the double eigenvalue problem

$$\begin{cases} -\Delta u = \lambda u^{-\gamma} + \mu f(u), & \text{in } \Omega \\ u > 0, & \text{in } \Omega \\ u = 0, & \text{on } \partial\Omega \end{cases} \qquad (\mathscr{P}_{\lambda,\mu})$$

we prove the following:

Theorem 3.1 *([5, Theorem 1.2]) Let* $\gamma \in]0,1[$ *and* $f : [0,+\infty[\to \mathbb{R}$ *be a continuous function with* $f(0) = 0$ *and* $f(t) > 0$ *for all* $t > 0$*. Suppose that there exist* $c > 0$ *and* $1 < q < 2^*$ *such that*

$$f(t) \le c(1 + t^{q-1}), \quad \text{for all } t \ge 0. \tag{13}$$

Let also assume the following conditions:

(H_1) $\displaystyle\lim_{t\to 0^+} \frac{F(t)}{t^2} = 0;$

(H_2) $\displaystyle\lim_{t\to +\infty} \frac{F(t)}{t^2} = 0.$

Set

$$\mu^* = \frac{1}{2}\inf\left\{ \frac{\int_\Omega |\nabla u(x)|^2\,dx}{\int_\Omega F(u(x))\,dx} : \int_\Omega F(u(x))\,dx > 0 \right\}.$$

Then, for each compact interval $[a,b] \subset]\mu^*,+\infty[$*, there exists* $r > 0$ *with the following property: for every* $\mu \in [a,b]$*, there exists* $\lambda^* > 0$ *such that for each* $\lambda \in [0,\lambda^*]$*, the problem* $(\mathscr{P}_{\lambda,\mu})$ *has at least three weak solutions belonging to* $int(C_0^1(\overline{\Omega})_+)$ *whose norms are less than* r*.*

The proof of the above results relies on the following abstract result by Ricceri:

Theorem A *([17, Theorem 4]) Let* (X,τ) *be a Hausdorff topological space, and* $P, Q : X \to \mathbb{R}$ *two sequentially lower semicontinuous functions. Assume that there is* $\sigma > \inf_X P$ *such that the set* $\overline{P^{-1}(]-\infty,\sigma[)}$ *is compact and first countable. Moreover, assume that there is a strict local minimum of P, say* x_0 *such that* $\inf_X P < P(x_0) < \sigma$*. Then, there exists* $\delta > 0$ *such that for each* $\mu \in [0,\delta]$*,* *the function $P + \mu Q$ has at least two τ_P local minimizers lying in $P^{-1}(]-\infty, \sigma[)$, where τ_P denotes the smallest topology on X which contains both τ and the family of sets $\{P^{-1}(]-\infty, \rho[)\}_{\rho\in\mathbb{R}}$.*

Denote by $\Phi, J, \Psi : W_0^{1,2}(\Omega) \to \mathbb{R}$ the functionals defined by

$$\Phi(u) = \frac{1}{2}\|u\|^2, \qquad \Psi(u) = \frac{1}{1-\gamma}\int_\Omega u_+^{1-\gamma}dx, \qquad J(u) = \int_\Omega F(u)dx.$$

Define also the energy functional associated to the problem $(\mathscr{P}_{\lambda,\mu})$, i.e. the functional $\mathscr{E} : W_0^{1,2}(\Omega) \to \mathbb{R}$ given by

$$\mathscr{E}(u) = \Phi(u) - \lambda\Psi(u) - \mu J(u).$$

Proposition 3.1 *([5, Propositions 2.1, 2.2]) Assume* (13) *and let* $\lambda, \mu > 0$. *If u is a local minimum of $\mathscr{E}$, then it is a weak solution of problem $(\mathscr{P}_{\lambda,\mu})$. Every weak solution of problem $(\mathscr{P}_{\lambda,\mu})$ belongs to $C^{1,\beta}(\overline{\Omega}) \cap \operatorname{int}C_+$, for some $\beta \in (0, 1)$.*

Proof Let $\rho > 0$ such that $\mathscr{E}(u) \leq \mathscr{E}(v)$ for every $v \in B_\rho(u)$. We claim that $u > 0$ a.e. in Ω.

For $t \in]0, 1[$ small enough, one has $u + tu_- \in B_\rho(u)$ and $(u + tu_-)_+ = u_+$. So,

$$\begin{aligned}
0 &\leq \frac{\mathscr{E}(u + tu_-) - \mathscr{E}(u)}{t}\\
&= \frac{1}{2}\left(\frac{\|u + tu_-\|^2 - \|u\|^2}{t}\right) - \mu\int_\Omega \frac{F(u + tu_-) - F(u)}{t}\\
&- \frac{\lambda}{1-\gamma}\int_\Omega \frac{(u + tu_-)_+^{1-\gamma} - u_+^{1-\gamma}}{t}\\
&= \frac{1}{2}\left(\frac{\|u + tu_-\|^2 - \|u\|^2}{t}\right) \to \int_\Omega \nabla u\nabla u_- = -\|u_-\|^2, \quad \text{as } t \to 0^+.
\end{aligned}$$

(Recall that $f(z) = 0$, for all $z \leq 0$).

From the above computation, it follows that $u_- = 0$, so $u \geq 0$ a.e. in Ω.

Assume that there exists a set of positive measure A such that $u = 0$ in A. Let $\varphi : \Omega \to \mathbb{R}$ be a function in $W_0^{1,2}(\Omega)$, positive in Ω. For $t > 0$ small enough, the function $u + t\varphi \in B_\rho(u)$ and $(u + t\varphi)^{1-\gamma} > u^{1-\gamma}$ a.e. in Ω, so

$$\begin{aligned}
0 &\leq \frac{\mathscr{E}(u + t\varphi) - \mathscr{E}(u)}{t}\\
&= \frac{1}{2}\left(\frac{\|u + t\varphi\|^2 - \|u\|^2}{t}\right) - \mu\int_\Omega \frac{F(u + t\varphi) - F(u)}{t}
\end{aligned}$$

$$
\begin{aligned}
&-\frac{\lambda}{(1-\gamma)t^{\gamma}}\int_A \varphi^{1-\gamma}-\frac{\lambda}{1-\gamma}\int_{\Omega\setminus A}\frac{(u+t\varphi)^{1-\gamma}-u^{1-\gamma}}{t}\\
&<\frac{1}{2}\left(\frac{\|u+t\varphi\|^2-\|u\|^2}{t}\right)-\mu\int_{\Omega}\frac{F(u+t\varphi)-F(u)}{t}\\
&-\frac{\lambda}{(1-\gamma)t^{\gamma}}\int_A \varphi^{1-\gamma}\to-\infty \text{ as } t\to 0^+.
\end{aligned}
$$

The contradiction ensures that $u>0$. Let us prove now that

$$
u^{-\gamma}\varphi\in L^1(\Omega)\quad\text{for all }\varphi\in W_0^{1,2}(\Omega) \tag{14}
$$

and

$$
\int_{\Omega}\nabla u\nabla\varphi-\mu\int_{\Omega}f(u)\varphi-\lambda\int_{\Omega}u^{-\gamma}\varphi\geq 0\quad\text{for all }\varphi\in W_0^{1,2}(\Omega),\ \varphi\geq 0. \tag{15}
$$

Choose $\varphi\in W_0^{1,2}(\Omega),\varphi\geq 0$. Fix a decreasing sequence $\{t_n\}\subseteq]0,1]$ with $\lim_n t_n=0$. The functions

$$
h_n(x)=\frac{(u(x)+t_n\varphi(x))^{1-\gamma}-u(x)^{1-\gamma}}{t_n}
$$

are measurable, non-negative and $\lim_n h_n(x)=(1-\gamma)u(x)^{-\gamma}\varphi(x)$ for a.e. $x\in\Omega$. From Fatou's lemma, we deduce

$$
\int_{\Omega}u^{-\gamma}\varphi\leq\frac{1}{1-\gamma}\liminf_n\int_{\Omega}h_n. \tag{16}
$$

As above,

$$
\begin{aligned}
0&\leq\frac{\mathscr{E}(u+t_n\varphi)-\mathscr{E}(u)}{t_n}\\
&=\frac{1}{2}\frac{\|u+t_n\varphi\|^2-\|u\|^2}{t_n}-\mu\int_{\Omega}\frac{F(u+t_n\varphi)-F(u)}{t_n}-\frac{\lambda}{1-\gamma}\int_{\Omega}h_n
\end{aligned}
$$

so, from (16) passing to the liminf in the above inequality we deduce at once condition (14) (it is enough to prove the integrability for a nonnegative test function) and

$$
\lambda\int_{\Omega}u^{-\gamma}\varphi\leq\int_{\Omega}\nabla u\nabla\varphi-\mu\int_{\Omega}f(u)\varphi,
$$

which is claim (15).

Let $\varepsilon \in]0,1[$ such that $(1+t)u \in B_\rho$ for all $t \in [-\varepsilon, \varepsilon]$. The function $\tilde{\xi}(t) = \mathscr{E}((1+t)u)$ has a local minimum at zero and

$$\begin{aligned} 0 = \tilde{\xi}'(0) &= \lim_{t\to 0} \frac{\mathscr{E}((1+t)u) - \mathscr{E}(u)}{t} \\ &= \int_\Omega |\nabla u|^2 - \lambda \int_\Omega u^{1-\gamma} - \mu \int_\Omega f(u)u. \end{aligned}$$

So,

$$\int_\Omega |\nabla u|^2 = \lambda \int_\Omega u^{1-\gamma} + \mu \int_\Omega f(u)u. \tag{17}$$

Let $\varphi \in W_0^{1,2}(\Omega)$ and plug into (15) the test function $v = (u+\varepsilon\varphi)_+$. Hence, by using (17) we have

$$\begin{aligned} 0 \le &\int_{\{u+\varepsilon\varphi\ge 0\}} \nabla u \nabla(u+\varepsilon\varphi) - \lambda \int_{\{u+\varepsilon\varphi\ge 0\}} u^{-\gamma}(u+\varepsilon\varphi) \\ &-\mu \int_{\{u+\varepsilon\varphi\ge 0\}} f(u)(u+\varepsilon\varphi) \\ = &\int_\Omega |\nabla u|^2 + \varepsilon \int_\Omega \nabla u \nabla\varphi - \lambda \int_\Omega u^{1-\gamma} - \varepsilon\lambda \int_\Omega u^{-\gamma}\varphi \\ &-\mu \int_\Omega f(u)u - \varepsilon\mu \int_\Omega f(u)\varphi \\ &- \int_{\{u+\varepsilon\varphi< 0\}} |\nabla u|^2 - \varepsilon \int_{\{u+\varepsilon\varphi< 0\}} \nabla u\nabla\varphi + \lambda \int_{\{u+\varepsilon\varphi< 0\}} u^{-\gamma}(u+\varepsilon\varphi) \\ &+\mu \int_{\{u+\varepsilon\varphi< 0\}} f(u)(u+\varepsilon\varphi) \\ \le &\ \varepsilon \left[\int_\Omega \nabla u\nabla\varphi - \lambda \int_\Omega u^{-\gamma}\varphi - \mu \int_\Omega f(u)\varphi \right] \\ &-\varepsilon \int_{\{u+\varepsilon\varphi< 0\}} \nabla u \nabla\varphi. \end{aligned}$$

Notice that as $\varepsilon \to 0$, the measure of the set $\{u+\varepsilon\varphi < 0\} \to 0$, so

$$\int_{\{u+\varepsilon\varphi<0\}} \nabla u\nabla\varphi \to 0.$$

Hence, dividing by ε, and passing to the limit as $\varepsilon \to 0$, we get that

$$\int_{\Omega} \nabla u \nabla \varphi - \lambda \int_{\Omega} u^{-\gamma} \varphi - \mu \int_{\Omega} f(u)\varphi \geq 0.$$

From the arbitrariness of φ, we get at once that u is a weak solution of $(\mathscr{P}_{\lambda,\mu})$.

The regularity of u and the fact that $u \in \mathrm{int}C_+$ follow easily from Theorem B.1 of [8] and from the Strong Maximum Principle respectively. □

For $\lambda > 0$, let u_λ be the unique global minimizer of the functional

$$u \to \frac{1}{2}\|u\|^2 - \frac{\lambda}{1-\gamma} \int_{\Omega} u_+^{1-\gamma} dx.$$

Proposition 3.2 *([5, Proposition 2.3]) Assume* (13) *and let* $\lambda, \mu > 0$. *Define* $g : \Omega \times \mathbb{R} \to [0, +\infty)$ *and* $\tilde{\Psi}$, $\mathscr{F} : W_0^{1,2}(\Omega) \to \mathbb{R}$ *by*

$$g(x,t) = \begin{cases} t^{-\gamma}, & \text{if } \ x \in \Omega \ \text{ and } \ t \geq u_\lambda(x) \\ \\ u_\lambda(x)^{-\gamma}, & \text{if } \ x \in \Omega \ \text{ and } \ t \leq u_\lambda(x), \end{cases}$$

$$\tilde{\Psi}(u) = \int_{\Omega} \int_0^{u_+(x)} g(x,t)dtdx,$$

and

$$\mathscr{F}(u) = \frac{1}{2}\|u\|^2 - \lambda\tilde{\Psi}(u) - \mu J(u)$$

respectively. Then, $\mathscr{F} \in C^1(W_0^{1,2}(\Omega))$ *and the following hold:*

(a) if u_0 *is a critical point of* $\mathscr{F}$, *then* $u_0 \geq u_\lambda$ *a.e. in* Ω;
(b) if u_0 *is a critical point of* $\mathscr{F}$, *then it is a weak solution of* $(\mathscr{P}_{\lambda,\mu})$;
(c) if $u_0 \in \mathrm{int}C_+$ *is a local minimizer of* $\mathscr{F}$ *in the* $C_0^1(\overline{\Omega})$*-topology, then* u_0 *is also a local minimizer of* $\mathscr{F}$ *in the* $W_0^{1,2}(\Omega)$*-topology.*

Proof The fact that $\mathscr{F} \in C^1(W_0^{1,2}(\Omega))$ follows from the proof of Lemma A.3 of [8] and its derivative at u is given by

$$\langle \mathscr{F}'(u), \ \varphi \rangle = \int_{\Omega} \nabla u \nabla \varphi - \mu \int_{\Omega} f(x,u)\varphi - \lambda \int_{\Omega} g(x,u)\varphi$$

for every $\varphi \in W_0^{1,2}(\Omega)$.

(a) Let u_0 be a critical point of $\mathscr{F}$. Choosing $(u_0 - u_\lambda)_-$ as test function, one has

$$-\int_{\{u_0<u_\lambda\}} \nabla u_0 \cdot (\nabla u_0 - \nabla u_\lambda) + \int_{\{u_0<u_\lambda\}} [\mu f(x,u_0) + \lambda u_\lambda^{-\gamma}](u_0 - u_\lambda) = 0.$$

Bearing in mind that u_λ is a global minimum of $u \to \frac{1}{2}\|u\|^2 - \frac{\lambda}{1-\gamma}\int_\Omega u_+^{1-\gamma}$, we also obtain that

$$-\int_{\{u_0<u_\lambda\}} \nabla u_\lambda \cdot (\nabla u_0 - \nabla u_\lambda) + \int_{\{u_0<u_\lambda\}} \lambda u_\lambda^{-\gamma}(u_0 - u_\lambda) = 0.$$

Hence, subtracting the two equalities,

$$\int_{\{u_0<u_\lambda\}} (\nabla u_0 - \nabla u_\lambda) \cdot (\nabla u_0 - \nabla u_\lambda) = \int_{\{u_0<u_\lambda\}} \mu f(x, u_0)(u_0 - u_\lambda) \leq 0.$$

Thus, $u_0 \geq u_\lambda$ almost everywhere in Ω.

(b) It follows from (a).

(c) Assume that $u_0 \in \mathrm{int}C_+$ is a local minimizer of $\mathscr{F}$ in the $C_0^1(\overline{\Omega})$-topology. Then, for $\varphi \in C_0^1(\overline{\Omega})$ and t small, one has

$$\begin{aligned}0 &\leq \lim_{t\to 0} \frac{\mathscr{F}(u_0 + t\varphi) - \mathscr{F}(u_0)}{t} \\ &= \int_\Omega \nabla u_0 \nabla \varphi - \mu \int_\Omega f(x, u_0)\varphi - \lambda \int_\Omega g(x, u_0)\varphi.\end{aligned}$$

Rewriting the above inequality replacing φ with $-\varphi$ we obtain

$$\int_\Omega \nabla u_0 \nabla \varphi - \mu \int_\Omega f(x, u_0)\varphi - \lambda \int_\Omega g(x, u_0)\varphi = 0.$$

By density, u_0 is a critical point of $\mathscr{F}$ in $W_0^{1,2}(\Omega)$. Thus $u_0 \geq u_\lambda$ (see (a)).

Suppose on the contrary that u_0 is *not* a local minimizer of $\mathscr{F}$ in the $W_0^{1,2}(\Omega)$-topology.

Choose $r \in (q, 2^*)$ and consider the closed convex sets

$$S_n = \{u \in W_0^{1,2}(\Omega) : \frac{1}{r}||u - u_0||_r^r \leq \frac{1}{n}\}, \quad n \geq 1$$

(here $||\cdot||_r$ stands for the $L^r(\Omega)$-norm). Since $\mathscr{F}$ is sequentially weakly lower semi-continuous and coercive on S_n, we may find v_n, $n \geq 1$, such that

$$v_n \in S_n, \quad \mathscr{F}(v_n) = \min_{u\in S_n} \mathscr{F}(u), \quad \mathscr{F}(v_n) < \mathscr{F}(u_0), \quad n \geq 1. \tag{18}$$

Claim $v_n \geq u_\lambda$, for all $n \geq 1$.

Arguing indirectly, suppose that for some $n \geq 1$, we have $(u_\lambda - v_n)_+ \not\equiv 0$. Set

$$w_t = v_n + t(u_\lambda - v_n)_+, \quad \xi(t) = \mathscr{F}(w_t), \quad t \in [0, 1].$$

Then on $\{u_\lambda > v_n\}$, we have

$$w_t - u_\lambda = (1-t)(v_n - u_\lambda) < 0, \quad \text{for all } t \in]0,1[.$$

Therefore, for $t \in]0,1[$,

$$\xi'(t) = \langle \mathscr{F}'(w_t), (u_\lambda - v_n)_+\rangle$$

$$= \int_{\{u_\lambda > v_n\}} \nabla w_t \cdot (\nabla u_\lambda - \nabla v_n) - \lambda \int_{\{u_\lambda > v_n\}} f(x, w_t)(u_\lambda - v_n)$$

$$- \lambda \int_{\{u_\lambda > v_n\}} g(x, w_t)(u_\lambda - v_n) \ \leq$$

$$\leq \int_{\{u_\lambda > v_n\}} \nabla w_t \cdot (\nabla u_\lambda - \nabla v_n) - \lambda \int_{\{u_\lambda > v_n\}} u_\lambda^{-\gamma}(u_\lambda - v_n)$$

$$= - \int_{\{u_\lambda > v_n\}} (\nabla w_t \ - \ \nabla u_\lambda) \cdot (\nabla v_n - \nabla u_\lambda) \quad \text{(due to the choice of } u_\lambda),$$

so,

$$(1-t)\xi'(t) \leq - \int_{\{u_\lambda > v_n\}} (\nabla w_t \ - \ \nabla u_\lambda) \cdot (\nabla w_t - \nabla u_\lambda)$$

$$< 0 \quad \text{(by the strong monotonicity of the Laplacian operator)}.$$

Consequently, ξ is strictly decreasing on $[0,1]$. In particular, we have

$$\xi(1) < \xi(0) \Rightarrow \mathscr{F}(w_1) < \mathscr{F}(v_n).$$

But since $u_0 \geq u_\lambda$, we may check that $|w_1 - u_0| \leq |v_n - u_0|$. Thus, $w_1 \in S_n$, which contradicts the fact that v_n is a global minimizer of $\mathscr{F}$ on S_n and finishes the proof of the claim.

Then the Lagrange multiplier rule gives rise to a sequence $k_n, \ n \geq 1$ such that

$$\mathscr{F}'(v_n) = k_n E'(v_n), \quad n \geq 1,$$

where $E(u) = ||u - u_0||_r^r / r \ , \ u \in W_0^{1,2}(\Omega)$.

Now the above claim combined with the definition of g yields that for all $n \geq 1$,

$$\begin{cases} -\Delta v_n(x) = \lambda f(x, v_n(x)) + \lambda v_n(x)^{-\gamma} + k_n|v_n(x) - u_0(x)|^{r-2}(\, v_n(x) - u_0(x)\,), \\ \qquad\qquad \text{a.e. in } \Omega, \\ v_n\,|_{\partial\Omega} = 0. \end{cases}$$

We also remark that $k_n \leq 0,\ n \geq 1$. Indeed, for each $n \geq 1$, the function

$$\zeta_n(t) = \mathscr{F}((1-t)v_n + tu_0), \quad t \in [0, 1]$$

attains its minimum at $t_0 = 0$ so, $\zeta_n'(0) \geq 0 \ \Rightarrow\ \langle \mathscr{F}'(v_n), u_0 - v_n\rangle \geq 0$, which implies $k_n||v_n - u_0||_r^r \leq 0$ and thus, $k_n \leq 0$.

Then we proceed as in the proof of Theorem 1.1 of [8, p. 701], to reach a contradiction. □

Proof of Theorem 3.1 We are going to apply Theorem A with $X = W_0^{1,2}(\Omega)$ and τ the weak topology on $W_0^{1,2}(\Omega)$. Let us prove that

$$\lim_{u\to 0} \frac{J(u)}{\Phi(u)} = 0. \tag{19}$$

Fix $\varepsilon > 0$. Hypothesis (H_1) together with the subcritical growth of f imply that for some constant $c_\varepsilon > 0$ and $\theta \in]\max\{2,\ q\},\ 2^*[$,

$$0 \leq F(t) \leq \frac{\varepsilon}{2}|t|^2 + c_\varepsilon|t|^\theta, \ \text{ for all } \ t \in \mathbb{R}.$$

It follows that for some $c_\varepsilon' > 0$,

$$0 \ \leq\ \frac{J(u)}{\Phi(u)} \ \leq\ \frac{\varepsilon}{\lambda_1} \ +\ c_\varepsilon'\ ||u||^{\theta-2}, \quad \text{for all } \ u \in W_0^{1,2}(\Omega) \setminus \{0\}.$$

Then

$$\limsup_{u\to 0} \frac{J(u)}{\Phi(u)} \leq \varepsilon$$

and since $\varepsilon > 0$ is arbitrary, (19) follows.

From (H_2), we easily deduce that

$$\lim_{\|u\|\to+\infty} \frac{J(u)}{\Phi(u)} = 0. \tag{20}$$

Set, for all $\mu > 0$,

$$P_\mu = \Phi - \mu J.$$

The functional P_μ is sequentially weakly lower semicontinuous and coercive (see (20)), whereas 0 turns out to be a (strong) strict local minimizer of P_μ (see (19)).

From [18, Theorem C] we get that 0 is a local minimizer of P_μ in the weak topology. Moreover, by the definition of μ^* (see the statement of Theorem 3.1), we obtain that for every $\mu > \mu^*$, 0 is not a global minimizer of P_μ. In fact, $\inf_X P_\mu < P_\mu(0) = 0$.

We point out that $\mu^* > 0$. Indeed, there exists a constant $c > 0$ such that

$$F(t) \leq c|t|^2 \text{ for all } t \in \mathbb{R}.$$

Thus,

$$\int_\Omega F(u(x))\,dx \leq c\|u\|_2^2 \leq c'\|u\|^2,$$

where $c' > 0$ involves also the Sobolev embedding constant. This implies that $\mu^* > 0$.

To proceed, fix $[a, b] \subset]\mu^*, +\infty[$ and choose $\sigma > 0$.

From the coercivity of P_μ it clearly follows that the sets $\overline{P_\mu^{-1}(]-\infty, \sigma[)}^{w}$ are compact and metrizable (thus, first countable) with respect to the weak topology. (Recall that the weak closure of a bounded subset of a separable reflexive Banach space is compact and metrizable with respect to the weak topology.) Notice that

$$\bigcup_{\mu\in[a,b]} \{u \in X : P_\mu(u) < \sigma\} \subseteq \{u \in X : \Phi(u) - bJ(u) < \sigma\} \subseteq B_\eta,$$

for some positive radius η (this follows from the fact that $J(u) \geq 0$ for every $u \in W_0^{1,2}(\Omega)$ and the coercivity of $\Phi - bJ$). Put also $c^* = \sup_{B_\eta}(\Phi - aJ)$ and let $r > \eta$ such that

$$\bigcup_{\mu\in[a,b]} \{u \in X : P_\mu(u) \leq c^* + 2\} \subseteq B_r. \tag{21}$$

Next, choose $\mu \in [a, b]$. Note that Ψ is sequentially weakly continuous but not differentiable in X since $0 < \gamma < 1$.

In order to obtain the uniform estimate of the norm of our solutions we need to introduce a function $\alpha \in C^1(\mathbb{R})$, bounded, such that $\alpha(t) = t$ for every t such that $|t| \leq \sup_{B_{2r}} \Psi$. Therefore,

$$(\alpha \circ \Psi)(u) = \Psi(u) \qquad \text{for every } u \in B_{2r}. \tag{22}$$

Now Theorem A guarantees the existence of some $\delta = \delta(\mu) > 0$ such that for every $\lambda \in [0, \delta]$, $P_\mu - \lambda(\alpha \circ \Psi)$ has two local minimizers in the τ_{P_μ} topology, say u_1, u_2, such that

$$u_1,\ u_2 \in P_\mu^{-1}(]-\infty, \sigma[) \ \subseteq B_\eta \ \subseteq \ B_r\ . \tag{23}$$

Since P_μ is continuous, the topology τ_{P_μ} is weaker than the strong topology and u_1, u_2 turn out to be local minimizers of the functional

$$\mathscr{E}_\alpha : X \to \mathbb{R}, \qquad \mathscr{E}_\alpha(u) = \frac{1}{2}\|u\|^2 - \lambda(\alpha \circ \Psi)(u) - \mu J(u).$$

Notice that if $\|u - u_i\| < r$, then, $\|u\| < \|u_i\| + r < 2r$ for $i = 1, 2$. Therefore, since from (22), $\mathscr{E}_\alpha = \mathscr{E}$ in B_{2r}, u_1, u_2 turn out to be local minimizers of $\mathscr{E}$.

Put $\lambda^* = \lambda^*(\mu) = \min\{\delta, (\sup_{\mathbb{R}} \alpha)^{-1}\}$ and fix $\lambda \in [0, \lambda^*]$.

From Proposition 3.1, u_1 and u_2 are weak solutions of $(\mathscr{P}_{\lambda,\mu})$ belonging to $\mathrm{int}C_+ \cap C^{1,\beta}(\overline{\Omega})$ for some $\beta \in]0, 1[$.

The existence of a third solution is obtained via regularization methods.

For $\lambda \in [0, \lambda^*]$, let $g, \tilde{\Psi}, \mathscr{F}$ as in Proposition 3.2 and let

$$\mathscr{F}_\alpha : W_0^{1,2}(\Omega) \to \mathbb{R}, \qquad \mathscr{F}_\alpha(u) = \frac{1}{2}\|u\|^2 - \lambda(\alpha \circ \tilde{\Psi})(u) - \mu J(u).$$

It is clear that since $g(x, t) \leq t^{-\gamma}$ for every $t > 0$, if $\|u\| \leq 2r$, one has

$$\tilde{\Psi}(u) \leq \Psi(u) \leq \sup_{B_{2r}} \Psi,$$

and $(\alpha \circ \tilde{\Psi})(u) = \tilde{\Psi}(u)$, so that $\mathscr{F}_\alpha$ coincides with $\mathscr{F}$ in B_{2r}.

From the strong comparison principle for singular problems (Theorem 2.3 of [9]), we deduce that $u_1 - u_\lambda \in \mathrm{int}C_+$ and $u_2 - u_\lambda \in \mathrm{int}C_+$. (Recall that $f(t) > 0$, for $t > 0$.)

Moreover, u_1 is a $C_0^1(\overline{\Omega})$-local minimizer of $\mathscr{E}$ and since $u_1 - u_\lambda \in \mathrm{int}C_+$ and $\mathrm{int}C_+$ is open in the $C_0^1(\overline{\Omega})$-topology, there exists a neighborhood V of u_1 in this topology such that $V \subseteq u_\lambda + \mathrm{int}C_+$ and $\mathscr{E}(u) \geq \mathscr{E}(u_1)$ for all $u \in V$.

Notice that for every $u \in u_\lambda + \mathrm{int}C_+$, we have that

$$\begin{aligned}
\mathscr{E}(u) = \frac{1}{2}\|u\|^2 - \lambda \int_\Omega \int_0^{u_\lambda(x)} t^{-\gamma} dt dx - \lambda \int_\Omega \int_{u_\lambda(x)}^{u(x)} t^{-\gamma} dt dx - \\
\lambda \int_\Omega \int_0^{u_\lambda(x)} u_\lambda(x)^{-\gamma} dt dx + \lambda \int_\Omega \int_0^{u_\lambda(x)} u_\lambda(x)^{-\gamma} dt dx - \mu J(u) = \\
\mathscr{F}(u) - \lambda \int_\Omega \int_0^{u_\lambda(x)} t^{-\gamma} dt dx + \lambda \int_\Omega \int_0^{u_\lambda(x)} u_\lambda(x)^{-\gamma} dt dx = \\
\mathscr{F}(u) + \text{const}.
\end{aligned}$$

By virtue of the above equality we obtain that u_1 is a $C_0^1(\overline{\Omega})$-local minimizer of $\mathscr{F}$. But then Proposition 3.2 implies that u_1 is also a $W_0^{1,2}(\Omega)$-local minimizer of $\mathscr{F}$. Similarly, u_2 turns out to be a $W_0^{1,2}(\Omega)$-local minimizer of $\mathscr{F}$. Moreover, since for every $\|u\| < 2r$ one has $\mathscr{F}(u) = \mathscr{F}_\alpha(u)$, u_1 and u_2 are actually $W_0^{1,2}(\Omega)$-local minimizers of $\mathscr{F}_\alpha$.

The functional $\mathscr{F}_\alpha$ is of class C^1 in $W_0^{1,2}(\Omega)$. Indeed, since $u_\lambda \geq \varepsilon_\lambda \varphi_1$, the functional $\tilde{\Psi}$ is of class C^1 in $W_0^{1,2}(\Omega)$. Therefore, $\alpha \circ \tilde{\Psi} \in C^1(W_0^{1,2}(\Omega))$ and the same is true for $\mathscr{F}_\alpha$. Also, since $\mathscr{F}_\alpha$ is coercive, it verifies in a standard way the well known Palais-Smale condition.

By Theorem 1 of [16] there exists a critical point for $\mathscr{F}_\alpha$, say u_3 such that

$$\mathscr{F}_\alpha(u_3) = \inf_{\gamma \in S} \sup_{t \in [0,1]} \mathscr{F}_\alpha(\gamma(t)),$$

where

$$S = \{\gamma \in C^0([0,1], W_0^{1,2}(\Omega)) : \ \gamma(0) = u_1, \ \gamma(1) = u_2\}.$$

In particular, if $\tilde{\gamma}(t) = tu_1 + (1-t)u_2, \ t \in [0,1]$, then $\tilde{\gamma} \in S$ and

$$\tilde{\gamma}(t) \in B_\eta, \quad \text{for all } \ t \in [0,1].$$

(Recall that $u_1, \ u_2 \in B_\eta$ (see (23)).)

So, by the definition of c^* and λ^*, one has

$$\begin{aligned}\mathscr{F}_\alpha(u_3) &\leq \sup_{t \in [0,1]} \mathscr{F}_\alpha(\tilde{\gamma}(t)) \\ &\leq \sup_{u \in B_\eta} [\Phi(u) - aJ(u)] + \mu^* \sup_{u \in B_\eta} (\alpha \circ \tilde{\Psi})(u) \\ &\leq c^* + 1.\end{aligned}$$

Therefore,

$$P_\mu(u_3) = \Phi(u_3) - \mu J(u_3) \leq c^* + 1 + \lambda(\alpha \circ \tilde{\Psi})(u_3) \leq c^* + 2,$$

and from (21)

$$u_3 \in B_r.$$

It is clear that u_3 is a critical point of $\mathscr{F}$ and from Proposition 3.2, $u_3 \geq u_\lambda$. Thus, from Propositions 3.1 and 3.2, $u_3 \in \operatorname{int} C_+$ is a positive solution of problem $(\mathscr{P}_{\lambda,\mu})$ and the proof is concluded. □

The above result has been extended in [3] with the aid of nonsmooth critical point theory to any exponent γ provided that the nonlinearity is multiplied by a suitable positive function a (see also assumption i) in Theorem 1.6).

Theorem 3.2 *([3, Theorem 1.1]) Let $\gamma > 0$, $f : [0, +\infty[\to \mathbb{R}$ be a continuous function with $f(0) = 0$ and $f(t) > 0$ for all $t > 0$ and $a : \Omega \to \mathbb{R}$ be a positive measurable function. Suppose that there exist $c > 0$ and $1 < q < 2^*$ such that*

$$f(t) \leq c(1+t^{q-1}), \quad \textit{for all } t \geq 0.$$

Let also assume conditions (H_1), (H_2) *and*

$(\widetilde{H})$ *there exists* $\bar{u} \in C_0^1(\overline{\Omega})$ *such that* $\bar{u} > 0$ *on* Ω *and* $a\bar{u}^{-\gamma} \in L^{(2^*)'}(\Omega)$.

Set

$$\mu^* = \frac{1}{2} \inf \left\{ \frac{\int_\Omega |\nabla u(x)|^2 \, dx}{\int_\Omega F(u(x)) \, dx} : \int_\Omega F(u(x)) \, dx > 0 \right\}.$$

Then, for each $\mu > \mu^*$ *there exists* $\lambda^* > 0$ *such that for each* $\lambda \in [0, \lambda^*]$, *the problem*

$$\begin{cases} -\Delta u = \lambda a(x) u^{-\gamma} + \mu f(u), & \textit{in } \Omega \\ u > 0, & \textit{in } \Omega \\ u = 0, & \textit{on } \partial\Omega \end{cases}$$

has at least three weak solutions.

In the proof of the above result we combine the abstract multiplicity result by Ricceri with techniques from non smooth analysis. To the best of our knowledge, this is the first contribution establishing three solutions in the higher dimensional case and for any $\gamma > 0$. It is to be mentioned that in such framework we can not expect to have solutions in $C^1(\overline{\Omega})$. Also, the arbitrariness of γ does not allow us to prove the uniform (with respect to λ) boundedness of the solutions.

4 Multiple Solutions

The existence of k solutions, with $k > 3$, as far as we know, has never been studied in the literature. Because of the presence for singular boundary value problems of the singular term, higher multiplicity results are not expected (when the perturbation f is zero, problem $(\mathscr{P}_\lambda)$ has a unique solution for every λ). We follow the approach of [12] where the existence of an arbitrarily big number of solutions is proved for a nonsingular perturbed semilinear elliptic problem involving oscillatory term. We propose here a multiplicity result under a suitable oscillatory behaviour of f at zero. Since the singularity blows up as $u \to 0$, we need to control it by reducing the range of the parameter λ. Also,we prove the existence of infinitely many solutions when the nonlinearity f exhibits an oscillatory behaviour at infinity. Such stronger result is motivated by the fact that the singularity plays a "minor" role as it tends to zero as $u \to \infty$ (see [4]).

Theorem 4.1 *([4, Theorem 1.5]) Let* $\gamma \in]0,1[$ *and* $f : [0,+\infty[\to \mathbb{R}$ *be a continuous function. Assume the following conditions:*

(H_3) there exists a sequence $\{t_n\} \subset \mathbb{R}^+$ such that $t_n \to 0^+$ and $f(t_n) < 0$ for every $n \in \mathbb{N}$;

(H_4) $-\infty < \liminf_{t\to 0^+} \frac{F(t)}{t^2} \leq \limsup_{t\to 0^+} \frac{F(t)}{t^2} = +\infty$.

Then, for each $k \in \mathbb{N}$, there exists $\lambda_k^\star > 0$ such that, for every $0 < \lambda < \lambda_k^\star$, problem $(\mathscr{P}_\lambda)$ has at least k essentially bounded weak solutions.

The proof of the above theorem is based on the following preliminary result.

Lemma 4.1 *([4, Lemma 4.1]) Let $f : [0, +\infty[\to \mathbb{R}$ be a continuous function. For $\lambda > 0$ assume that there exist $0 < a < b$ such that*

$$f(t) + \lambda t^{-\gamma} \leq 0 \qquad \textit{for every } t \in [a, b].$$

Define $h_\lambda :]0, +\infty[\to \mathbb{R}$ by

$$h_\lambda(t) = \begin{cases} f(t) + \lambda t^{-\gamma} & \textit{if } 0 < t < a \\ f(a) + \lambda a^{-\gamma} & \textit{if } t \geq a \end{cases}$$

and set

$$H_\lambda(t) = \int_0^{t_+} h_\lambda(s)ds, \quad t \in \mathbb{R}.$$

Then, the functional $\mathscr{E}_\lambda : W_0^{1,2}(\Omega) \to \mathbb{R}$ defined by

$$\mathscr{E}_\lambda(u) = \frac{1}{2}\|u\|^2 - \int_\Omega H_\lambda(u(x))dx$$

has a global minimizer $u_\lambda \in W_0^{1,2}(\Omega) \bigcap L^\infty(\Omega)$ such that $\|u_\lambda\|_\infty \leq a$. Moreover, u_λ turns out to be a weak solution of problem $(\mathscr{P}_\lambda)$.

Proof Let $F : [0, +\infty[\to \mathbb{R}$, be the function $F(s) = \int_0^{s_+} f(t)dt$. Since $0 < \gamma < 1$, H_λ is well defined and continuous on $\mathbb{R}$. In particular,

$$H_\lambda(t) = \begin{cases} 0 & \text{if } t \leq 0 \\ F(t) + \dfrac{\lambda}{1-\gamma} t^{1-\gamma} & \text{if } 0 < t < a \\ H_\lambda(a) + h_\lambda(a)(t-a) & \text{if } t \geq a. \end{cases}$$

Moreover, $H_\lambda(t_+) = H_\lambda(t)$, for all $t \in \mathbb{R}$ and $(H_\lambda)'(t) = h_\lambda(t)$, for all $t > 0$.

The functional $\mathscr{E}_\lambda$ is well defined on $W_0^{1,2}(\Omega)$, sequentially weakly lower semicontinuous and coercive. Thus, it has a global minimizer u_λ.

We can assume that $u_\lambda \leq a$. Indeed, if

$$v_\lambda = \begin{cases} u_\lambda & \text{if } 0 < u_\lambda < a \\ a & \text{if } u_\lambda \geq a, \end{cases}$$

then $v_\lambda \in W_0^{1,2}(\Omega)$ and in view of $h_\lambda(a) \le 0$ we have the following inequality

$$\mathscr{E}(u_\lambda) = \frac{1}{2}\int_\Omega |\nabla u_\lambda|^2 - \int_{\{u_\lambda \le a\}} H_\lambda(u) - \int_{\{u_\lambda > a\}} H_\lambda(a)$$
$$-h_\lambda(a)\int_{\{u_\lambda > a\}} (u_\lambda - a) \ge \mathscr{E}(v_\lambda).$$

The proof that u_λ is a weak solution of problem $(\mathscr{P}_\lambda)$ follows as in Proposition 3.1. □

Remark 4.1 If $\lambda = 0$, h can be defined in zero and the above conclusion holds with u non negative weak solution of $(\mathscr{P}_0)$.

Proof of Theorem 4.1 The proof of this result closely follows the idea of Theorem 1.2 of [12]. For completeness we give the details.

From (H_4) there exist $M_0 < 0$ and $\delta > 0$ such that

$$\frac{F(t)}{t^2} > M_0, \qquad \text{for every } 0 < t < \delta.$$

Fix $x_0 \in \Omega$ and $0 < r < R$ such that $\overline{B(x_0, R)} \subset \Omega$. Choose $M_1 > 0$ large enough such that

$$\frac{1}{2}\omega_N \frac{(R^N - r^N)}{(R-r)^2} - M_1\omega_N r^N - M_0\omega_N(R^N - r^N) < 0,$$

where ω_N is the volume of the unit ball in $\mathbb{R}^N$.

Hypothesis (H_4) also enables us to choose a sequence of positive numbers $\{\xi_n\}$ such that

$$\xi_n \to 0^+, \quad \frac{F(\xi_n)}{\xi_n^2} > M_1, \qquad \text{for every } n \in \mathbb{N}.$$

By virtue of hypothesis (H_3) and by continuity, we can construct three sequences of positive numbers $\{a_n\}$, $\{b_n\}$ and $\{\lambda_n\}$ such that $a_n \to 0^+$, $b_n \to 0^+$, $\lambda_n \downarrow 0^+$, $a_n < b_n < a_{n-1}$, $\xi_n \le a_n < \delta$ for all n and

$$f(t) + \lambda t^{-\gamma} \le 0 \qquad \text{for every } t \in [a_n, b_n],\ \lambda \in [0, \lambda_n],\ \ n \in \mathbb{N}.$$

In particular, we deduce that

$$f(t) \le 0 \qquad \text{for every } t \in [a_n, b_n],\ n \in \mathbb{N}.$$

For every $n \in \mathbb{N}$ and for each $\lambda \in [0, \ \lambda_n]$, define $h_{n,\lambda} :]0, +\infty[\to \mathbb{R}$ by

$$h_{n,\lambda}(t) = \begin{cases} f(t) + \lambda t^{-\gamma} & \text{if } 0 < t < a_n \\ f(a_n) + \lambda a_n^{-\gamma} & \text{if } t \geq a_n \end{cases}$$

and

$$H_{n,\lambda}(t) = \int_0^{t_+} h_{n,\lambda}(s)ds, \quad t \in \mathbb{R}.$$

Denote by $\mathscr{E}_{n,\lambda} : W_0^{1,2}(\Omega) \to \mathbb{R}$ the functionals defined by

$$\mathscr{E}_{n,\lambda}(u) = \frac{1}{2}\|u\|^2 - \int_\Omega H_{n,\lambda}(u(x))dx$$

and notice that, if $\|u\|_\infty \leq a_n$,

$$\mathscr{E}_{n,\lambda}(u) = \mathscr{E}_{n,0}(u) - \frac{\lambda}{1-\gamma}\int_\Omega u^{1-\gamma}.$$

From Lemma 4.1 we deduce that for each $n \in \mathbb{N}$ and for each $\lambda \in [0, \lambda_n]$, there exists a global minimizer of $\mathscr{E}_{n,\lambda}$, denoted by $u_{n,\lambda}$, such that $\|u_{n,\lambda}\|_\infty \leq a_n$, which is also a weak solution of $(\mathscr{P}_\lambda)$. Notice that since $a_{n+1} < a_n$, we have that for $\lambda < \lambda_{n+1}$, $\mathscr{E}_{n,\lambda}(u_{n,\lambda}) \leq \mathscr{E}_{n,\lambda}(u_{n+1,\lambda}) = \mathscr{E}_{n+1,\lambda}(u_{n+1,\lambda})$.

Applying again Lemma 4.1 for $\lambda = 0$ and Remark 4.1, we deduce also the existence of a sequence $\{u_{n,0}\}$ of non negative weak solutions of the following problem

$$\begin{cases} -\Delta u = f(u), & \text{in } \Omega \\ u = 0, & \text{in } \partial\Omega \end{cases} \tag{$\mathscr{P}_0$}$$

such that for each $n \in \mathbb{N}$, $u_{n,0}$ is a global minimizer of the functional $\mathscr{E}_{n,0}$ with $\|u_{n,0}\|_\infty \leq a_n$.

We prove now that up to a subsequence, $\{u_{n,0}\}$ has pairwise distinct terms.

Define on Ω the continuous functions w_n, $n \in \mathbb{N}$ by

$$w_n(x) = \begin{cases} \xi_n & \text{if } x \in B(x_0, r) \\ \xi_n \dfrac{R - |x - x_0|}{R - r} & \text{if } x \in B(x_0, R) \setminus B(x_0, r) \\ 0 & \text{if } x \in \Omega \setminus B(x_0, R). \end{cases}$$

Then, $w_n \in W_0^{1,2}(\Omega)$, $0 \le w_n \le \xi_n \le a_n < \delta$ and $\|w_n\|^2 = \omega_N \frac{(R^N-r^N)}{(R-r)^2}\xi_n^2$. Thus,

$$\begin{aligned}
\mathscr{E}_{n,0}(w_n) &= \frac{1}{2}\omega_N \frac{(R^N-r^N)}{(R-r)^2}\xi_n^2 - \int_\Omega F(w_n) \\
&= \frac{1}{2}\omega_N \frac{(R^N-r^N)}{(R-r)^2}\xi_n^2 - \int_{B(x_0,r)} F(\xi_n) - \int_{B(x_0,R)\setminus B(x_0,r)} F(w_n) \\
&\le \left[\frac{1}{2}\omega_N \frac{(R^N-r^N)}{(R-r)^2} - M_1\omega_N r^N - M_0\omega_N(R^N-r^N)\right]\xi_n^2 < 0
\end{aligned}$$

by the above choice of M_1. Thus, $\mathscr{E}_{n,0}(u_{n,0}) \le \mathscr{E}_{n,0}(w_n) < 0$ for every $n \in \mathbb{N}$. Moreover, from the inequalities

$$0 > \mathscr{E}_{n,0}(u_{n,0}) \ge -a_n \max_{[0,a_1]} |f||\Omega|,$$

we deduce that

$$\lim_n \mathscr{E}_{n,0}(u_{n,0}) = \lim_n \mathscr{E}_{n,0}(w_n) = 0.$$

From above we conclude that there exists a subsequence which we still denote by $\{u_{n,0}\}$ of pairwise distinct solutions of $(\mathscr{P}_0)$.

Choose now, as in [12], an increasing sequence $\{\theta_n\}$ of negative numbers tending to zero, such that

$$\theta_n < \mathscr{E}_{n,0}(u_{n,0}) < \theta_{n+1}, \quad n \in \mathbb{N}.$$

We notice that

$$\mathscr{E}_{n,\lambda}(u_{n,\lambda}) \le \mathscr{E}_{n,\lambda}(u_{n,0}) < \mathscr{E}_{n,0}(u_{n,0}) < \theta_{n+1},$$

and

$$\mathscr{E}_{n,\lambda}(u_{n,\lambda}) = \mathscr{E}_{n,0}(u_{n,\lambda}) - \frac{\lambda}{1-\gamma}\int_\Omega (u_{n,\lambda})^{1-\gamma} \ge \mathscr{E}_{n,0}(u_{n,0}) - \frac{\lambda}{1-\gamma}|\Omega| a_n^{1-\gamma}, \ \ n \in \mathbb{N}.$$

For each $n \in \mathbb{N}$ and since $\mathscr{E}_{n,0}(u_{n,0}) > \theta_n$, we can choose

$$\lambda < \min\{\lambda_n, \tilde{\lambda}_n\}$$

where

$$\tilde{\lambda}_n = (1-\gamma)\frac{\mathscr{E}_n(u_{n,0}) - \theta_n}{a_n^{1-\gamma}|\Omega|}$$

to get

$$\theta_n < \mathscr{E}_{n,\lambda}(u_{n,\lambda}) < \theta_{n+1}. \tag{24}$$

Fix $k \in \mathbb{N}$ and set $\lambda_k^\star = \min\{\lambda_1, \lambda_2, \dots \lambda_k, \tilde{\lambda}_1, \tilde{\lambda}_2, \dots \tilde{\lambda}_k\}$. If $0 < \lambda \leq \lambda_k^\star$, then the functions $u_{1,\lambda}, u_{2,\lambda}, \dots, u_{k,\lambda}$ are weak solutions of problem $(\mathscr{P}_\lambda)$. They are distinct. Indeed, if $u_{i,\lambda} = u_{j,\lambda}$ for some $i < j$, then $\mathscr{E}_{i,\lambda}(u_{i,\lambda}) = \mathscr{E}_{i,\lambda}(u_{j,\lambda}) = \mathscr{E}_{j,\lambda}(u_{j,\lambda})$, against (24). The proof is concluded. □

Example Define

$$f(t) = \begin{cases} \sqrt{t}\max\{0, \sin\frac{1}{t}\} + t^2\min\{0, \sin\frac{1}{t}\} & \text{if } t > 0 \\ \\ 0 & \text{if } t = 0. \end{cases}$$

In the next result we prove the existence of a sequence of solutions avoiding the parameter (i.e. putting $\lambda = 1$) as the singular term itself gives a small contribution at infinity. However we need to strengthen the sign condition on f.

For the problem

$$\begin{cases} -\Delta u = u^{-\gamma} + f(u), & \text{in } \Omega \\ u > 0, & \text{in } \Omega \\ u = 0, & \text{on } \partial\Omega \end{cases} \tag{$\mathscr{P}$}$$

we prove

Theorem 4.2 *([4, Theorem 1.6]) Let* $\gamma \in]0, 1[$, $f : [0, +\infty[\to \mathbb{R}$ *be a continuous function. Assume the following conditions:*

(H$_5$*) there exist* $l < 0$ *and a sequence* $\{t_n\} \subset \mathbb{R}^+$ *such that* $t_n \to +\infty$ *and* $f(t_n) \leq l$*;*

(H$_6$*)* $-\infty < \liminf\limits_{t\to+\infty} \dfrac{F(t)}{t^2} \leq \limsup\limits_{t\to+\infty} \dfrac{F(t)}{t^2} = +\infty.$

Then, there exists a sequence $\{u_n\}$ *of essentially bounded weak solutions of* $(\mathscr{P})$ *such that* $\lim_n \|u_n\|_\infty = +\infty$.

Proof From (H_6) there exist $M_0 < 0$ and $\delta > 0$ such that

$$\frac{F(t)}{t^2} > M_0 \qquad \text{for every } t > \delta.$$

Fix $x_0 \in \Omega$ and $0 < r < R$ such that $\overline{B(x_0, R)} \subset \Omega$, and choose M_1 and a sequence $\{\xi_n\} \subset \mathbb{R}^+$ with $\xi_n \to +\infty$ such that

$$\frac{1}{2}\omega_N \frac{(R^N - r^N)}{(R-r)^2} - M_1\omega_N r^N - M_0\omega_N(R^N - r^N) < 0$$

and

$$\frac{F(\xi_n)}{\xi_n^2} > M_1 \qquad \text{for every } \ n \in \mathbb{N}.$$

Eventually passing to a subsequence we can suppose that $\delta < \xi_n \leq t_n$ for every $n \in \mathbb{N}$ and

$$f(t_n) + t_n^{-\gamma} < 0 \qquad \text{for every } n \in \mathbb{N}$$

(see hypothesis (H_5)).

By continuity we can construct two sequences of positive numbers $\{a_n\}$ and $\{b_n\}$ such that $a_n \to +\infty$, $b_n \to +\infty$, $a_n < b_n < a_{n+1}$, $\xi_n \leq a_n$ and

$$f(t) + t^{-\gamma} \leq 0 \qquad \text{for every } \ t \in [a_n, b_n], \ n \in \mathbb{N}.$$

For every $n \in \mathbb{N}$ define $h_n :]0, +\infty[\to \mathbb{R}$ by

$$h_n(t) = \begin{cases} f(t) + t^{-\gamma} & \text{if } 0 < t < a_n \\ f(a_n) + a_n^{-\gamma} & \text{if } t \geq a_n \end{cases}$$

and

$$H_n(t) = \int_0^{t_+} h_n(s)ds.$$

Then, according to Lemma 4.1, the functional $\mathscr{E}_n : W_0^{1,2}(\Omega) \to \mathbb{R}$ defined by

$$\mathscr{E}_n(u) = \frac{1}{2}\|u\|^2 - \int_\Omega H_n(u(x))dx$$

has a global minimizer u_n such that $\|u_n\|_\infty \leq a_n$. Also, u_n turns out to be a weak solution of $(\mathscr{P})$.

Let us prove that $\lim_n \mathscr{E}_n(u_n) = -\infty$. Observe that the sequence $\{\mathscr{E}_n(u_n)\}$ is decreasing. Indeed, for every $n \in \mathbb{N}$, since $\|u_n\|_\infty \leq a_n < a_{n+1}$,

$$\mathscr{E}_{n+1}(u_{n+1}) \leq \mathscr{E}_{n+1}(u_n) = \mathscr{E}_n(u_n).$$

As before, define on Ω the continuous functions w_n, $n \in \mathbb{N}$ by

$$w_n(x) = \begin{cases} \xi_n & \text{if } x \in B(x_0, r) \\ \xi_n \dfrac{R - |x - x_0|}{R - r} & \text{if } x \in B(x_0, R) \setminus B(x_0, r) = D \\ 0 & \text{if } x \in \Omega \setminus B(x_0, R). \end{cases}$$

Then $w_n \in W_0^{1,2}(\Omega)$, $0 \le w_n \le \xi_n \le a_n$ and $\|w_n\|^2 = \omega_N \frac{(R^N-r^N)}{(R-r)^2}\xi_n^2$ for all $n \in \mathbb{N}$.

Moreover, for all $n \in \mathbb{N}$, we have

$$\begin{aligned}
\mathscr{E}_n(w_n) &= \frac{1}{2}\|w_n\|^2 - \int_\Omega H_n(w_n)dx \\
&= \frac{1}{2}\omega_N \frac{(R^N - r^N)}{(R-r)^2}\xi_n^2 - \int_\Omega F(w_n) - \frac{1}{-\gamma+1}\int_\Omega w_n^{-\gamma+1} \\
&< \frac{1}{2}\omega_N \frac{(R^N - r^N)}{(R-r)^2}\xi_n^2 - \int_{B(x_0,r)} F(\xi_n) - \int_{D\cap\{w_n>\delta\}} F(w_n) - \int_{D\cap\{w_n\le\delta\}} F(w_n) \\
&\le \left[\frac{1}{2}\omega_N \frac{(R^N - r^N)}{(R-r)^2} - M_1\omega_N r^N - M_0\omega_N(R^N - r^N)\right]\xi_n^2 + \omega_N(R^N - r^N)\max_{[0,\delta]}|F|.
\end{aligned}$$

By the choice of M_1, $\lim_n \mathscr{E}_n(w_n) = -\infty$, which immediately implies $\lim_n \mathscr{E}_n(u_n) = -\infty$. In particular, by passing eventually to a subsequence, we may assume that u_n, $n \in \mathbb{N}$, are pairwisely distinct.

Finally, suppose that $\{\|u_n\|_\infty\}$ is bounded, i.e. there exists a constant M_2 such that $\|u_n\|_\infty \le M_2$ for all $n \in \mathbb{N}$. Fix $\bar{n}$ such that $a_{\bar{n}} > M_2$. Then for every $n \ge \bar{n}$, we have $u_n < a_{\bar{n}} \le a_n$, so $H_{\bar{n}}(u_n(\cdot)) = H_n(u_n(\cdot))$ and hence,

$$\mathscr{E}_n(u_n) = \mathscr{E}_{\bar{n}}(u_n) \ge \mathscr{E}_{\bar{n}}(u_{\bar{n}}),$$

which is in contradiction with the previous limit.

It follows that $\{\|u_n\|_\infty\}$ is unbounded so, we may extract a subsequence which tends to $+\infty$, as $n \to \infty$. The proof is concluded. □

Example Define for $t \ge 0$

$$f(t) = t^2(1/2 + \sin t).$$

References

1. M. Coclite, G. Palmieri, On a singular nonlinear Dirichlet problem. Comm. Partial Differ. Equ. **14**, 1315–1327 (1989)
2. M.G. Crandall, P.H. Rabinowitz, L. Tartar, On a Dirichlet problem with a singular nonlinearity. Comm. Partial Differ. Equ. **2**, 193–222 (1977)
3. F. Faraci, G. Smyrlis, Three solutions for a class of higher dimensional singular problems. Nonlinear Differ. Equ. Appl. **23**, 14 (2016)
4. F. Faraci, G. Smyrlis, On a singular semilinear elliptic problem: multiple solutions via critical point theory. Topol. Methods Nonlinear Anal. (2018, to appear)
5. F. Faraci, G. Smyrlis, Three solutions for a singular quasilinear elliptic problem. Proc. Edinb. **657** Math. Soc. (2018, to appear)
6. W. Fulks, J.S. Maybee, A singular non-linear equation. Osaka Math. J. **12**, 1–19 (1960)

7. M. Ghergu, V. Rădulescu, *Singular Elliptic Problems: Bifurcation and Asymptotic Analysis*. Oxford Lecture Series in Mathematics and Its Applications, vol. 37 (Oxford University Press, Oxford, 2008)
8. J. Giacomoni, K. Saoudi, $W_0^{1,p}$ versus C^1 local minimizer for a singular and critical functional. J. Math. Anal. Appl. **363**, 697–710 (2010)
9. J. Giacomoni, I. Schindler, P. Takàč, Sobolev versus Hölder minimizers and global multiplicity for a singular and quasilinear equation. Ann. Sc. Norm. Super. Pisa Cl. Sci. **6**, 117–158 (2007)
10. S.M. Gomes, On a singular nonlinear elliptic problem. SIAM J. Math. Anal. **17**, 1359–1369 (1986)
11. N. Hirano, C. Saccon, N. Shioji, Brezis - Nirenberg type theorems and multiplicity of positive solutions for a singular elliptic problem. J. Differ. Equ. **245**, 1997–2037 (2008)
12. A. Kristály, Detection of arbitrarily many solutions for perturbed elliptic problems involving oscillatory terms. J. Differ. Equ. **245**, 3849–3868 (2008)
13. A. Lair, A. Shaker, Classical and weak solutions of a singular semilinear elliptic problem. J. Math. Anal. Appl. **211**, 371–385 (1997)
14. A. Lazer, P.J. McKenna, On a singular nonlinear elliptic boundary value problem. Proc. AMS **111**, 721–730 (1991)
15. K. Perera, E.A.B. Silva, Existence and multiplicity of positive solutions for singular quasilinear problems. J. Math. Anal. Appl. **323**, 1238–1252 (2006)
16. P. Pucci, J. Serrin, A mountain pass theorem. J. Differ. Equ. **60**, 142–149 (1985)
17. B. Ricceri, Sublevel sets and global minima of coercive functionals and local minima of their perturbation. J. Nonlinear Convex Anal. **5**, 157–168 (2004)
18. B. Ricceri, A further three critical points theorem. Nonlinear Anal. **71**, 4151–4157 (2009)
19. J. Shi, M. Yao, On a singular nonlinear semilinear elliptic problem. Proc. Royal Soc. Edinb. Sect. A **128**, 1389–1401 (1998)
20. Y. Sun, S. Wu, Y. Long, Combined effects of singular and superlinear nonlinearities in some singular boundary value problems. J. Differ. Equ. **176**, 511–531 (2001)
21. Z. Zhang, Critical points and positive solutions of singular elliptic boundary value problems. J. Math. Anal. Appl. **302**, 476–483 (2005)
22. L. Zhao, Y. He, P. Zhao, The existence of three positive solutions of a singular p-Laplacian problem. Nonlinear Anal. **74**, 5745–5753 (2011)

The Pilgerschritt (Liedl) Transform on Manifolds

Wolfgang Förg-Rob

1 Introduction: The Main Idea

Starting point was the question of finding iterative roots respectively iteration groups. Therefore, let X be an arbitrary set and $f : X \to X$ a bijective mapping. The problem is to find an iteration group $(f_t : X \to X)_{t \in \mathbb{R}}$ such that $f_0 = \mathrm{id}_X$ and $f_{t+s} = f_t \circ f_s$ for $t, s \in \mathbb{R}$. As it is (and was) well known, in general this problem has no solution.

Roman Liedl, a mathematician in Innsbruck, had an idea for topological groups in the late seventies of the last century. He came up with an idea to solve the problem of finding homomorphisms (one parameter subgroups through a given element) for such groups and introduced a transform he called “Pilgerschritt transform”, because it reminded to a method of medieval pilgrims to enlarge the way of pilgrimage (two steps forward, one step back). The background for his ideas is given by the Volterra product integral and questions in iteration theory.

In Lie groups for endpoints not ‘far away’ from the unit, there is a usual tool available: The logarithm as the inverse of the exponential. However, the logarithm is given by a (convergent) power series, and using this series up to a given power this gives an approximation of the homomorphism through that point in the group G.

The new idea of Roman Liedl was the following: Choose an *arbitrary* path $\varphi : [0, 1] \to G$ connecting the unit e with the given element g, and transform the path in a deterministic way to a path $\widetilde{\varphi}$, such that $\widetilde{\varphi}$ is the restriction of the achieved homomorphism $h : \mathbb{R} \to G$, or at least the sequence $\varphi, \widetilde{\varphi}, \widetilde{\widetilde{\varphi}}, \ldots$ converges to this restriction.

W. Förg-Rob (✉)
Institute of Mathematics, University of Innsbruck, Innsbruck, Austria
e-mail: wolfgang.foerg-rob@uibk.ac.at

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_11

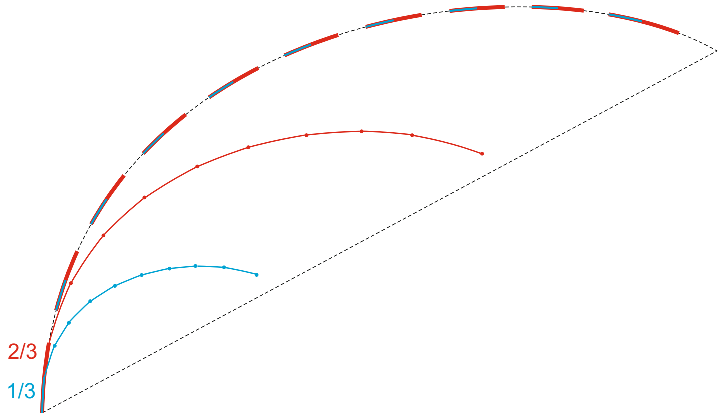
2 The Pilgerschritt Transform on Groups

Let G be a topological group, $g \in G$ and $\varphi : [0, 1] \to G$ a (continuous) path connecting the unit element e with the given element g, i.e., $\varphi(0) = e$ and $\varphi(1) = g$. Then we define a new path $\widetilde{\varphi}$ by the following process (compare it with a similarity deformation!).
Let $\mathscr{Z} = (0 = t_0 < t_1 < \ldots < t_m = 1)$ be a partition of the interval $[0, 1]$, and let $\tau \in [0, 1]$ be a real number. The Pilgerschritt product with respect to $\mathscr{Z}$ and τ is given as the product

$$\pi(\varphi, \mathscr{Z}, \tau) = \Big(\varphi(t_{m-1} + \tau(t_m - t_{m-1})) \cdot \varphi(t_{m-1})^{-1}\Big) \cdot \ldots \cdot \Big(\varphi(t_0 + \tau(t_1 - t_0)) \cdot \varphi(t_0)^{-1}\Big)$$

If the limit of this expression $\pi(\varphi, \mathscr{Z}, \tau)$ exists, when the mesh size of $\mathscr{Z}$ tends to 0, this limit will be called the Pilgerschritt transform of φ, i.e.

$$\widetilde{\varphi}(\tau) = \lim_{|\mathscr{Z}| \to 0} \pi(\varphi, \mathscr{Z}, \tau).$$



For a detailed description we refer to the original papers [4–6].
Of course, the question arises when this limit does exist. Suppose that the group G has a differentiable structure (Lie group or Banach Lie group), and the path φ is continuously differentiable. A Taylor expansion of the product terms in π gives

$$\begin{aligned} &\big(\varphi(t_{k-1} + \tau(t_k - t_{k-1})) \cdot \varphi(t_{k-1})^{-1}\big) = \\ &= \big(\varphi(t_{k-1}) + \tau(t_k - t_{k-1})\, \varphi'(t_{k-1}) + R\big) \cdot \varphi(t_{k-1})^{-1} = \\ &= e + \tau(t_k - t_{k-1})\, \varphi'(t_{k-1})\, \varphi(t_{k-1})^{-1} + R_1 \end{aligned}$$

where R respectively R_1 denote remainder terms. An easy calculation shows that for taking the limit $|\mathscr{Z}| \to 0$ these remainder terms are negligible, and we end up with the product integral in the sense of Volterra and with coefficient function $t \mapsto \tau\varphi'(t)\, \varphi(t)^{-1}$.

Another interpretation can be made by looking at the differential equation

$$y' = \tau\varphi'(t)\ \varphi(t)^{-1} \cdot y.$$

A usual Euler method gives rise to the product above (without remainder terms). Thus in Lie groups and Banach Lie groups the Pilgerschritt transform of a $\mathscr{C}^1$-path φ may be defined equivalently by the following process:

1. Solve the differential equation $y' = \tau\varphi'(t)\ \varphi(t)^{-1} \cdot y$ for $\tau \in [0, 1]$ and the initial condition $y(0) = e$.
2. Denote this solution by $\widehat{\varphi}(t, \tau)$.
3. Put $\widetilde{\varphi}(\tau) = \widehat{\varphi}(1, \tau)$.

In order to give an answer to the question: Whenever φ is a $\mathscr{C}^1$-path in a Lie group or a Banach Lie group, the Pilgerschritt transform exists.
First results on properties of the transformed path $\widetilde{\varphi}$ can be given as follows.

1. $\widetilde{\varphi}(0) = \varphi(0), \widetilde{\varphi}(1) = \varphi(1)$
2. If φ is the restriction of a homomorphism $h : \mathbb{R} \to G$ to the interval $[0, 1]$, then $\widetilde{\varphi} = \varphi$
3. If φ is the restriction of a homomorphism $h : \mathbb{R} \to G$ to the interval $[0, 1]$ up to a transform of the parameter, then $\widetilde{\varphi} = h|[0, 1]$
4. $\widetilde{\varphi}$ is homotopic to φ in the group G
5. $\widetilde{\varphi}$ is a $\mathscr{C}^\infty$-function

Answers to the question whether the sequence $\varphi, \widetilde{\varphi}, \widetilde{\widetilde{\varphi}}, \ldots$ converges to the restriction of a homomorphism could be given, here is a short overview:

1. If the group G is abelian, then $\widetilde{\varphi}$ itself is the restriction of the homomorphism.
2. If the group G is nilpotent, then the sequence ends up with the restriction of a homomorphism after a finite number of steps.
3. If the group G is solvable, then the sequence converges to the restriction of a homomorphism under the condition that the endpoint $\varphi(1)$ is ‘close’ to e.
4. In general, the sequence converges to the restriction of a homomorphism under the condition that $\varphi(t)$ is close to e and $\varphi'(t)$ is small.

For results in detail on the Pilgerschritt transform on Lie groups see for example the overview report [2] and the cited literature there.

3 The Pilgerschritt Transform on Manifolds

While working on the Pilgerschritt transform, R. Liedl asked the members of his group at the university of Innsbruck, whether this method could be used also to compute geodesic lines on manifolds. The background of his question was very simple: One-parameter-subgroups of Lie groups are just the geodesic lines according to the connection on the Lie group defined by right resp. left translation. In the paper [1] a suitable definition (according to the description of this method on Lie groups) and the proof of (local) existence of this transform could be given.

The main questions which arise are the following:

- What are the properties of the path $\widetilde{\varphi}$?
- Does the sequence $\varphi, \widetilde{\varphi}, \widetilde{\widetilde{\varphi}}, \widetilde{\widetilde{\widetilde{\varphi}}}, \ldots$ exist?
- Does the sequence $\varphi, \widetilde{\varphi}, \widetilde{\widetilde{\varphi}}, \widetilde{\widetilde{\varphi}}, \ldots$ converge to a geodesic line connecting starting point and endpoint of φ?

We need a manifold M ($\mathscr{C}^\infty$) with a linear connection ∇. The manifold is modelled over $\mathbb{R}^d$—but it also may be over an arbitrary Banach space, the proofs need not be changed in the later case. Furthermore, $\gamma\colon [0, 1] \to M$ is a given path (we assume it to be continuously differentiable) on this manifold with starting point $s = \gamma(0)$ and endpoint $p = \gamma(1)$. On $\mathbb{R}^d$ we use a norm $\| - \|$ (all the norms are equivalent), on a Banach space V the given norm, and on the space of (linear, multilinear) operators the induced norm.

The original definition on groups was given:
Let $\mathscr{Z} = (0 = t_0 < t_1 < \ldots < t_m = 1)$ be a partition of the interval $[0, 1]$, and let $\tau \in [0, 1]$ be a real number. The Pilgerschritt product with respect to $\mathscr{Z}$ and τ is given as the product

$$\begin{aligned}&\pi(\varphi, \mathscr{Z}, \tau) = \\ &= \left(\varphi(t_{m-1} + \tau(t_m - t_{m-1})) \cdot \varphi(t_{m-1})^{-1}\right) \cdot \ldots \cdot \left(\varphi(t_0 + \tau(t_1 - t_0)) \cdot \varphi(t_0)^{-1}\right).\end{aligned}$$

But there are several problems: What should the product be on a manifold? Remembering that the right and left connection on a Lie group are defined by the multiplication (i.e., transportation via multiplication), we must replace the product of elements by parallel transport. On the other hand parallel transport can be defined on manifolds by the connection ∇ along paths—which path should it be for the multiplication?
Therefore we go back to the equivalent definition on groups via differential equation and the parallel transport.

Let us remind the definition of 'parallel' on manifolds:
Let $\gamma\colon [a, b] \to M$ be a $\mathscr{C}^1$-path and $X\colon [a, b] \to TM$ be a $\mathscr{C}^1$-field tangent along γ (that means that $X(t) \in T_{\gamma(t)}M$ for all $t \in [a, b]$). Then X is called parallel along γ, if $\nabla_{\gamma'(t)}X(t) = 0$ for all $t \in [a, b]$.

Using a local coordinate system mapping an open set in M to an open set $U \subseteq \mathbb{R}^d$ the linear connection ∇ is given on U by the so called Christoffel symbols Γ_{ij}^k. As we do not need to have a Riemann connection, we assume that the functions $\Gamma_{ij}^k\colon U \to \mathbb{R}$ are $\mathscr{C}^\infty$, but they need not to be symmetric, i.e., in general we have $\Gamma_{ij}^k(m) \neq \Gamma_{ji}^k(m)$ for $m \in U$.

For sake of simplicity (to avoid a lot of indices) we use the following terminology: For each $m \in U$ denote by $G(m)$ the bilinear function

$$G(m)\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d\colon (u, w) \mapsto \left(\sum_{i=1}^d \sum_{j=1}^d \Gamma_{ij}^k(p)\, u_i w_j\right)_{1 \le k \le d}.$$

If we now denote by η the expression of γ in coordinates and similarly v that one of the vector field X, the condition of 'parallel' may be expressed as the differential equation

$$v'(t) = -G(\eta(t))(\eta'(t), v(t)) \quad \text{for all } t \in [a, b] \text{ such that } \eta(t) \text{ is defined.}$$

As for a given curve γ this equation locally is a *linear* differential equation (in the coordinate expression v), usual theorems on linear differential equations give rise to the following results:

(a) For any $t_0 \in [a, b]$ and any $X_0 \in T_{\gamma(t)}M$ there exists a unique parallel vector field X along γ, such that $X(t_0) = X_0$.
(b) The mapping (*parallel displacement along* γ) $P_{\gamma,t_1,t_0} \colon T_{\gamma(t_0)}M \to T_{\gamma(t_1)}M \colon X_0 \mapsto X(t_1)$—where X is the unique vector field parallel along γ with $X(t_0) = X_0$—is a linear isomorphism, whose inverse is given by P_{γ,t_0,t_1}.

With this notation we use an idea of Kobayashi-Nomizu (cf. [3]) and define

Definition Let $\gamma \colon [0, 1] \to M$ be a $\mathscr{C}^1$-path. Then define

$$\operatorname{der}(\gamma) \colon [0, 1] \to T_{\gamma(0)}M \colon t \mapsto P_{\gamma,0,t}(\gamma'(t)).$$

Of course, $\operatorname{der}(\gamma)$ is a continuous function. The notion 'der' has been chosen according to the expression 'derivative'.

A first interesting result is the following:

Theorem 1 *Let* $\gamma_1, \gamma_2 \colon [0, 1] \to M$ *be* $\mathscr{C}^1$*-paths such that* $\gamma_1(0) = \gamma_2(0)$. *If* $\operatorname{der}(\gamma_1)(t) = \operatorname{der}(\gamma_2)(t)$ *for all* $t \in [0, 1]$, *then* $\gamma_1(t) = \gamma_2(t)$ *for all* $t \in [0, 1]$.

Proof By the usual arguments on compactness and connectedness of the interval $[0, 1]$ it suffices to show that γ_1 and γ_2 coincide in a neighbourhood of 0. Thus choose a coordinate system near $\gamma_1(0) = \gamma_2(0)$. We use the expressions η_1 respectively η_2 for the paths in these coordinates,
$f(t) = \operatorname{der}(\gamma_1)(t) = \operatorname{der}(\gamma_2)(t) \in T_{\gamma_1(0)}M$ with coordinate expression $g(t)$, and let $R_1(t)$ be the coordinate expression for the parallel displacement $P_{\gamma_1,t,0}$ along γ_1 and similarly $R_2(t)$ for γ_2.
Then in a neighbourhood of 0 we have the equations (for $i \in \{1, 2\}$)

$$\begin{aligned} \eta_i'(t) &= R_i(t)(g(t)) \\ R_i'(t) &= -G(\eta_i(t))(\eta_i'(t), -) \circ R_i(t) \end{aligned}$$

that is, we have

$$\begin{aligned} \eta_i'(t) &= R_i(t)(g(t)) \\ R_i'(t) &= -G(\eta_i(t))(R_i(t)(g(t)), -) \circ R_i(t) \end{aligned}$$

But this means, that both pairs (η_1, R_1) and (η_2, R_2) fulfill the same system of differential equation (with differentiable right hand side) to the same initial values, as $\eta_1(0) = \eta_2(0)$ and $R_1(0) = \mathrm{Id} = R_2(0)$. Therefore, the functions are equal.

This result just gave uniqueness. On the other hand, suppose that for some point $s \in M$ a continuous function $f: [0, 1] \to T_sM$ is given. Does there exist a $\mathscr{C}^1$-path γ such that $\gamma(0) = s$ and $\mathrm{der}(\gamma) = f$? In general the answer will be "no", but locally we can state the following

Theorem 2 *Let $s \in M$. Then there exists a neighbourhood W of $0 \in T_sM$ such that for all continuous functions $f: [0, 1] \to W$ a $\mathscr{C}^1$-path γ exists with the properties $\gamma(0) = s$ and $\mathrm{der}(\gamma) = f$.*

Proof Fix a coordinate system x of M around s—and let U denote the image in $\mathbb{R}^d$. We may assume that s is mapped to $0 \in U$, and G describes the linear connection ∇ on U.
We use the notation of the preceding theorem. Then we have to show that for 'small' continuous functions $g: [0, 1] \to \mathbb{R}^d$ the system of differential equations

$$\begin{aligned} \eta'(t) &= R(t)(g(t)) \\ R'(t) &= -G(\eta(t))(R(t)(g(t)), -) \circ R(t) \end{aligned}$$

has a solution to the initial value $\eta(0) = 0$, $R(0) = \mathrm{Id}$ which is defined on the whole interval $[0, 1]$.

The usual proofs of the theorem by Picard-Lindelöf give information about the interval of existence of the solutions if we know bounds and Lipschitz constants for the right hand side: Let $r > 0$ be such that the ball $B_r(0)$ is contained in U, the function G is bounded on $B_r(0)$ by a constant M_1 and the derivative G' is bounded by M_2. Furthermore, let K be the ball with radius $\frac{1}{2}$ around Id in the space of linear functions from $\mathbb{R}^d$ to $\mathbb{R}^d$. Then all the elements of K are invertible and their norm is bounded by $\frac{3}{2}$. Now we have a look on the set $B_r(0) \times K$ and the right hand side of the differential equation

$$(\sigma, \varrho) \mapsto (\varrho(g(t)), -G(\sigma)(\varrho(g(t))) \circ \varrho).$$

Using the sum norm on $\mathbb{R}^d \times L(\mathbb{R}^d, \mathbb{R}^d)$ and the mean value theorem we get
$L = \max(\{1+3M_1, \frac{9}{4} M_2\}) \cdot \|g\|_\infty$ as a Lipschitz constant (according to the theorem of Picard-Lindelöf) and

$$M = \left(\frac{3}{2} + \frac{9}{4} M_1\right) \cdot \|g\|_\infty \text{ as an upper bound for the right hand side.}$$

Thus if we choose

$$\varepsilon = \frac{1}{\frac{3}{2} + \frac{9}{4} M_1} \cdot \min(\{r, \frac{1}{2}\}),$$

for any continuous function g such that $\|g\|_\infty < \varepsilon$, a solution of the differential equation exists which is defined on $[0, 1]$.

The system of differential equations we just dealt with gives rise to two interesting applications:

1. A $\mathscr{C}^1$-path $\gamma\colon [0,1] \to M$ is a geodesic line iff the function $\operatorname{der}(\gamma)$ is constant.
2. Suppose that $(\gamma_n)_{n\in\mathbb{N}}\colon [0,1] \to M$ is a sequence of $\mathscr{C}^1$-paths starting at the same point $s = \gamma_n(0)$ (for all n) such that all the images of these paths are contained in a neighbourhood of s homeomorphic to a ball in the $\mathbb{R}^d$ via a coordinate system and the sequence $(\operatorname{der}(\gamma_n))_{n\in\mathbb{N}}$ converges *uniformly* to a function

   $f\colon [0,1] \to T_sM$. Then the sequence $(\gamma_n)_{n\in\mathbb{N}}$ converges uniformly to a path $\gamma\colon [0,1] \to M$, and $\operatorname{der}(\gamma) = f$.

After this preparation we are able to define the Pilgerschritt transform of a path—according to the definition in Lie groups via a differential equation:

Definition Let $\gamma\colon [0,1] \to M$ be a $\mathscr{C}^1$-path, and let be

$$f = \operatorname{der}(\gamma)\colon [0,1] \to T_{\gamma(0)}M.$$

If there exists a $\mathscr{C}^1$-path $\widehat{\gamma}(\tau,-)\colon [0,1] \to M$ for each $\tau \in [0,1]$ such that

$$\widehat{\gamma}(\tau,0) = \gamma(0) \text{ and } \operatorname{der}(\widehat{\gamma}(\tau,-)) = \tau\cdot f$$

then we call the function

$$\widetilde{\gamma}\colon [0,1] \to M\colon \tau \mapsto \widehat{\gamma}(\tau,1)$$

the Pilgerschritt transform of the path γ.

In order to proof that the Pilgerschritt sequence $\gamma, \widetilde{\gamma}, \widetilde{\widetilde{\gamma}}, \ldots$ converges (under suitable conditions) to a geodesic line, we will show that the sequence $(f_n)_{n\in\mathbb{N}}$ with $f = f_0 = \operatorname{der}(\gamma)$, $f_1 = \widetilde{f} = \operatorname{der}(\widetilde{\gamma})$, $f_2 = \widetilde{f_1} = \operatorname{der}(\widetilde{\widetilde{\gamma}}), \ldots$ converges uniformly to a constant.

First of all we start with some elementary properties of the transformed path:

Theorem 3 *Let $\gamma\colon [0,1] \to M$ be a $\mathscr{C}^1$-path such that $\widehat{\gamma}(\tau,t)$ exists for all $t,\tau \in [0,1]$. Then we have*

1. *$\widetilde{\gamma}(0) = \gamma(0)$ and $\widetilde{\gamma}(1) = \gamma(1)$.*
2. *$\widetilde{\gamma}$ is a $\mathscr{C}^\infty$-function.*
3. *If γ is a geodesic line, then $\widetilde{\gamma} = \gamma$.*

These properties follow immediately from the defining differential equations.

From now on we want to deal with the sequence $(f_0 = f, f_1 = \widetilde{f}, f_2 = \widetilde{f_1}, \ldots)$. Therefore, $s \in M$ is our starting point for the paths, $f\colon [0,1] \to T_sM$ is continuous, and we use the notation as in Theorem 2 in order to guarantee the existence of paths according to f. Furthermore, η denotes the coordinate expression of a path γ, and g the coordinate expression of f.

Thus we will deal with the coordinate expressions of our functions.

The proof of Theorem 2 gives us a constant ε, such that for any continuous function $g\colon [0,1] \to \mathbb{R}^d$ with the property $\|g\|_\infty < \varepsilon$ there exists a $\mathscr{C}^1$-path η with $g = \mathrm{der}(\eta)$ (here and in the further we mix up the 'der'-notation with its coordinate expression, but there should be no reason of confusion).

If such a function $g\colon [0,1] \to \mathbb{R}^d$ fulfills the property $\|g\|_\infty < \varepsilon$, then this inequality also holds for each of the functions $\tau \cdot g$ for $\tau \in [0,1]$ which implies that the Pilgerschritt transform exists, and we may compute $\widetilde{g}$. As we know from Theorem 3, $\widetilde{g}$ is a $\mathscr{C}^\infty$-function. In order to study the behaviour of the sequence $g_1 = g, g_2 = \widetilde{g}, g_3 = \widetilde{\widetilde{g}}, \ldots$ without loss of generality we may assume that g itself is a $\mathscr{C}^\infty$-function.

Now we want to compute $\widetilde{g}$ in some sense "directly" from g. In Theorem 2 we had used the system of differential equations

$$\begin{aligned} \eta'(t) &= R(t)(g(t)) \\ R'(t) &= -\,G(\eta(t))(R(t)(g(t)), -) \circ R(t) \end{aligned}$$

in order to compute the path η from the function g. Now we add a parameter $\tau \in [0,1]$ in order to describe the construction of $\widehat{\eta}$ from $\tau \cdot g$:

$$\begin{aligned} \frac{\partial}{\partial t}\widehat{\eta}(\tau,t) &= \tau\, R(\tau,t)(g(t)) \\ \frac{\partial}{\partial t}R(\tau,t) &= -\tau\, G(\eta(t))(R(\tau,t)(g(t)), -) \circ R(\tau,t) \end{aligned}$$

This is just to produce $\widehat{\eta}$. At the end, we have to transport back parallel along the curve $\tau \mapsto \widehat{\eta}(\tau,t)$ the vector $\dfrac{\partial}{\partial \tau}\widehat{\eta}(\tau,t)$, especially for $t = 1$, because this will give the Pilgerschritt transform $\widetilde{g}$.

In order to distinguish between parallel transport along lines $\tau = \text{const.}$, which we denoted by $R(\tau,t)$, we use the notation $S(\tau,t)$ to denote the parallel transport along the lines $t = \text{const.}$ Thus we get

$$\widehat{g}(\tau,t) = S(\tau,t)^{-1}\left(\frac{\partial}{\partial \tau}\widehat{\eta}(\tau,t)\right) \text{ and finally } \widetilde{g}(\tau) = \widehat{g}(\tau,1).$$

As a first result we get an expression for $\widetilde{g}(0)$:

Theorem 4 *Let $g\colon [0,1] \to \mathbb{R}^d$ fulfill the property $\|g\|_\infty < \varepsilon$. Then the Pilgerschritt transform $\widetilde{g}$ exists, and $\widetilde{g}(0) = \int\limits_0^1 g(t)\,\mathrm{d}t$.*

Proof The existence of $\widetilde{g}$ was just discussed before.
Now $\widehat{\eta}(0,t) = 0 = \eta(0)$, and therefore $S(0,t) = \mathrm{Id}$ for $t \in [0,1]$. Thus

$$\widetilde{g}(0) = \frac{\partial}{\partial \tau}\widehat{\eta}(0,1).$$

Now we use the differential equations

$$\frac{\partial}{\partial t}\widehat{\eta}(\tau,t) = \tau\, R(\tau,t)(g(t))$$
$$\frac{\partial}{\partial t}R(\tau,t) = -\tau\, G(\eta(t))(R(\tau,t)(g(t)),-)\circ R(\tau,t)$$

Differentiating the first equation with respect to τ we get

$$\frac{\partial^2}{\partial t\,\partial\tau}\widehat{\eta}(\tau,t) = R(\tau,t)(g(t)) + \tau\,\frac{\partial}{\partial\tau}R(\tau,t)(g(t))$$

Inserting $\tau = 0$ we get

$$\frac{\partial}{\partial t}\left(\frac{\partial}{\partial\tau}\widehat{\eta}\right)(0,t) = R(0,t)(g(t)) = g(t),\ \text{because} R(0,t) = \mathrm{Id}\text{ for } t\in[0,1].$$

As $\widehat{\eta}(\tau,0) = 0$ we have $\frac{\partial}{\partial\tau}\widehat{\eta}(0,0) = 0$, and an easy integration gives the desired result.

Now we can come to the convergence theorem. At first we proof the convergence under a special hypothesis, and in the following theorem we proof that this hypothesis is satisfied. As we have seen that the Pilgerschritt transform is a $\mathscr{C}^\infty$-function, without loss of generality we may assume that the original function g is $\mathscr{C}^1$.

Theorem 5 *Suppose that there exists a real number δ, $0<\delta\le\varepsilon$, such that for all $\mathscr{C}^1$-functions $g\colon[0,1]\to\mathbb{R}^d$ with the property $\|g(0)\|+\|g'\|_\infty<\delta$ we have that $\widetilde{g}$ exists and, furthermore, $\|\widetilde{g}'\|_\infty \le \frac{1}{2}\|g'\|_\infty$. Then the following list of propositions are true for functions g that fulfill $\|g(0)\|+\|g'\|_\infty<\delta$:*

(a) The transform $\widetilde{g}$ exists (and the existence is not only an assumption).
(b) We have $\|\widetilde{g}(0)\|+\|\widetilde{g}'\|_\infty<\delta$.
(c) The sequence $(g(0),\widetilde{g}(0),\widetilde{\widetilde{g}}(0),\ldots)$ is a Cauchy sequence.
(d) The sequence $(g,\widetilde{g},\widetilde{\widetilde{g}},\ldots) = (g_0,g_1,g_2,\ldots)$ converges uniformly to a constant.

Proof These are easy computations. For sake of simplicity let us abbreviate $\omega = \|g'\|_\infty$.

(a) For $t\in[0,1]$ we have $g(t) = g(0)+\int\limits_0^t g'(\xi)\,\mathrm{d}\xi$ and therefore

$$\|g(t)\| \le \|g(0)\| + \int\limits_0^t \|g'(\xi)\|\,\mathrm{d}\xi \le \|g(0)\| + t\omega < \delta,\ \text{thus}$$

$\|g\|_\infty \le \delta \le \varepsilon$. Thus $\widetilde{g}$ exists by Theorem 4.

(b)

$$\|\widetilde{g}(0)\| + \|\widetilde{g}'\|_\infty \le \left\| \int_0^1 g(\xi)\,\mathrm{d}\xi \right\| + \frac{1}{2}\,\|g'\|_\infty \le \int_0^1 (\|g(0)\| + \xi\omega)\,\mathrm{d}\xi + \frac{1}{2}\omega =$$
$$= \|g(0)\| + \frac{1}{2}\omega + \frac{1}{2}\omega < \delta.$$

(c) By induction, we have $\|g_n'\|_\infty \le \frac{1}{2^n}\,\|g_0'\|_\infty$, and, therefore,

$$\|g_{n+1}(0) - g_n(0)\| \le \frac{1}{2}\,\|g_n'\|_\infty \le \frac{1}{2^{n+1}}\,\|g_0'\|_\infty.$$

Summing up shows that the sequence $(g_n(0))_{n\in\mathbb{N}}$ is a Cauchy sequence and therefore convergent.

(d) As the sequence $(g_n')_{n\in\mathbb{N}}$ uniformly tends to 0 and the sequence $(g_n(0))_{n\in\mathbb{N}}$ is convergent, we know that the sequence $(g_n)_{n\in\mathbb{N}}$ uniformly tends to a constant, and this constant is just the limit of the sequence $(g_n(0))_{n\in\mathbb{N}}$.

The only problem we now have to solve is to show the existence of such a number δ as stated above.

Theorem 6 *There exists a number δ, $0 < \delta \le \varepsilon$, such that for all $\mathscr{C}^1$-functions $g\colon [0,1] \to \mathbb{R}^d$ with the property $\|g(0)\| + \|g'\|_\infty < \delta$ the transform $\widetilde{g}$ exists and we have $\|\widetilde{g}'\|_\infty \le \frac{1}{2}\,\|g'\|_\infty$.*

Proof These are the 'hard' (but straightforward) computations:
Like in Theorem 2 let $r > 0$ be a radius such that G and G' are bounded on $B_r(0)$ by constants M_0 and M_1. Furthermore, we suppose that also G'' and G''' are bounded by constants M_2 and M_3. Let K be the ball around id in $L(\mathbb{R}^d, \mathbb{R}^d)$ with radius $\frac{1}{2}$, and let the number ε be chosen $0 < \varepsilon < r$ according to Theorem 2 such that all the paths and Pilgerschritt transforms exist.
Furthermore, suppose that we are given a number δ, $0 < \delta \le \varepsilon$, and our original function g fulfills the condition $\|g(0)\| + \|g'\|_\infty < \delta$. Now we will find bounds for $\|\widetilde{g}'\|_\infty$ and show that δ can be chosen such that $\|\widetilde{g}'\|_\infty \le \frac{1}{2}\,\|g'\|_\infty$.
According to the notation before, we have to 'compute' the pathes $\widehat{\eta}(\tau,t)$ for $t, \tau \in [0,1]$ and then the function $\widehat{g}(\tau,t) = S(\tau,t)^{-1}\left(\frac{\partial\widehat{\eta}}{\partial\tau}(\tau,t)\right)$ in order to get an estimate for $\frac{\partial\widehat{g}}{\partial\tau}(\tau,1)$.

$\widehat{\eta}$ is given by the differential equations

$$\frac{\partial}{\partial t}\widehat{\eta}(\tau,t) = \tau\, R(\tau,t)(g(t))$$
$$\frac{\partial}{\partial t}R(\tau,t) = -\tau\, G(\eta(t))(R(\tau,t)(g(t)), -) \circ R(\tau,t)$$

To make the problem 'easier' (it does not seem so, but it really is) we introduce a new variable for $\frac{\partial}{\partial t}\widehat{\eta}$ and get a system of three differential equations—for sake of simplicity let us denote the variables by $a_1 = \widehat{\eta}(\tau, t)$, $a_2 = \frac{\partial}{\partial t}\widehat{\eta}(\tau, t)$, $a_3 = R(\tau, t)$:

$$\begin{aligned}
&\frac{\partial a_1}{\partial t}(\tau, t) = a_2(\tau, t)\\
&\frac{\partial a_2}{\partial t}(\tau, t) = -\, G(a_1(\tau, t))(a_2(\tau, t), a_2(\tau, t)) + \tau\, a_3(\tau, t)\, g'(t)\\
&\frac{\partial a_3}{\partial t}(\tau, t) = -\, G(a_1(\tau, t))(a_2(\tau, t), -) \circ a_3(\tau, t)
\end{aligned}$$

to the initial value $(0, \tau g(0), \mathrm{id}) \in \mathbb{R}^d \times \mathbb{R}^d \times L(\mathbb{R}^d, \mathbb{R}^d)$.
For sake of simplicity we use the notation $A = (a_1, a_2, a_3)$ in order to treat these functions simultaneously.
Let us split the right hand side of the differential equation:
We denote $Z_1(t, \begin{pmatrix} x \\ y \\ z \end{pmatrix}) = \begin{pmatrix} y \\ -\, G(x)(y, y) \\ -\, G(x)(y, -) \circ z \end{pmatrix}$ and $Z_2(t, \begin{pmatrix} x \\ y \\ z \end{pmatrix}) = \begin{pmatrix} 0 \\ \tau\, z g'(t) \\ 0 \end{pmatrix}$.
Thus, shortly speaking, our differential equation is given by

$$A'(\tau, t) = Z_1(t, A(\tau, t)) + Z_2(t, A(\tau, t))$$

We also take into account the 'short' differential equation

$$B'(\tau, t) = Z_1(t, B(\tau, t))$$

with the solution $B = (b_1, b_2, b_3)$ to the same initial value $(0, \tau g(0), \mathrm{id})$.

By the perturbation formula of Gröbner and Alekseev (see [7] e.g.) we get

$$A(\tau, t) = B(\tau, t) + \int_0^t \Psi(t, \xi, A(\tau, \xi)) Z_2(\xi, A(\tau, \xi))\, \mathrm{d}\xi,$$

where Ψ is the fundamental matrix to the linear system of ODEs with matrix $\frac{\partial Z_1}{\partial(x, y, z)}$ at the time t with initial value $A(\tau, \xi)$ at time ξ. For further investigations we need this matrix:

$$P(x, y, z) = \frac{\partial Z_1}{\partial(x, y, z)} =$$

$$= \begin{pmatrix} 0 & \mathrm{Id} & 0 \\ -\, G'(x)(-)(y, y) & -\, G(x)(-, y) - G(x)(y, -) & 0 \\ -\, G'(x)(-)(y, \sim) \circ z & -\, G(x)(-, \sim) \circ z & -\, G(x)(y, -) \end{pmatrix}$$

Abbreviating the integral by $I = I(\tau, t) = (I_1, I_2, I_3)$ and using the notation $S(\tau, t)$ as before we get

$$\begin{aligned}
&\frac{\partial \widehat{g}}{\partial \tau}(\tau, t) = \frac{\partial}{\partial \tau}\left(S(\tau, t)^{-1}\,\frac{\partial \widehat{\eta}}{\partial \tau}(\tau, t)\right) = \\
&= S(\tau, t)^{-1}\left(\frac{\partial^2 \widehat{\eta}}{\partial \tau^2}(\tau, t) - \frac{\partial S}{\partial \tau}(\tau, t) \circ S(\tau, t)\frac{\partial \widehat{\eta}}{\partial \tau}(\tau, t)\right) = \\
&= S(\tau, t)^{-1}\left(\frac{\partial^2 \widehat{\eta}}{\partial \tau^2}(\tau, t) + G(\widehat{\eta}(\tau, t))\left(\frac{\partial \widehat{\eta}}{\partial \tau}(\tau, t), \frac{\partial \widehat{\eta}}{\partial \tau}(\tau, t)\right)\right) = \\
&= S(\tau, t)^{-1}\left(\frac{\partial^2 a_1}{\partial \tau} + G(a_1)\left(\frac{\partial a_1}{\partial \tau}, \frac{\partial a_1}{\partial \tau}\right)\right) = \\
&= S(\tau, t)^{-1}\left(\frac{\partial^2 b_1}{\partial \tau} + \frac{\partial^2 I_1}{\partial \tau} + G(a_1)\left(\frac{\partial b_1}{\partial \tau}, \frac{\partial b_1}{\partial \tau}\right) + G(a_1)\left(\frac{\partial I_1}{\partial \tau}, \frac{\partial b_1}{\partial \tau}\right) + \right.\\
&\left. + \; G(a_1)\left(\frac{\partial b_1}{\partial \tau}, \frac{\partial I_1}{\partial \tau}\right) + G(a_1)\left(\frac{\partial I_1}{\partial \tau}, \frac{\partial I_1}{\partial \tau}\right)\right) = \\
&= S(\tau, t)^{-1}\left(\frac{\partial^2 b_1}{\partial \tau} + G(b_1)\left(\frac{\partial b_1}{\partial \tau}, \frac{\partial b_1}{\partial \tau}\right) + \frac{\partial^2 I_1}{\partial \tau}\right.\\
&+ \left(G(a_1)\left(\frac{\partial b_1}{\partial \tau}, \frac{\partial b_1}{\partial \tau}\right) - G(b_1)\left(\frac{\partial b_1}{\partial \tau}, \frac{\partial b_1}{\partial \tau}\right)\right) + \\
&\left. + \; G(a_1)\left(\frac{\partial I_1}{\partial \tau}, \frac{\partial b_1}{\partial \tau}\right) + G(a_1)\left(\frac{\partial b_1}{\partial \tau}, \frac{\partial I_1}{\partial \tau}\right) + G(a_1)\left(\frac{\partial I_1}{\partial \tau}, \frac{\partial I_1}{\partial \tau}\right)\right)
\end{aligned}$$

Now the equation for B is the equation of a geodesic line. Thus we have $\frac{\partial^2 b_1}{\partial \tau} + G(b_1)\left(\frac{\partial b_1}{\partial \tau}, \frac{\partial b_1}{\partial \tau}\right) = 0$, as it was stated in Theorem 3. Our duty is to give estimates for the remaining terms.

First let us shrink the ball in which the paths should run: Let ϱ, $0 < \varrho \leq r$ be the radius for η (for a_1 and b_1), $\sigma > 0$ be a radius for the derivative (for a_2 and b_2), and μ a radius for a_3 and b_3, $0 < \mu \leq \frac{1}{2}$. Furthermore, let $\delta > 0$ at the moment be so small that for $\|g(0)\| + \|g'\|_\infty < \delta$ the solution A and B run in the set $W = B_\varrho(0) \times B_\sigma(0) \times B_\mu(\text{id})$ in $\mathbb{R}^d \times \mathbb{R}^d \times L(\mathbb{R}^d, \mathbb{R}^d)$. We use the sum norm on this space.

We go to find the estimates in several steps:

1. The matrix $\dfrac{\partial Z_1}{\partial(x, y, z)}$ is bound on the set W by
   $L_0(\varrho, \sigma, \mu) = \max\{M_1\sigma(1 + \sigma + \mu), 1 + M_0(1 + 2\sigma + \mu), M_0\sigma\}$
   Using the constant initial value $(0, \tau g(0), \text{id})$ as an approximate solution of B, we just take the difference

$$\left\| 0 - Z_1 \begin{pmatrix} 0 \\ \tau g(0) \\ \text{id} \end{pmatrix} \right\| = \left\| \begin{pmatrix} \tau g(0) \\ -G(0)(\tau g(0), \tau g(0)) \\ -G(0)(\tau g(0), -) \end{pmatrix} \right\| \leq$$

$$\leq \delta(1 + M_0\delta + M_0) = L_1(\varrho, \sigma, \mu, \delta).$$

From these two bounds by Gronwall's Lemma we get

$$\left\| \begin{pmatrix} b_1(\tau, t) \\ b_2(\tau, t) \\ b_3(\tau, t) \end{pmatrix} - \begin{pmatrix} 0 \\ \tau g(0) \\ \text{id} \end{pmatrix} \right\| \leq \frac{L_1(\varrho, \sigma, \mu, \delta)}{L_0(\varrho, \sigma, \mu)} \left(\mathrm{e}^{tL_0} - 1 \right).$$

2. $\dfrac{\partial B}{\partial \tau}$:
   For this function we have the differential equation
   $\dfrac{\partial}{\partial t}\dfrac{\partial B}{\partial \tau} = \dfrac{\partial Z_1}{\partial(x, y, z)}(B) \cdot \dfrac{\partial B}{\partial \tau}$ with the initial value
   $\dfrac{\partial B}{\partial \tau}(\tau, 0) = (0, g(0), 0)$. Thus

$$\left\| \frac{\partial B}{\partial \tau}(\tau, t) \right\| \leq \delta\, \mathrm{e}^{tL_0}$$

3. The difference $G(a_1) - G(b_1)$ is bounded ('mean value theorem') by $M_1 \, \|I_1\|$. As the matrix $\dfrac{\partial Z_1}{\partial(x, y, z)}$ is bounded by $L_0(\varrho, \sigma, \mu)$, the fundamental solution $\Psi(t, \xi)$ is bounded by $\mathrm{e}^{(t-\xi)L_0}$, and therefore we get

$$\|I_1\| \leq \|I\| \leq \frac{\mathrm{e}^{tL_0} - 1}{L_0} (1 + \mu)\|g'\|_\infty.$$

4. $\dfrac{\partial I_1}{\partial \tau}$: We have

$$I = I(\tau, t) = \int\limits_0^t \Psi(t, \xi, A(\tau, \xi)) \begin{pmatrix} 0 \\ \tau a_3(\tau, \xi) g'(\xi) \\ 0 \end{pmatrix} \mathrm{d}\xi.$$

Now $\dfrac{\partial \Psi(t, \xi, A(\tau, \xi))}{\partial \tau} = \displaystyle\int\limits_\xi^t \Psi(t, \zeta, A(\tau, \xi)) \, \frac{\partial P(A(\tau, \zeta))}{\partial \tau} \, \Psi(\zeta, \xi, A(\tau, \xi)) \, \mathrm{d}\zeta$

As P has L_0 as an upper bound, $\Psi(t,\xi,A(\tau,\xi))$ is bounded by $\mathrm{e}^{(t-\xi)L_0}$, and we have to find an estimate for $\dfrac{\partial P(A(\tau,\zeta))}{\partial\tau}$.

$$P(A(\tau,\zeta)) =$$
$$= \begin{pmatrix} 0 & \mathrm{Id} & 0 \\ -A'(a_1)(-)(a_2,a_2) & -A(a_1)(a_2,-)-A(a_1)(-,a_2) & 0 \\ -A'(a_1)(-)(a_2,\sim)\circ a_3 & -A(a_1)(-,\sim)\circ a_3 & -A(a_1)(a_2,\sim) \end{pmatrix}$$

Thus, by elementary differentiation (using linearity of many expressions) we get a linear operator acting on the vector $\dfrac{\partial A}{\partial\tau}$, and in norm this linear operator (matrix) can be majorized (by doing it componentwise) by the constant $L_2(\varrho,\sigma,\mu,\delta) = \max\{M_2\,\sigma\,(1+\sigma+\mu)+M_1(1+2\sigma+\mu), M_1(1+2\sigma+\mu)+2M_0, M_0+M_1\,\sigma\}$, and so we have

$$\left\| \frac{\partial P(A(\tau,\zeta))}{\partial\tau} \right\| \le L_2\cdot \left\| \frac{\partial A}{\partial\tau} \right\| .$$

In order to give an estimate for $\dfrac{\partial A}{\partial\tau}$ we go back to the defining differential equation

$$\begin{aligned}
\frac{\partial a_1}{\partial t}(\tau,t) &= a_2(\tau,t) \\
\frac{\partial a_2}{\partial t}(\tau,t) &= -G(a_1(\tau,t))(a_2(\tau,t),a_2(\tau,t)) + \tau\, a_3(\tau,t)\, g'(t) \\
\frac{\partial a_3}{\partial t}(\tau,t) &= -G(a_1(\tau,t))(a_2(\tau,t),-)\circ a_3(\tau,t)
\end{aligned}$$

to the initial value (0, $\tau g(0)$, id).
So we get

$$\begin{aligned}
\frac{\partial}{\partial t}\frac{\partial a_1}{\partial\tau} &= \frac{\partial a_2}{\partial\tau} \\
\frac{\partial}{\partial t}\frac{\partial a_2}{\partial\tau} &= -G'(a_1)\left(\frac{\partial a_1}{\partial\tau}\right)(a_2,a_2) - G(a_1)\left(\frac{\partial a_2}{\partial\tau},a_2\right) \\
&\quad -G(a_1)\left(a_2,\frac{\partial a_2}{\partial\tau}\right) + \tau\frac{\partial a_3}{\partial\tau}g' + a_3\,g' \\
\frac{\partial}{\partial t}\frac{\partial a_3}{\partial\tau} &= -G'(a_1)\left(\frac{\partial a_1}{\partial\tau}\right)(a_2,-)\circ a_3 - G(a_1)\left(\frac{\partial a_2}{\partial\tau},-\right)\circ a_3 \\
&\quad -G(a_1)(a_2,-)\circ\frac{\partial a_3}{\partial\tau}
\end{aligned}$$

a linear inhomogenous equation for $\frac{\partial A}{\partial \tau}$ to the initial value $(0, g(0), 0)$. Let us denote the linear operator in this equation by Φ and its fundamental solution by Ω. Thus by the formula of variation of constants we get

$$\frac{\partial A}{\partial \tau}(\tau, t) = \Omega(\tau, t) \cdot \begin{pmatrix} 0 \\ g(0) \\ 0 \end{pmatrix} + \int_0^t \Omega(\tau, \xi) \begin{pmatrix} 0 \\ a_3(\tau, \xi)\, g'(\xi) \\ 0 \end{pmatrix} \mathrm{d}\xi$$

which gives the estimate

$$\left\| \frac{\partial A}{\partial \tau} \right\| \leq \delta \cdot \left(\mathrm{e}^{tL_3} + \frac{\mathrm{e}^{tL_3} - 1}{L_3}(1+\mu) \right), \text{ where}$$
$$L_3(\varrho, \sigma, \mu, \delta) = \max\{M_1\sigma(1+\sigma+\mu), 1 + M_0(1+2\sigma+\mu), \delta + M_0\sigma\}.$$

Now we are able to estimate $\frac{\partial I}{\partial \tau}$:

$$I = \int_0^t \Psi(t, \xi, A(\tau, \xi)) \begin{pmatrix} 0 \\ \tau a_3(\tau, \xi) g'(\xi) \\ 0 \end{pmatrix} \mathrm{d}\xi.$$

Thus we have

$$\begin{aligned}\frac{\partial I}{\partial \tau} = &\int_0^t \int_\xi^t \Psi(t, \zeta, A(\tau, \xi)) \frac{\partial P(A(\tau, \zeta))}{\partial \tau} \Psi(\zeta, \xi, A(\tau, \xi)) \\ &\times \begin{pmatrix} 0 \\ \tau a_3(\tau, \xi) g'(\xi) \\ 0 \end{pmatrix} \mathrm{d}\zeta\, \mathrm{d}\xi + \\ &+ \int_0^t \Psi(t, \xi, A(\tau, \xi)) \begin{pmatrix} 0 \\ a_3(\tau, \xi) g'(\xi) \\ 0 \end{pmatrix} \mathrm{d}\xi + \\ &+ \int_0^t \Psi(t, \xi, A(\tau, \xi)) \begin{pmatrix} 0 \\ \tau \frac{\partial a_3}{\partial \tau}(\tau, \xi) g'(\xi) \\ 0 \end{pmatrix} \mathrm{d}\xi,\end{aligned}$$

and we can give an estimate

$$\left\| \frac{\partial I}{\partial \tau} \right\| \leq \int_0^t \int_\xi^t \mathrm{e}^{(t-\zeta)L_0} L_2 \cdot \left(\delta \cdot \left(\mathrm{e}^{tL_3} + \frac{\mathrm{e}^{tL_3}}{L_3}(1+\mu) \right) \right)$$
$$\times \mathrm{e}^{(\zeta-\xi)L_0}(1+\mu)\|g'\|_\infty \, \mathrm{d}\zeta \, \mathrm{d}\xi +$$
$$+ \int_0^t \mathrm{e}^{(t-\xi)L_0}(1+\mu)\|g'\|_\infty \, \mathrm{d}\xi + \int_0^t \mathrm{e}^{(t-\xi)L_0}\delta \cdot \left(\mathrm{e}^{tL_3} + \frac{\mathrm{e}^{tL_3}}{L_3}(1+\mu) \right) \|g'\|_\infty \, \mathrm{d}\xi.$$

5. $\dfrac{\partial^2 I_1}{\partial \tau^2}$:
Once more we use the expression

$$\frac{\partial I}{\partial \tau} = \int_0^t \int_\xi^t \Psi(t,\zeta,A(\tau,\xi)) \frac{\partial P(A(\tau,\zeta))}{\partial \tau}$$
$$\times \Psi(\zeta,\xi,A(\tau,\xi)) \begin{pmatrix} 0 \\ \tau a_3(\tau,\xi) g'(\xi) \\ 0 \end{pmatrix} \mathrm{d}\zeta \, \mathrm{d}\xi +$$
$$+ \int_0^t \Psi(t,\xi,A(\tau,\xi)) \begin{pmatrix} 0 \\ a_3(\tau,\xi) g'(\xi) \\ 0 \end{pmatrix} \mathrm{d}\xi +$$
$$+ \int_0^t \Psi(t,\xi,A(\tau,\xi)) \begin{pmatrix} 0 \\ \tau \dfrac{\partial a_3}{\partial \tau}(\tau,\xi) g'(\xi) \\ 0 \end{pmatrix} \mathrm{d}\xi,$$

and hence we get

$$\frac{\partial^2 I}{\partial \tau^2} =$$
$$= \int_0^t \int_\xi^t \int_\zeta^t \Psi(t,\omega,A) \frac{\partial P(A(\tau,\omega))}{\partial \tau} \Psi(\omega,\zeta,A) \frac{\partial P(A(\tau,\zeta))}{\partial \tau} \Psi(\zeta,\xi,A) \cdot$$
$$\cdot \begin{pmatrix} 0 \\ \tau a_3(\tau,\xi) g'(\xi) \\ 0 \end{pmatrix} \mathrm{d}\omega \, \mathrm{d}\zeta \, \mathrm{d}\xi +$$
$$+ \int_0^t \int_\xi^t \int_\xi^\zeta \Psi(t,\xi,A) \frac{\partial P(A(\tau,\zeta))}{\partial \tau} \Psi(\zeta,\omega,A) \frac{\partial P(A(\tau,\omega))}{\partial \tau} \Psi(\omega,\xi,A) \cdot$$
$$\cdot \begin{pmatrix} 0 \\ \tau a_3(\tau,\xi) g'(\xi) \\ 0 \end{pmatrix} \mathrm{d}\omega \, \mathrm{d}\zeta \, \mathrm{d}\xi +$$

$$+\int_0^t \int_\xi^t \Psi(t,\zeta,A)\frac{\partial^2 P(A(\tau,\zeta))}{\partial\tau^2}\Psi(\zeta,\xi,A)\begin{pmatrix}0\\ \tau a_3(\tau,\xi)g'(\xi)\\ 0\end{pmatrix}\mathrm{d}\zeta\,\mathrm{d}\xi+$$

$$+2\int_0^t \int_\xi^t \Psi(t,\zeta,A)\frac{\partial P(A(\tau,\zeta))}{\partial\tau}\Psi(\zeta,\xi,A)\begin{pmatrix}0\\ a_3(\tau,\xi)g'(\xi)\\ 0\end{pmatrix}\mathrm{d}\zeta\,\mathrm{d}\xi+$$

$$+2\int_0^t \int_\xi^t \Psi(t,\zeta,A)\frac{\partial P(A(\tau,\zeta))}{\partial\tau}\Psi(\zeta,\xi,A)\begin{pmatrix}0\\ \tau\frac{\partial a_3}{\partial\tau}(\tau,\xi)g'(\xi)\\ 0\end{pmatrix}\mathrm{d}\zeta\,\mathrm{d}\xi+$$

$$+2\int_0^t \Psi(t,\xi,A(\tau,\xi))\begin{pmatrix}0\\ \frac{\partial a_3}{\partial\tau}(\tau,\xi)g'(\xi)\\ 0\end{pmatrix}\mathrm{d}\xi+$$

$$+\int_0^t \Psi(t,\xi,A(\tau,\xi))\begin{pmatrix}0\\ \tau\frac{\partial^2 a_3}{\partial\tau^2}(\tau,\xi)g'(\xi)\\ 0\end{pmatrix}\mathrm{d}\xi$$

To give bounds for these seven terms almost all estimates are present—except bounds for $\frac{\partial^2 P(A)}{\partial\tau^2}$ and $\frac{\partial^2 a_3}{\partial\tau^2}$. We have to find now:

(a) $\frac{\partial^2 a_3}{\partial\tau^2}$:

For $\frac{\partial A}{\partial\tau}$ we had the differential equation

$$\frac{\partial}{\partial t}\frac{\partial A}{\partial\tau} = \Phi(A)\cdot\frac{\partial A}{\partial\tau} + \begin{pmatrix}0\\ a_3\,g'\\ 0\end{pmatrix}$$

From this we derive the equation for $\frac{\partial^2 A}{\partial\tau^2}$:

$$\frac{\partial}{\partial t}\frac{\partial^2 A}{\partial\tau^2} = \Phi(A)\cdot\frac{\partial^2 A}{\partial\tau^2} + \frac{\partial\Phi(A)}{\partial\tau}\cdot\frac{\partial A}{\partial\tau} + \begin{pmatrix}0\\ \frac{\partial a_3}{\partial\tau}\,g'\\ 0\end{pmatrix}.$$

This is an inhomogenous equation with matrix of coefficients Φ and initial condition $(0,0,0)$, the inhomogeneity is given by

$$\frac{\partial\Phi(A)}{\partial\tau}\cdot\frac{\partial A}{\partial\tau} + \begin{pmatrix}0\\ \frac{\partial a_3}{\partial\tau}\,g'\\ 0\end{pmatrix}$$

According to the initial condition we get $\frac{\partial^2 A}{\partial \tau^2} = \int_0^t \Omega(\tau, \xi) \cdot \text{inhomog.}\, d\xi$,

and so we have (the constant L_4 depends on the bilinear part of $\frac{\partial \Phi(A)}{\partial \tau}$)

$$\left\| \frac{\partial^2 A}{\partial \tau^2} \right\| \leq \int_0^t e^{\xi L_3}\, d\xi \left(L_4(\varrho, \sigma, \mu, \delta) \left\| \frac{\partial A}{\partial \tau} \right\|^2 + 2 \left\| \frac{\partial a_3}{\partial \tau} \right\| \|g'\|_\infty \right) \leq \delta^2 \cdot L_5$$

for some constant L_5.

(b) $\frac{\partial^2 P(A)}{\partial \tau^2}$:

The matrix $P(A)$ is just the matrix we used to estimate $\frac{\partial^2 A}{\partial \tau^2}$ except the term $a_3\, g'$ in the second line. Therefore we get the estimate

$$\left\| \frac{\partial^2 P(A)}{\partial \tau^2} \right\| \leq \delta^2 \cdot L_5.$$

We had seen that $\frac{\partial P(A)}{\partial \tau}$ is bounded by $\delta \cdot L_6$ for some constant L_6 depending on ϱ, σ and μ. Thus for such constants $L_7 \ldots L_{13}$ we have the following estimates

- $\delta^2\, \|g'\|_\infty \cdot L_7$ for the first and the second integral
- $\delta^2\, \|g'\|_\infty \cdot L_8$ for the third integral
- $2\,\delta\, \|g'\|_\infty \cdot L_9$ for the fourth integral
- $2\,\delta^2\, \|g'\|_\infty \cdot L_{10}$ for the fifth integral
- $2\,\delta\, \|g'\|_\infty \cdot L_{11}$ for the sixth integral
- $\delta^2\, \|g'\|_\infty \cdot L_{12}$ for the seventh integral

Summing up, we find a constant L_{13} such that $\frac{\partial^2 I}{\partial \tau^2}$ is bound by $\delta\, \|g'\|_\infty\, L_{13}$.

6. We also need an estimate for $S(\tau, t)$:
   $S(\tau, t)$ is a solution of the differential equation
   $\frac{\partial S(\tau, t)}{\partial \tau} = -\,G(a_1)(\frac{\partial a_1}{\partial \tau}, -) \circ S(\tau, t)$ to the initial value $S(0, t) = \text{id}$.
   As we had the estimate $\left\| \frac{\partial A}{\partial \tau} \right\| \leq \delta \cdot L_{14}$ for some constant, we get
   $\|S(\tau, t) - \text{id}\| \leq e^{\delta\, M_0\, L_{14}} - 1$.
   Thus a suitable small choice of δ guarantees that $S(\tau, t) \in B_\mu(\text{id})$ and also $S(\tau, t)^{-1} \in B_\mu(\text{id})$.
7. Combining all these results and using the bilinearity of $G(x)(-, \sim)$ we see that $\|\tilde{g}'\|_\infty$ is bounded by a constant times δ times $\|g'\|_\infty$. A suitable small choice of δ gives the desired result when constant times δ is smaller than $\frac{1}{2}$.

4 Summary

Theorem 2 shows that if we choose the starting path γ close to the initial point s (in the sense that $\|\mathrm{der}(\gamma)\|_\infty$ is small) then the Pilgerschritt transform $\widetilde{\gamma}$ exists. Furthermore, Theorem 6 shows the existence of a number $\delta > 0$, such that for all starting paths γ with $f = \mathrm{der}(\gamma)$ the Pilgerschritt sequence $\gamma, \widetilde{\gamma}, \widetilde{\widetilde{\gamma}}, \ldots$ converges to a geodesic line if the function $f : [0, 1] \to T_sM$ fulfills the condition $\|f(0)\| + \|f'\|_\infty < \delta$.

References

1. W. Förg-Rob, Zentrale Ähnlichkeit auf Gruppen und Mannigfaltigkeiten mit liearem Zusammenhang. Dissertation, University of Innsbruck, 1980
2. W. Förg-Rob, The Pilgerschritt (Liedl) transform. ESAIM Proc. **7**, 1–10 (2014)
3. S. Kobayashi, K. Nomizu, *Foundations of Differential Geometry*, vols. 1–2 (Interscience, Geneva, 1963/1969)
4. R. Liedl, Eine Methode zur stetigen Iteration der Gruppenoperation und zur Berechnung der Homotopieklasse eines Weges in einer abelschen reellen Liegruppe, Institut für Mathematik, Universität Innsbruck, 1977
5. R. Liedl, Über eine Methode zur Lösung der Translationsgleichung. Berichte der mathematisch-statistischen Sektion im Forschungszentrum Graz, Nr. 84, 1978
6. R. Liedl, *Non-commutative Calculus and Pilgerschritt Transformation* (Institut für Mathematik, Universität Innsbruck, Innsbruck, 1979)
7. G. Wanner, H. Reitberger, On the perturbation formulas of Gröbner and Alekseev. Buletinul Institutului Politehnic Din Iaşi, tomul XIX (XXIII), Fasc. 1–2 (1973)

On Some Mathematical Models Arising in Lubrication Theory

D. Goeleven and R. Oujja

1 Introduction

The lubrication theory is the source of many mathematical problems [1, 3, 6, 7, 12, 14, 17, 20, 21, 30]. The problems studied in these papers are related to the study of the free boundary problem of hydrodynamic lubrication from mechanical engineering. Both models describe the flow of a lubricant in some thin space and the mechanism of cavitation, i.e. the formation of air bubbles inside the lubricant.

The three-dimensional displacement of a Newtonian fluid in a laminar flow is governed by the Navier-Stokes equations. However, in the case of displacement in a privileged direction, i.e. when a dimension (the thickness in this case) is very small compared to other dimensions, Navier-Stokes equations are considerably simplified and reduce to two-dimensional Reynolds equation (see [7]).

Among hydrodynamic lubrication mechanisms we consider the shaft-bearing system formed by a cylindrical shaft which rotates with an angular velocity ω within a cylindrical fixed bush. The bush and the shaft are separated by a very thin layer of lubricant which role is to cushion the effects of friction or heating. Theses effects can cause damages on the contact surfaces. To ensure the correct functioning of the mechanism, a lubricant supply is kept constant through a circumferential groove (Fig. 1).

The narrow gap between the cylinders is occupied by a lubricant and the pressure p of this lubricant satisfy the Reynolds equation. Let r, θ, x be the cylindrical coordinates with origin in the bottom of the cylinder and suppose that θ is measured from the line of maximum clearance of the centers. As a consequence of the thin-film hypothesis, the pressure depends on the angular coordinate θ and the height of

D. Goeleven (✉) · R. Oujja
University of La Réunion, PIMENT EA4518, Saint-Denis Messag, La Réunion, France
e-mail: daniel.goeleven@univ-reunion.fr; rachid.oujja@univ-reunion.fr

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization
and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_12

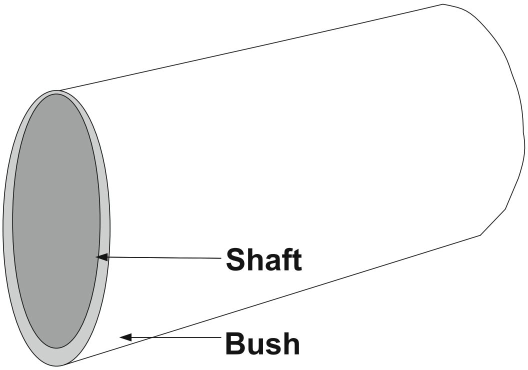


Fig. 1 Schematic representation of the bearing

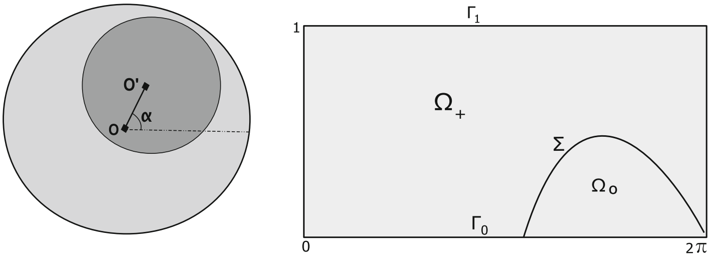


Fig. 2 A cross of the bearing and the domain Ω

the cylinders x and does not depend on the normal coordinate r. Therefore, the mathematical model can be formulated on the set $\Omega = [0, 2\pi] \times [0, 1]$ which represents the lateral area of the shaft.

Let r_s be the radius of the shaft and r_b the radius of the bush, e the distance between the axes of the cylinders and $\eta = \frac{e}{r_s - r_b}$, $0 \leq \eta < 1$ the eccentricity ratio of the bearing. The thickness of the thin fluid film is represented by the function

$$h(\theta) = (r_s - r_b)(1 - \eta\cos(\theta - \alpha)), \tag{1}$$

where α is the angle of vector $\overrightarrow{OO'}$ ($\overrightarrow{OO'} = (e\cos\alpha, e\sin\alpha)$) (Fig. 2).

When considering the non-coincidence of the axes of the cylinders (i.e. $\eta > 0$), overpressure areas and low pressure areas are produced and this results in the appearance of bubbles phenomenon known as cavitation. Mathematical models of the phenomenon of cavitation consider that the lubricant pressure satisfies the following Reynolds equation in the lubricated area Ω_+:

$$-\frac{1}{r_b}\frac{\partial}{\partial\theta}\Big(\frac{h^3}{12\mu r_b}\frac{\partial p}{\partial\theta}\Big) - \frac{\partial}{\partial x}\Big(\frac{h^3}{12\mu}\frac{\partial p}{\partial x}\Big) = -\frac{r_b\omega}{2}\frac{dh}{d\theta}, \tag{2}$$

where μ is a constant viscosity coefficient, while in the remaining area Ω_0 (cavitated zone) the pressure vanishes. Equation (2) is based on heuristic reasoning [30] and mathematics [7]. By taking $r_b = 1$, $h := \frac{h}{r_s - rb}$ and $p := \frac{(r_s - r_b)^2 p}{6\mu\omega}$ we can write the Reynolds equation in the dimensionless form

$$\frac{\partial}{\partial\theta}\big(h^3\frac{\partial p}{\partial\theta}\big) + \frac{\partial}{\partial x}\big(h^3\frac{\partial p}{\partial x}\big) = \frac{dh}{d\theta}, \tag{3}$$

where $h(\theta) = 1 - \eta\cos(\theta - \alpha)$.

2 Reynolds Free Boundary Problem

The Reynolds free boundary problem consists to find the pressure $p(X)$ $(X = (\theta, x))$ and the regions Ω_+ and Ω_0 such that

$$\frac{\partial}{\partial\theta}\big(h^3\frac{\partial p}{\partial\theta}\big) + \frac{\partial}{\partial x}\big(h^3\frac{\partial p}{\partial x}\big) = \frac{dh}{d\theta}, \quad p > 0 \quad \text{in } \Omega_+, \tag{4}$$

$$p = 0 \quad \text{in } \Omega_0, \tag{5}$$

$$p = \frac{\partial p}{\partial n} = 0 \quad \text{on } \Sigma = \Omega_0 \cap \Omega_+, \tag{6}$$

$$p(\theta, 0) = p(\theta, 1) = 0, \qquad 0 \leq \theta \leq 2\pi \tag{7}$$

and

$$p(2\pi, x)) = p(0, x), \qquad 0 \leq x \leq 1. \tag{8}$$

Note that the free boundary Σ is an additional unknown of the problem. Equations (4)–(6) may be summarized as

$$p \geq 0 \quad \text{and} \quad p(div(h^3\nabla p) - \frac{dh}{d\theta}) = 0 \quad \text{in} \quad \Omega.$$

We set

$$V = \{\varphi \in H^1(\Omega) : \varphi(., 0) = \varphi(., 1) = 0, \varphi \text{ is } 2\pi\text{-periodic}\}$$

and

$$K = \{\varphi \in V, \varphi \geq 0\}.$$

Problem (4)–(8) can be formulated as the following variational inequality:

Find $p \in K$ such that

$$\int_{\Omega} h^3 \nabla p \cdot \nabla(\varphi - p)\, dX \geq \int_{\Omega} h \frac{\partial}{\partial\theta}(\varphi - p)\, dX, \quad \forall \varphi \in K. \tag{9}$$

Existence and uniqueness results of (9) are now well known. There exists a unique $p(\theta, x)$ satisfying (9) and a no empty cavitated area $\Omega_0 \neq \emptyset$ (see [22]).

Numerical approach of inequality (9) can be placed in the frame of some approximation methods for variational inequalities based on classical results for monotone operators given in [11, 29]. For this purpose we give the following formulation of (9):

Find $p \in V$ such that

$$\int_{\Omega} h^3 \nabla p . \nabla(\varphi - p)\, dX + I_K(\varphi) - I_K(p) \geq \int_{\Omega} h \frac{\partial}{\partial\theta}(\varphi - p)\, dX, \forall \varphi \in V, \tag{10}$$

where I_K is the indicatrix function of the non empty closed and convex set K.

From the definition of the sub-differential ∂I_K we have

$$\beta \in \partial I_K(p) \iff I_K(\varphi) - I_K(p) \geq < \beta, \varphi - p >, \ \forall \varphi \in V.$$

It follows then from (10) that

$$\beta = Div(h^3 \nabla p) - \frac{\partial h}{\partial\theta} \in \partial I_K(p) \tag{11}$$

and we obtain the following equivalent formulation:

Find $p \in V$ such that

$$\int_{\Omega} h^3 \nabla p . \nabla \varphi\, dX + \int_{\Omega} \beta \varphi\, dX = -\int_{\Omega} \frac{dh}{d\theta} \varphi\, dX, \forall \varphi \in V \tag{12}$$

and

$$\beta \in \partial I_K(p). \tag{13}$$

Following [10] we introduce the multiplier $\gamma = \beta - \omega p$ where $\omega > 0$ is a positive parameter and we get the formulation:

Find $p \in V$ such that

$$\int_{\Omega} h^3 \nabla p . \nabla \varphi\, dX + \omega \int_{\Omega} p \varphi\, dX = -\int_{\Omega} \gamma \varphi\, dX - \int_{\Omega} \frac{\partial h}{\partial\theta} \varphi\, dX, \forall \varphi \in V \tag{14}$$

and

$$\gamma \in \partial I_K(p) - \omega p. \tag{15}$$

Let $\lambda > 0$, suppose that $\lambda\omega < 1$ and set $T = \partial I_K - \omega I$. We have

$$I + \lambda T = (1 - \lambda\omega)I + \lambda\partial I_K. \tag{16}$$

It can be proved that for all $f \in V$ there exists a unique $y \in V$ such that

$$f \in (I + \lambda T)(y).$$

The single-valued map

$$J_\lambda^T = (I + \lambda T)^{-1}$$

is the resolvent operator of T and the map

$$T_\lambda = \frac{I - J_\lambda^T}{\lambda}$$

is the Moreau-Yosida approximation of T. The map T_λ is single-valued and $\frac{1}{\lambda}$-Lipschitz continuous. Moreover it satisfies the following property:

Lemma 1 *For all y and u in V, we have the equivalence property:*

$$u \in T(y) \iff u = T_\lambda(y + \lambda u). \tag{17}$$

Taking into account (17) in (14)–(15) we get the final formulation:
Find $p \in V$ such that

$$\int_\Omega h^3 \nabla p . \nabla\varphi \, dX + \omega \int_\Omega p\varphi \, dX = - \int_\Omega \gamma\varphi \, dX - \int_\Omega \frac{dh}{d\theta}\varphi \, dX, \forall\varphi \in V \tag{18}$$

and

$$\gamma = T_\lambda(p + \lambda\gamma). \tag{19}$$

2.1 Iterative Algorithm

To compute the solution (p, γ) of (18)–(19), we apply the following iterative method:

(0) Start with some arbitrary value of the multiplier γ_0.
(1) For γ_j known, compute p_j solution to

$$\int_\Omega h^3 \nabla p_j.\nabla\varphi\, dX + \omega \int_\Omega p_j \varphi\, dX = -\int_\Omega \gamma_j \varphi\, dX - \int_\Omega \frac{dh}{d\theta}\varphi\, dX, \forall\varphi \in V. \tag{20}$$

(2) Update multiplier γ_j as

$$\gamma_{j+1} = T_\lambda(p_j + \lambda\gamma_j). \tag{21}$$

(3) Go to (1) until stop criterion is reached.

Theorem 1 *For* $\lambda \geq \dfrac{1}{2\omega}$ *we have:*

$$\lim_{j\to\infty} \|p_j - p\| = 0.$$

Proof The mapping T_λ is $\dfrac{1}{\lambda}$-Lipschitz and thus

$$\|\gamma - \gamma_{j+1}\|^2 = \|T_\lambda(p + \lambda\gamma) - T_\lambda(p_j + \lambda\gamma_j)\|^2 \leq \frac{1}{\lambda^2}\|(p + \lambda\gamma) - (p_j + \lambda\gamma_j)\|^2$$

$$= \frac{1}{\lambda^2}\|(p - p_j) + \lambda(\gamma - \gamma_j)\|^2$$

$$= \frac{1}{\lambda^2}\|p - p_j\|^2 + \frac{2}{\lambda}(p - p_j, \gamma - \gamma_j) + \|\gamma - \gamma_j\|^2.$$

Therefore

$$\|\gamma - \gamma_j\|^2 - \|\gamma - \gamma_{j+1}\|^2 \geq -\frac{1}{\lambda^2}\|p - p_j\|^2 - \frac{2}{\lambda}(p - p_j, \gamma - \gamma_j). \tag{22}$$

From (14) and (20) we have

$$\int_\Omega h^3 \nabla(p - p_j).\nabla\varphi\, dX + \omega \int_\Omega (p - p_j)\varphi\, dX = -\int_\Omega (\gamma - \gamma_j)\varphi\, dX, \quad \forall\varphi \in V.$$

And by taking $\varphi = p - p_j$ we obtain

$$\omega\|p - p_j\|^2 \leq \int_\Omega h^3|\nabla(p - p_j)|^2\, dX + \omega \int_\Omega (p - p_j)^2\, dX$$

$$= -\int_\Omega (\gamma - \gamma_j)(p - p_j)\, dX.$$

Now by substituting this inequality in (22) we obtain

$$\|\gamma - \gamma_j\|^2 - \|\gamma - \gamma_{j+1}\|^2 \geq -\frac{1}{\lambda^2}\|p - p_j\|^2 + \frac{2\omega}{\lambda}\|p - p_j\|^2$$

$$= \frac{1}{\lambda}(2\omega - \frac{1}{\lambda})\|p - p_j\|^2.$$

If $\lambda \geq \dfrac{1}{2\omega}$ then we get

$$\|\gamma - \gamma_j\|^2 - \|\gamma - \gamma_{j+1}\|^2 \geq 0.$$

The sequence $(\|\gamma - \gamma_j\|^2)_{j\geq 0}$ is then decreasing and positive. Therefore

$$\lim_{j\to\infty} \|\gamma_j - \gamma\|^2 = 0$$

and finally

$$\lim_{j\to\infty} \|p_j - p\|^2 = 0.$$

□

Remark 1 Note that for $\frac{1}{2} \leq \lambda\omega < 1$, the condition in (16) and the hypothesis of Theorem 1 are satisfied.

2.2 An Adaptive Finite Element Method

Our purpose in this section is to apply a P1-Galerkin finite element method so as to discretize Eq. (20) (see [13]). Let be $\mathscr{T}$ a regular triangulation of the domain into triangles and h an abstract discretization parameter. The discretization method consists in the construction of a finite-dimensional space

$$V_h := \{\psi \in \mathscr{C}(\Omega) \cap V : T \in \mathscr{T}, \quad \psi|_T \text{ is affine}\}.$$

The discrete solution $p_h \in V_h$ is defined as

$$\int_\Omega h^3 \nabla p_h.\nabla\varphi\, dX + \omega \int_\Omega p_h\varphi\, dX = -\int_\Omega \gamma_j\varphi\, dX - \int_\Omega \frac{dh}{d\theta}\varphi\, dX, \forall\varphi \in V_h. \tag{23}$$

As mentioned above, high pressure variations occur in the bearing and cause the appearance of cavitated areas. The accuracy of discrete solution p_h depends then on the triangulation $\mathscr{T}$ in the sense that singularities and high variations of p_h have to be resolved by the triangulation. For this purpose we apply an adaptive method where triangulation $\mathscr{T}$ is improved automatically by use of a mesh-refinement where high variations occur. Moreover, the solution p_h can become smoother in some area of domain when iteration proceeds and in this case, certain elements from $\mathscr{T}$ are removed and mesh coarsening is done.

More precisely to get a refined triangulation from the current triangulation, we first solve the equation to get the solution on the current triangulation. The error is estimated using the solution and used to mark a set of triangles that are to be refined or coarsened. Triangles are refined or coarsened in such a way to keep regularity of the triangulations. This method is based on the following error estimator introduced by Babuska and Rheinboldt [4] and used in most works on convergence and optimality.

Theorem 2 *Given a triangulation $\mathscr{T}$ and let be p_h the solution of the discrete problem. There exists a constant $C > 0$ and an error estimator $\eta_h > 0$ depending on p_h such that*

$$\|p_j - p_h\|_{H^1(\Omega)} \leq C\eta_h.$$

Proof Let $\varphi \in V$ be given. We have with arbitrary $\varphi_h \in V_h$

$$\int_\Omega h^3 \nabla(p_j - p_h).\nabla\varphi \, dX + \omega \int_\Omega (p_j - p_h)\varphi \, dX =$$

$$\int_\Omega h^3 \nabla(p_j - p_h).\nabla(\varphi - \varphi_h) \, dX + \omega \int_\Omega (p_j - p_h)(\varphi - \varphi_h) \, dX$$

$$= -\int_\Omega \gamma_j(\varphi - \varphi_h) \, dX - \int_\Omega \frac{dh}{d\theta}(\varphi - \varphi_h) \, dX$$

$$- \int_\Omega h^3 \nabla p_h.\nabla(\varphi - \varphi_h) \, dX - \omega \int_\Omega p_h(\varphi - \varphi_h) \, dX$$

$$= \sum_{T\in\mathscr{T}} \Big[-\int_T \gamma_j(\varphi - \varphi_h) \, dX - \int_T \frac{dh}{d\theta}(\varphi - \varphi_h) \, dX + \int_T Div(h^3\nabla p_h)(\varphi - \varphi_h) \, dX$$

$$-\omega \int_T p_h(\varphi - \varphi_h) \, dX + \frac{1}{2}\int_{\partial T\setminus\partial\Omega} [\frac{\partial(h^3 p_h)}{\partial n}](\varphi - \varphi_h) dS \Big]$$

$$= \sum_{T\in\mathscr{T}} \Big[\int_T \Big(Div(h^3\nabla p_h) - \omega p_h - \gamma_j - \frac{dh}{d\theta}\Big)(\varphi - \varphi_h) \, dX$$

$$+ \frac{1}{2}\int_{\partial T\setminus\partial\Omega} \Big[\frac{\partial(h^3 p_h)}{\partial n}\Big](\varphi - \varphi_h) dS \Big].$$

We obtain two kinds of residuals. Indeed $Div(h^3\nabla p_h) - \omega p_h - \gamma_j - \frac{dh}{d\theta}$ is a pointwise residual and $\left[\frac{\partial(h^3 p_h)}{\partial n}\right]$ is a measure of regularity of the discrete solution. By applying the Cauchy Schwarz inequality we get

$$\int_\Omega h^3\nabla(p_j - p_h).\nabla\varphi \, dX + \omega\int_\Omega (p_j - p_h)\varphi \, dX \leq$$

$$\sum_{T\in\mathscr{T}} \Big[\|Div(h^3\nabla p_h) - \omega p_h - \gamma_j - \frac{\partial h}{\partial\theta}\|_T \|\varphi - \varphi_h\|_T + \|\frac{1}{2}\Big[\frac{\partial(h^3 p_h)}{\partial n}\Big]\|_{\partial T^*}\|\varphi - \varphi_h\|_{\partial T^*}\Big]$$

where $\partial T^* = \partial T \setminus \partial\Omega$. We now chose $\varphi_h = C_h\varphi$ with Clement interpolation operator $C_h : V \to V_h$ (see [15]) which verifies the interpolation estimate

$$\|\varphi - C_h\varphi\|_T + d_T^{1/2}\|\varphi - C_h\varphi\|_{\partial T} \leq C d_T\|\nabla\varphi\|_{\Omega_T},$$

with Ω_T denoting the set of neighboring elements of T and d_T its diameter.

It follows that

$$\int_\Omega h^3\nabla(p_j - p_h).\nabla\varphi dX + \omega\int_\Omega (p_j - p_h)\varphi dX$$

$$\leq \sum_{T\in\mathscr{T}} C\Big[d_T\|Div(h^3\nabla p_h) - \omega p_h - \gamma_j - \frac{\partial h}{\partial\theta}\|_T + \frac{d_T^{1/2}}{2}\|\Big[\frac{\partial(h^3 p_h)}{\partial n}\Big]\|_{\partial T^*}\Big]\|\nabla\varphi\|_{\Omega_T}$$

$$\leq \sum_{T\in\mathscr{T}} 2C\Big[d_T^2\|Div(h^3\nabla p_h) - \omega p_h - \gamma_j - \frac{\partial h}{\partial\theta}\|_T^2 + \frac{d_T}{4}\|\Big[\frac{\partial(h^3 p_h)}{\partial n}\Big]\|_{\partial T^*}^2\Big]^{1/2}\|\nabla\varphi\|_{\Omega_T}$$

$$\leq C\Big(\sum_{T\in\mathscr{T}}\eta_T^2\Big)^{1/2}\Big(\sum_{T\in\mathscr{T}}\|\nabla\varphi\|_{\Omega_T}^2\Big)^{1/2} \leq C\eta_h\|\nabla\varphi\|,$$

where $\eta_T^2 = \Big[4d_T^2\|Div(h^3\nabla p_h) - \omega p_h - \gamma_j - \frac{\partial h}{\partial\theta}\|_T^2 + d_T\|\Big[\frac{\partial(h^3 p_h)}{\partial n}\Big]\|_{\partial T^*}^2\Big]$. We have used in the last step that the number of neighbors of any triangle T is bounded due to the uniform shape-regularity of the meshes. Taking $\varphi = p_j - p_h$ we obtain the global upper bounded:

$$\min((1-\eta)^3, \omega)\|p_j - p_h\|_{H^1(\Omega)} \leq C\eta_h,$$

with the residual-based error estimator $\eta_h = (\sum_{T\in\mathscr{T}}\eta_T^2)^{1/2}$ □

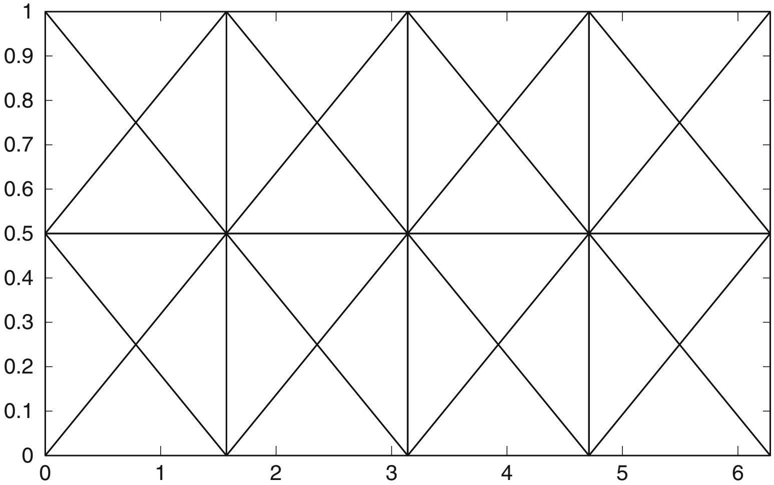


Fig. 3 Initial triangulation $\mathscr{T}_0$

2.3 Numerical Results

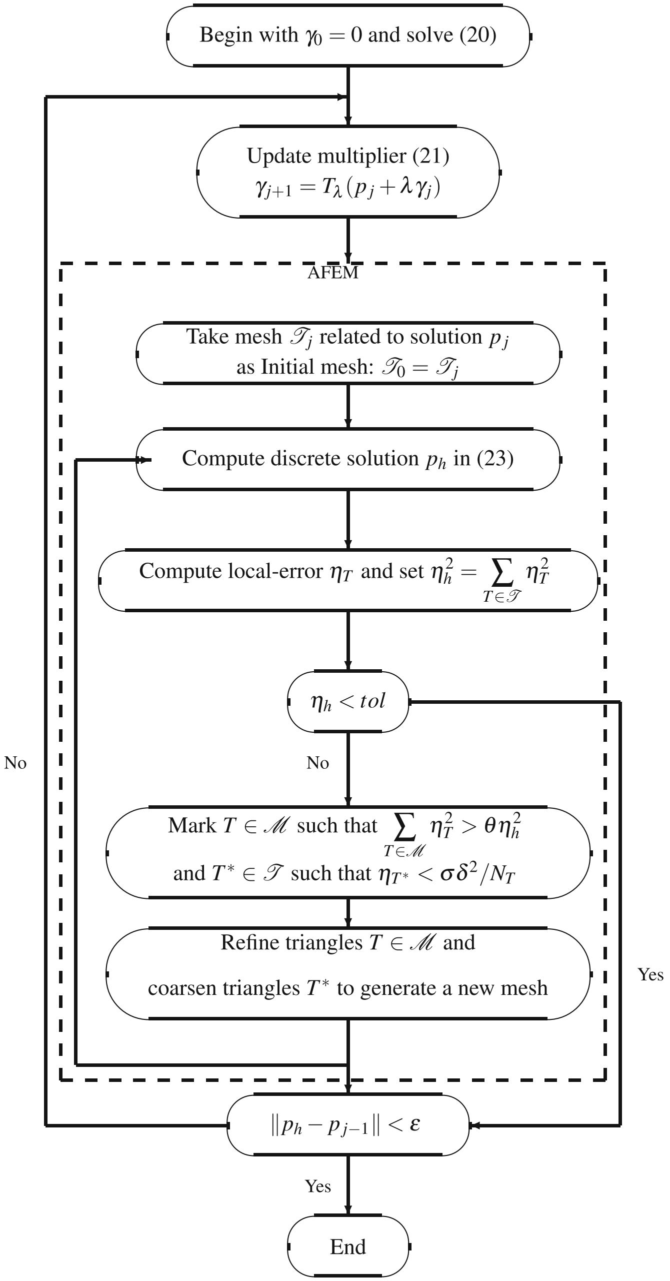
For our numerical simulations, we take $\omega = 0.1$ and $\lambda = 5$. Then both the approximation condition (16) and Theorem 1 convergence hypothesis are satisfied. We start algorithm (20)–(21) with $\gamma_0 = 0$ and take triangulation given in Fig. 3 as initial mesh for the adaptive finite element method in the first step. At each step $j + 1$ we take the final triangulation $\mathscr{T}_j$ obtained in the step j as initial triangulation for the adaptive method.

Let $\mathscr{T}$ be a given triangulation. If the error estimator $\eta_h < \delta$, where δ is a fixed tolerance then the corresponding solution p_h is accepted and the adaptive method stopped. Otherwise, we apply the Dorfler criterion [16] to mark elements $T \in \mathscr{T}$ for refinement. This criterion seeks to determine the minimal set $\mathscr{M} \subset \mathscr{T}$ such that

$$\theta \left(\sum_{T \in \mathscr{T}} \eta_T^2 \right) \leq \sum_{T \in \mathscr{M}} \eta_T^2,$$

for some parameter $\theta \in]0, 1[$.

For coarsening we mark elements $T \in \mathscr{T}$ such that $\eta_T^2 < \sigma \dfrac{\delta^2}{N}$, where $\sigma \in]0, 1[$, δ is a fixed tolerance and N_T the number of nodes of triangulation $\mathscr{T}$. We take here $\sigma = \delta = 1$. A stop criterion ϵ is fixed also for the algorithm (20)–(21). We summarize the method in the diagram below. We give in Figs. 4 and 5 the numerical simulations for $\alpha = \pi$ and $\alpha = 0$ respectively.


Begin with $\gamma_0 = 0$ and solve (20)
Update multiplier (21)
$\gamma_{j+1} = T_\lambda(p_j + \lambda\gamma_j)$
AFEM
Take mesh $\mathscr{T}_j$ related to solution p_j
as Initial mesh: $\mathscr{T}_0 = \mathscr{T}_j$
Compute discrete solution p_h in (23)
Compute local-error η_T and set $\eta_h^2 = \sum_{T\in\mathscr{T}} \eta_T^2$
$\eta_h < tol$
No
Mark $T \in \mathscr{M}$ such that $\sum_{T\in\mathscr{M}} \eta_T^2 > \theta\eta_h^2$
and $T^* \in \mathscr{T}$ such that $\eta_{T^*} < \sigma\delta^2/N_T$
Refine triangles $T \in \mathscr{M}$ and
coarsen triangles T^* to generate a new mesh
No
Yes
$\|p_h - p_{j-1}\| < \varepsilon$
Yes
End

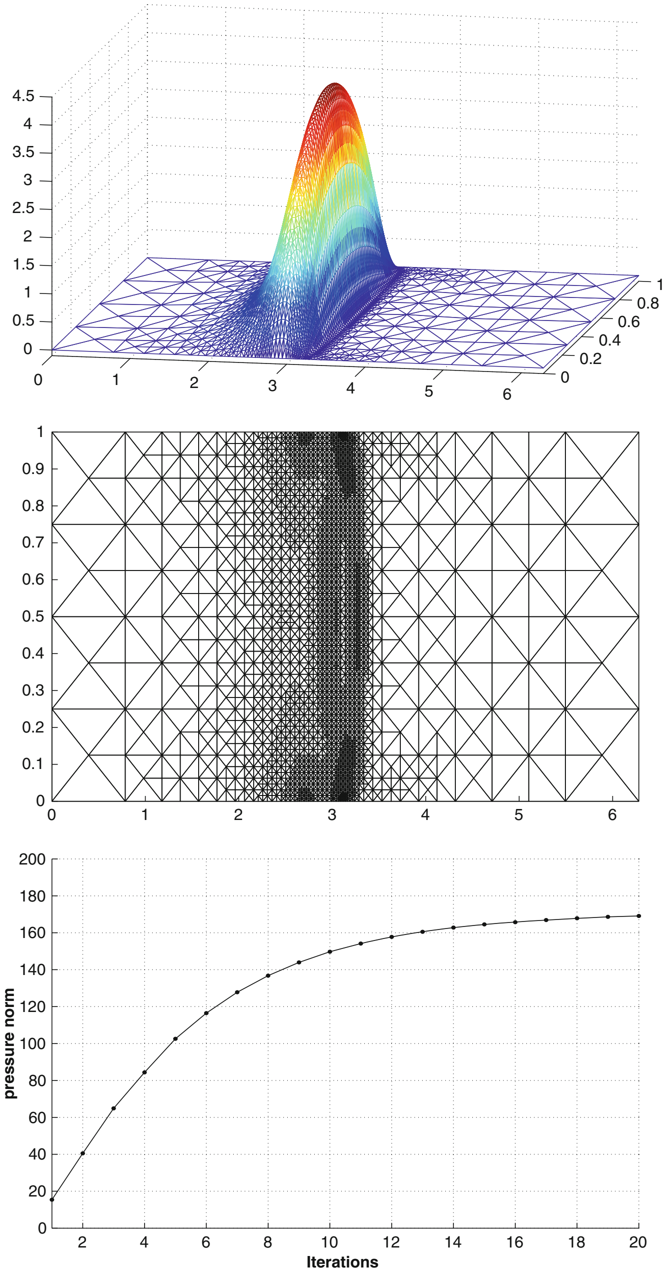


Fig. 4 Pressure p_h, the final mesh $\mathscr{T}_h$ and the pressure norm $\|p_j\|$ evolution ($1 \leq j \leq 20$) for $\alpha = \pi$

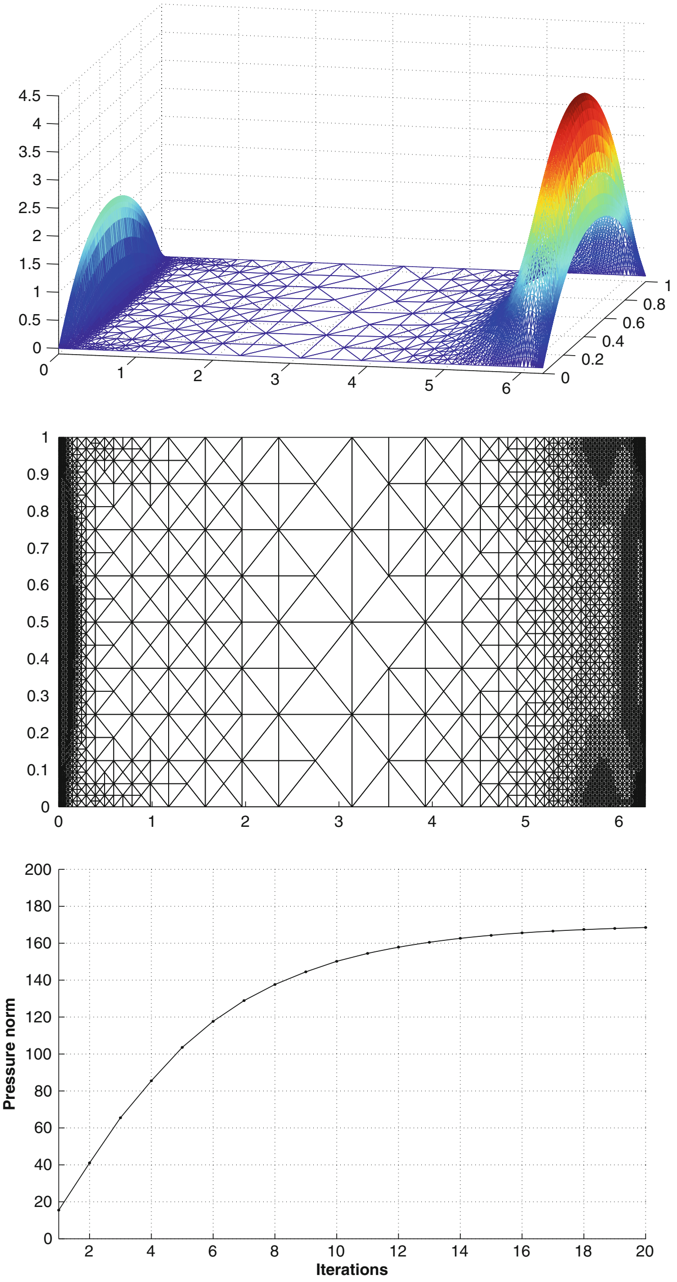


Fig. 5 Pressure p_h, the final mesh $\mathscr{T}_h$ and the pressure norm $\|p_j\|$ evolution $\|p_j\|$ $(1 \leq j \leq 20)$ for $\alpha = 0$

3 Elrod-Adams Free Boundary Problem

The difference between the models introduced in modeling the phenomenon of cavitation resides in the condition imposed on the interface that separates the cavity from the lubricated zone [6, 17, 32]. The Reynolds model is based on the continuity of the flow across the free boundary.

Another model supposed more realistic in the majority of situations is the Elrod-Adams one [1, 5, 18, 33]. This model considers that the cavitation zone $\Omega^0 = \{(x, y) \in \Omega | \quad p(x, y) = 0\}$ is a mixture of air and fluid. It introduces a new variable $\gamma(x, y)$ which represents the concentration of lubricant existing in a neighborhood of (x, y). That is a saturation function that takes values between 0 and 1 in cavitation and is equal to 1 in the active part $\Omega^+ = \{(x, y) \in \Omega | \quad p(x, y) > 0\}$. We denote by $\Gamma_0 = (0, 2\pi) \times \{0\}$ and $\Gamma_1 = (0, 2\pi) \times \{1\}$ the upper and lower parts of the boundary of the domain. The problem is to find (p, γ) defined on the domain Ω, with $p(x, y)$ is a $2\pi x$-periodic function, $p \geq 0$ and $0 \leq \gamma \leq 1$ such that

$$\frac{\partial}{\partial x}(h^3\frac{\partial p}{\partial x}) + \frac{\partial}{\partial y}(h^3\frac{\partial p}{\partial y}) = \frac{dh}{dx}, \quad p > 0, \quad \gamma = 1 \text{ sur } \Omega^+, \tag{24}$$

$$\frac{\partial(h\gamma)}{\partial x} = 0, \quad p = 0, \quad 0 \leq \gamma \leq 1 \text{ on } \Omega_0, \tag{25}$$

$$h^3\frac{\partial p}{\partial n} = (1-\gamma)h\cos(n, x), \quad p = 0 \text{ on } \Sigma = \partial\Omega^+ \cap \partial\Omega^0 \tag{26}$$

and

$$p = 0 \quad on \quad \Gamma_0, \quad p = p_a > 0 \quad on \quad \Gamma_. \tag{27}$$

Where n is the normal outside vector on Σ. The free boundary Σ is another unknown problem. For more details on how to obtain theses equations see [17, 31].

Let be the following subsets:

$$V_0 = \{\xi \in H^1(\Omega), \quad \xi|_{\Gamma_0} = 0, \quad \xi|_{\Gamma_1} = 0, \quad et \quad \xi \quad 2\pi x\text{-}periodic\} \tag{28}$$

and

$$V_a = \{\xi \in H^1(\Omega), \quad \phi|_{\Gamma_0} = 0, \quad \phi|_{\Gamma_1} = p_a, \quad et \quad \phi \quad 2\pi x\text{-}periodic\}. \tag{29}$$

In [1] the following weak formulation is established:

Problem $\mathscr{Q}$

Find $p \in V_a$ and $\gamma \in L^\infty(\Omega)$ such that

$$\int_\Omega h^3 \nabla p \nabla \xi = \int_\Omega h\gamma\xi_x, \qquad \forall \xi \in V_0, \tag{30}$$

$$H(p) \leq \gamma \leq 1 \text{ a.e. in } \Omega \tag{31}$$

and

$$p \geq 0. \tag{32}$$

where H is the Heaviside graph:

$$H(x) = \begin{cases} 1 \quad if \quad x > 0 \\ (0,1) \quad if \quad x = 0 \\ 0 \quad if \quad x < 0. \end{cases} \tag{33}$$

The existence and uniqueness of a weak solution is proved in [3]. Note that some qualitative properties of the free boundary have also been obtained in [2].

We are here interested in the numerical approximation of the problem $\mathscr{Q}$. For this purpose, we apply a numerical method which approaches the pressure p and the free boundary through a process of successive approximations. At each stage of this process, a linear equation is solved and the non-homogeneous term of this linear equation "seeks" the free boundary of the problem.

3.1 A One-Dimensional Problem

This formulation corresponds to a long bearing. In such systems pressure variations with respect to the vertical variable can be neglected and the free boundary problem can be formulated on the interval $I = [0, 2\pi]$ (see [26]).

Problem $\mathscr{M}$ Find $p \in H^1(I)$ and $\gamma \in L^\infty(I)$, $2\pi-$ periodic such that

$$\int_0^{2\pi} h^3 p'\xi' = \int_0^{2\pi} h\gamma\xi', \qquad \forall \xi \in H^1(I), \tag{34}$$

$$p \geq 0, \quad H(p) \leq \gamma \leq 1 \text{ a.e. in } I \tag{35}$$

and

$$\int_0^{2\pi} p = a \ (a > 0). \tag{36}$$

Equation (36) is a load condition introduced to have uniqueness of the solution p. The numerical method that we apply is based on the property (25) of the second member $h\gamma$ in the Eq. (34). Let p_0 be the solution of he linear problem:

$$\frac{d}{dx}(h^3\frac{dp_0}{dx}) = \frac{dh}{dx} \text{ in } \Omega,$$

$$p_0(0) = p_0(2\pi)$$

and

$$\int_0^{2\pi} p_0 = a \quad (a > 0).$$

Let a_0 and b_0 be two points in the interval $[0, 2\pi]$ satisfying $\pi < a_0 < b_0$, $p_0'(a_0) = 0$ (modulo 2π) and $p_0(b_0) = 0$. Let us set

$$g_1(x) = h(x) + (h(a_0) - h(x))\chi(a_0, b_0),$$

where $\chi(a_0, b_0)$ is the characteristic function of interval $[a_0, b_0]$. We consider the problem:

$$\text{Find } p_1 \in V_1 \text{ such that}$$

$$\int_0^{2\pi} h^3 p_1'(x)\xi'(x)dx = \int_0^{2\pi} g_1(x)\xi'(x)dx, \quad \forall \xi \in V_1 \tag{37}$$

and

$$\int_0^{2\pi} p_1 = a, \tag{38}$$

where $V_1 = \{\xi \in H^1(I), \quad \xi\ is\ 2\pi - \text{periodic}\}$.

Proposition 1 *There exist* $\pi < a_1 < a_0$ *and* $b_1 > b_0$ *(modulo* 2π*) such that* $p_1'(a_1) = 0$, $p_1(b_1) = 0$ *and the solution* p_1 *is an increasing function in the interval* $[a_1, b_1]$.

Suppose that for $n \in \mathbb{N}$ there exist two points a_n and b_n such that $\pi < a_n < b_n$ (modulo 2π) and p_n is an increasing function on the interval $[A_n, b_n]$ with $p_n'(a_n) = 0$ and $p_n(b_n) = 0$. Let us set

$$g_{n+1}(x) = h(x) + (h(a_n) - h(x))\chi(a_n, b_n)$$

and let us consider the variational problem:

Find $p_{n+1} \in V_1$ such that

$$\int_0^{2\pi} h^3 p'_{n+1}(x)\xi'(x)dx = \int_0^{2\pi} g_{n+1}(x)\xi'(x)dx, \quad \forall\xi \in V_1 \tag{39}$$

and

$$\int_0^{2\pi} p_{n+1} = a. \tag{40}$$

Proposition 2 *There exist $\pi < a_{n+1} < a_n$ and $b_{n+1} > b_n$ (modulo 2π) such that $p'_{n+1}(a_{n+1}) = 0$, $p_{n+1}(b_{n+1}) = 0$ and the solution p_{n+1} is an increasing function on $[a_{n+1}, b_{n+1}]$.*

We thus construct a decreasing, minimized sequence $(a_n)_{n\leq 0}$ and a growing sequence $(b_n)_{n\leq 0}$ (Modulo 2π). Then we obtain the following main result:

Theorem 3 *The sequence $(p_n)_{n\geq 0}$ converges uniformly towards the solution p of the free boundary problem $\mathscr{M}$.*

As an example, we show in Fig. 6 the evolution of the sequence $(p_n)_{n\geq 0}$ for a separation function $h(x) = 1 + 0.7\cos(x)$ and a mass constant $a = 0.5$.

3.2 Approximation of the Elrod-Adams Problem

In this section we extend the above method to the general problem (30). Let $q \in V_a$ be the solution of the variational equation

$$\int_\Omega h^3\nabla q\nabla\xi = \int_\Omega h\xi_x, \quad \forall \quad \xi \in V_0. \tag{41}$$

The solution q exists and is unique moreover $q \in C^{1,\alpha}(\Omega)$ for some $\alpha > 0$ (see [23]). We prove that this solution q satisfies the following conservation condition of load with respect to the variable y:

$$\int_0^{2\pi} h^3(x)q(x,y)dx = p_a y, \quad \forall y \in (0,1). \tag{42}$$

We begin by approaching the problem by the line method. We Discretize the interval [0, 1] in N equal parts $y_i = i\Delta y$, $(i = 0, 1, \ldots, N+1)$ of equal step $\Delta y = 1/N$. The space V_0 is approximated by a finite dimension subspace of linear functions with respect to the variable y on each interval (y_i, y_{i+1}). By setting $q_i(x) =$

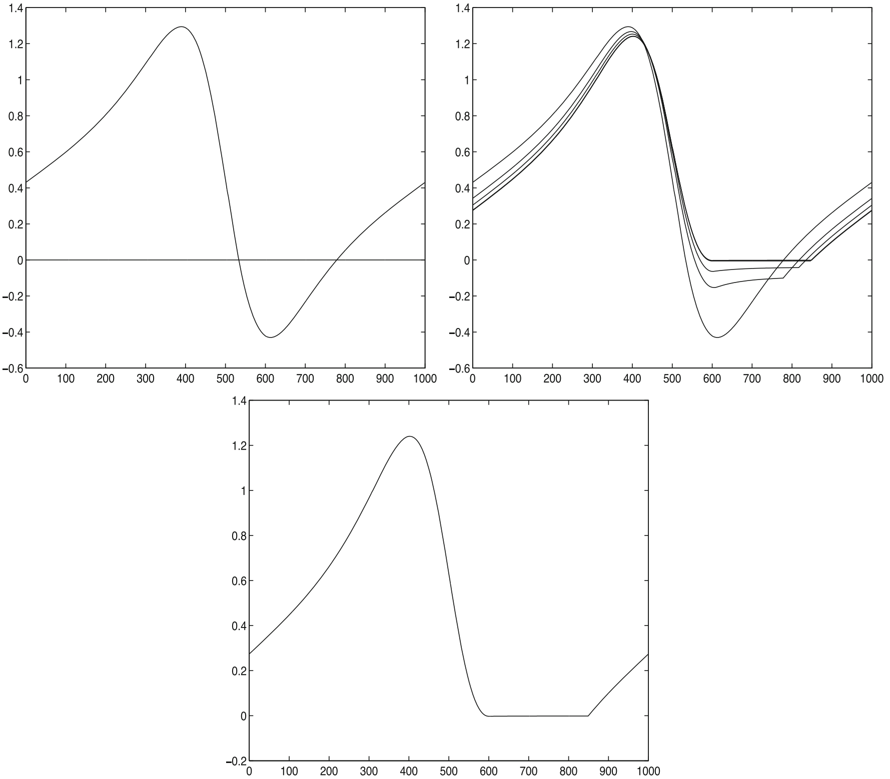


Fig. 6 Pressure evolution

$q_N(x, y_i), (i = 1, \dots, N)$ we obtain the solution of Eq. (42) for the resolution of the following system:

$$\int_0^{2\pi} h^3 q_i' \xi' + \frac{2}{\Delta y^2} \int_0^{2\pi} h^3 q_i \xi = \int_0^{2\pi} h \xi' + \frac{1}{\Delta y^2} \int_0^{2\pi} h^3 [q_{i+1} + q_{i-1}] \xi, \tag{43}$$

$$\int_0^{2\pi} h^3 q_i = i p_a y \int_0^{2\pi} h^3, (i = 1, \dots, N) \tag{44}$$

and

$$q_0(x) = 0, \quad q_{N+1}(x) = p_a. \tag{45}$$

We apply Gauss-Seidel's iterative method to solve the system (43)–(45). We start with $q_i^0(x) = p_a i \Delta y \, (i = 0, 1, \dots, N + 1)$ and for $k = 1, 2, \dots,$ we solve the system:

$$\int_0^{2\pi} h^3 q_i^{k'} \xi' + \frac{2}{\Delta y^2}\int_0^{2\pi} h^3 q_i^k \xi = \int_0^{2\pi} h\xi' + \frac{1}{\Delta y^2}\int_0^{2\pi} h^3 [q_{i+1}^{k-1} + q_{i-1}^k]\xi, \tag{46}$$

$$\int_0^{2\pi} h^3 q_i^k = ip_a y \int_0^{2\pi} h^3, \, (i = 1, \dots, N) \tag{47}$$

and

$$q_0^k(x) = 0, \quad q_{N+1}^k(x) = p_a. \tag{48}$$

To compute an approximation of the solution of the problem (30), we apply a combination of the above Gauss-Seidel method and the method presented in the previous section related to the one-dimensional case. For this, we take $q_i^0(x) = q(x, y_i), (i = 1, \dots, N)$, where q is the solution of the system (46)–(48) and for $k = 1, 2, \dots$, we solve the system:

$$\int_0^{2\pi} h^3 q_i^{k'} \xi' + \frac{2}{\Delta y^2}\int_0^{2\pi} h^3 q_i^k \xi = \int_0^{2\pi} h_{k-1}\xi' + \frac{1}{\Delta y^2}\int_0^{2\pi} h^3 [q_{i+1}^{k-1} + q_{i-1}^k]\xi, \tag{49}$$

$$\int_0^{2\pi} h^3 q_i^k = ip_a y \int_0^{2\pi} h^3, \, (i = 1, \dots, N) \tag{50}$$

and

$$q_0^k(x) = 0, \quad q_{N+1}^k(x) = p_a, \tag{51}$$

where $h_{k-1}(x) = h(x) + (h(a_{k-1}) - h(x))\chi([a_{k-1}, b_{k-1}])$, with $a_{k-1} \in (\pi, 2\pi)$ and $b_{k-1} > a_{k-1}$ (modulo 2π) such that $(q_i^{k-1})'(a_{k-1}) = 0$, $q_i^{k-1}(b_{k-1}) = 0$ and q_i^{k-1} is an increasing function on the interval (a_{k-1}, b_{k-1}). Here $\chi([a_{k-1}, b_{k-1}])$ is the characteristic function of interval $[a_{k-1}, b_{k-1}]$.

Let us set

$$W_i = \{\xi \in H^1(0, 2\pi), \quad \xi \quad 2\pi\textit{-periodic} \quad \textit{and} \quad \int_0^{2\pi} h^3 \xi = ip_a y \int_0^{2\pi} h^3\}$$

and $W = W_1 \times \dots \times W_N$ endowed with norm $\|U\| = \max_{1 \le i \le N} \|U_i\|_i$. We have stated in [28] the following proposition:

Proposition 3 *There exist* $q^* = (q_1^*, \dots, q_N^*) \in W$ *and* $h^* = (h_1^*, \dots, h_N^*) \in (L^\infty(0, 2\pi))^N$ *such that*

$$q^k \rightharpoonup q^* \textit{ weakly in } W \tag{52}$$

and

$$h^k \to h^* \textit{ weakly-* in } (L^\infty(0, 2\pi))^N. \tag{53}$$

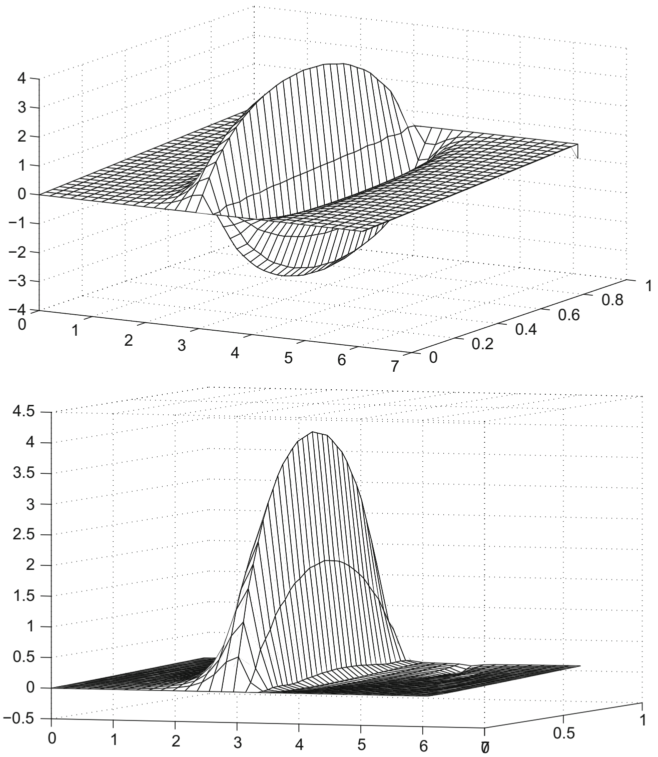


Fig. 7 Initial and final pressure by method (49)–(50)

We consider in the following example the function $h(x) = 1 + 0.9\cos(x)$ and a power supply $p_a = 0.1$. For each iteration k we compute the relative error $\dfrac{|q_k - q_{k-1}\,|}{|q_k\,|}$. Here $|.\,|$ denotes the norm of the space $L^2(\Omega)$ (Figs. 7 and 8).

3.3 Elasto-Hydrodynamic Problem

In several devices used in mechanical engineering, elastic cushions are used and high pressures can cause deformations of the pad. Therefore an additional variable is added to the problem and the separation function is in this case written as

$$h(x, y) = h_0(x) + w(x, y),$$

where $h_0(x)$ is the initial separation and $w(x, y)$ is a function that represents the deformation of the pad. This function satisfies the classical equation of deformable plates (see [14])

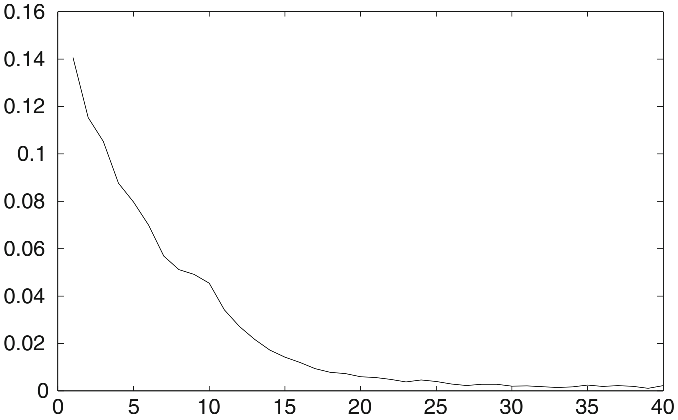


Fig. 8 Evolution of the relative error $\frac{\|q_k - q_{k-1}\|}{\|q_k\|}$

$$\eta \Delta^2 w = p \quad in \quad \Omega, \tag{54}$$

where p is the pressure and η a parameter that represents the elasticity of the pad. The following boundary conditions are also considered

$$w = \Delta w = 0 \quad on \quad \Gamma_0 \cup \Gamma_1, \ w \quad and \quad \Delta w \quad 2\pi\text{-}periodic. \tag{55}$$

These conditions have been introduced in [8] in order to guarantee the positivity of the deformation w. To solve (54)–(55) we propose the following mixed formulation applied in [19]:

$$-\Delta \psi = p \quad in \quad \Omega, \tag{56}$$

$$\psi = 0 \quad on \quad \Gamma_0 \cup \Gamma_1, \quad \psi \quad 2\pi\text{-}periodic, \tag{57}$$

$$-\eta \Delta w = \psi \quad in \quad \Omega \tag{58}$$

and

$$w = 0 \quad on \quad \Gamma_0 \cup \Gamma_1, \quad w \quad 2\pi\text{-}periodic. \tag{59}$$

In addition to the spaces introduced in (28) and (29), we define the set

$$M_0 = \{\varphi \in H^2(\Omega) : \quad \varphi = \Delta\varphi = 0 \quad on \ \ \Gamma_0 \cup \Gamma_1, \ \ \varphi \, and \, \Delta\varphi \quad 2\pi\text{-}periodic\}.$$

A general formulation of the elasto-hydrodynamic coupled problem is given by:

$$\text{Find}\quad (p,\theta,\psi,w)\in H^1(\Omega)\times L^\infty(\Omega)\times M_0\times M_0 \text{ such that}$$

$$\int_\Omega h^3\nabla p\nabla\xi=\int_\Omega h\theta\xi_x,\quad \xi\in V_0, \tag{60}$$

$$\theta\in H(p),\quad h=h_0+w,\quad p\in V_a, \tag{61}$$

$$\int_\Omega \nabla\phi\nabla\varphi=\int_\Omega p\varphi,\quad \forall\varphi\in M_0 \tag{62}$$

and

$$\eta\int_\Omega \nabla w\nabla\varphi=\int_\Omega \phi\varphi,\quad \forall\varphi\in M_0. \tag{63}$$

Equations (60)–(61) represent the hydro-dynamic part and (62)–(63) the elastic one.

The solution of (60)–(63) is approximated by an iterative method which consists in decoupling the hydro-dynamic and elastic parts of the system:

- Start the algorithm with initial values p_0,θ_0,w_0.
- At each step n and for a given thickness function h_{n-1}, we compute the solution p_n,θ_n of the problem (60)–(61).

  Then we solve successively the linear problems (62) and (63).
- We update the thickness function with the formula $h_n=h_0+w_n$ and return in the next step.

By the principle of the low maximum taking into account the boundary conditions we show that $w_n\geq 0$. It is shown that the sequences $(p_n)_{n\geq 0}$, $(\theta_n)_{n\geq 0}$ and $(w_n)_{n\geq 0}$ are bounded in their respective spaces and by applying the compactness arguments we prove their convergences (see [24]).

We apply a variant of the algorithm introduced in (49)–(50). The separation in this case is a function of two variables x and y. Starting from $q_i^0(x)=p_a i\Delta y$, $\forall i=0,1,\ldots,N+1$, at each step k we solve the system

$$\int_0^{2\pi} h_i^3 q_i^{k'}\xi'+\frac{1}{2\Delta y^2}\int_0^{2\pi}\left(h_{i+1}^3+2h_i^3+h_{i-1}^3\right)q_i^k\xi \tag{64}$$

$$=\int_0^{2\pi} h_i^{k-1}\xi'+\frac{1}{2\Delta y^2}\int_0^{2\pi}[\left(h_i^3+h_{i+1}^3\right)q_{i+1}^{k-1}+\left(h_i^3+h_{i-1}^3\right)q_{i-1}^k]\xi, \tag{65}$$

$$\int_0^{2\pi} h^3 q_i^k=ip_a y\int_0^{2\pi} h^3,\ (i=1,\ldots,N) \tag{66}$$

and

$$q_0(x)=0,\quad q_{N+1}(x)=p_a. \tag{67}$$

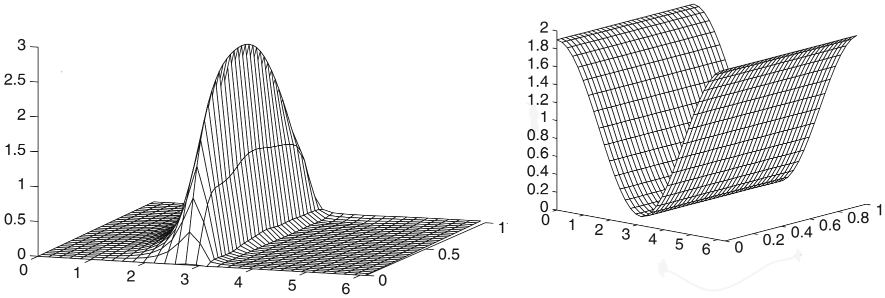


Fig. 9 Pressure p and the separation for a parameter $\eta = 1000$

as in (49)–(50) we take $h_i^{k-1}(x) = h_i(x) + (h_i(a_{k-1}) - h_i(x))\chi([a_{k-1}, b_{k-1}])$, where a_{k-1} and b_{k-1} are two points in the interval $(\pi, 2\pi)$ satisfying $b_{k-1} > a_{k-1}$ (modulo 2π), $(q_i^{k-1})'(a_{k-1}) = 0$, $q_i^{k-1}(b_{k-1}) = 0$ and q_i^{k-1} is an increasing function in (a_{k-1}, b_{k-1}). Here $\chi([a_{k-1}, b_{k-1}])$ is the characteristic function of $[a_{k-1}, b_{k-1}]$.

We apply the Gauss-Seidel method to solve the systems (62) and (63). So for a pressure $(p_i)_{0\leq i\leq N+1}$, obtained by (64)–(67), one solves successively

$$\int_0^{2\pi} \phi_i^{k'}\varphi' + \frac{2}{\Delta y^2}\int_0^{2\pi} \phi_i^k\varphi = \int_0^{2\pi} p_i\varphi' + \frac{1}{\Delta y^2}\int_0^{2\pi} [\phi_{i+1}^{k-1} + \phi_{i-1}^k]\varphi, \tag{68}$$

$$\phi_0(x) = 0, \quad \phi_{N+1}(x) = 0, \tag{69}$$

$$\eta\int_0^{2\pi} w_i^{k'}\varphi' + \frac{2\eta}{\Delta y^2}\int_0^{2\pi} w_i^k\varphi = \int_0^{2\pi} \phi_i\varphi' + \frac{\eta}{\Delta y^2}\int_0^{2\pi} [w_{i+1}^{k-1} + w_{i-1}^k]\varphi \tag{70}$$

and

$$w_0(x) = 0, \quad w_{N+1}(x) = 0. \tag{71}$$

Here the test function φ belongs to $H^1(0, 2\pi)$ and is 2π-periodic.

We present in Figs. 9 and 10 the numerical simulations of the pressure and the deformation for a rigid device which corresponds to a coefficient $\eta = 1000$ and then for a device which corresponds to a coefficient $\eta = 0.01$.

4 The Evolution Free Boundary Problem

The free evolution boundary problem is based on a mathematical formulation in the domain $Q = [0, T]\times\Omega$ with $T > 0$ and $\Omega = [0, 2\pi]\times[0, 1]$. The device is supplied across the border $\Sigma_a = [0, 2\pi] \times \{1\} \times [0, T]$ where we suppose $p = p_a > 0$. We put $p = 0$ on the boundary $\Sigma_0 = [0, 2\pi] \times \{0\} \times [0, T]$ (Fig. 11).

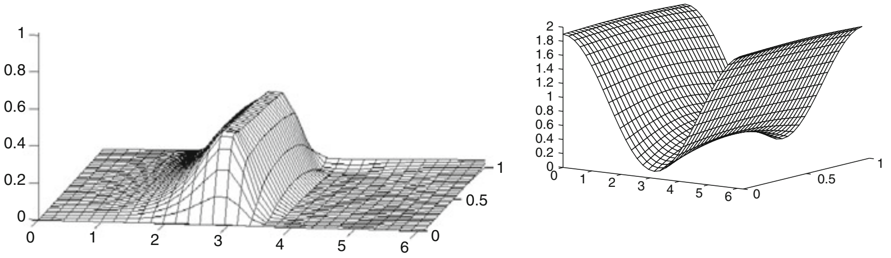
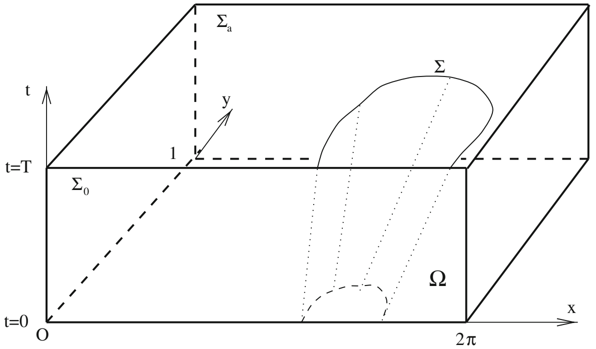


Fig. 10 Pressure p and the separation for a parameter $\eta = 0.01$

Fig. 11 The domain Q



In the part occupied by the fluid, the pressure satisfies the dynamic equation of Reynolds:

$$\frac{\partial h}{\partial t} - div(h^3 \nabla p) = -\frac{\partial h}{\partial x} \quad and \quad \gamma = 1 \quad if \quad p > 0. \tag{72}$$

The saturation function γ satisfies the following conservation law in the cavitation area:

$$\frac{\partial (h\gamma)}{\partial x} + \frac{\partial (h\gamma)}{\partial t} = 0 \quad and \quad 0 \leq \gamma \leq 1 \quad if \quad p = 0. \tag{73}$$

On the free boundary $\Sigma = \overline{[p = 0]} \cap \overline{[p > 0]}$, we take $p = 0$ and the flux satisfies the conservation condition:

$$h^3 \frac{\partial p}{\partial n} = (1 - \gamma) h \cos(n, i), \tag{74}$$

where n is the unit normal vector at the free boundary and (n, i) the angle between n and the unit vector i. The function $h(x, y, t)$ represents the separation between the surfaces. We suppose that $h \in C^\infty(Q)$ is periodic in x and satisfies

$$h(x, y, 0) = h(x, y, T), \quad \forall (x, y) \in (0, 2\pi) \times (0, 1).$$

The strong formulation above can be supplemented by considering an initial value for γ. Let $p_0 \in L^2(\Omega)$ with $p_0 \geq 0$ and let $\gamma_0 \in H(p_0)$ (see formula (33)). The following weak formulation of the problem is given:

Problem $\mathscr{P}$

$$\text{Find} \quad (p, \gamma) \in L^2(0, T; H^1(\Omega)) \times L^\infty(Q) \text{ such that}$$

$$p \geq 0, \quad \gamma \in H(p) \quad a.e. \quad in \quad Q, \tag{75}$$

$$-\int_Q h\gamma\xi_t + \int_Q h^3 \nabla p \nabla \xi = \int_Q h\gamma\xi_x + \int_\Omega h(.,.,0)\gamma_0\xi(.,.,0), \tag{76}$$

$$\forall \xi \in H^1(Q)\, 2\pi x\text{-periodic},\ \xi(x, y, T) = 0 \text{ a.e. in } \Omega, \quad \xi = 0 \quad on \quad \Sigma_0 \cup \Sigma_a$$

and

$$p = 0 \quad on \quad \Sigma_0, \quad p = p_a \quad on \quad \Sigma_a. \tag{77}$$

4.1 Existence, Uniqueness and Continuity of the Solution

Theorem 4 *Problem $\mathscr{P}$ has a unique solution $(p, \gamma) \in L^2(0, T; H^1(\Omega)) \times L^\infty(\Omega)$.*

The proof of this result is given in [1, 25]. It is based on classical elliptic regularization techniques. The function of Heaviside H is approximated by a regular function and the related regularized problem is studied. A priori estimates are stated by applying compactness arguments and we obtain the convergence to the solution of the problem $\mathscr{P}$.

To complete the study of the problem, we established in [25] the following monotonic property of $h\gamma$:

Theorem 5 *Let (p, γ) be the solution of $\mathscr{P}$ and χ the characteristic function of the set $[p > 0]$, then*

$$\big(h\gamma\big)_t + \big(h\gamma\big)_{x_1} - (h_{x_1} + h_t)\chi \geq 0.$$

We use this theorem to demonstrate the following strong continuity of γ:

Theorem 6 *Let (p, γ) be the solution of problem $\mathscr{P}$, then*

$$h\gamma \in C^0([0, T], L^q(\Omega)), \quad \forall q \in [1, \infty).$$

The main result established in [25] concerns the uniqueness of the solution of the problem $\mathscr{P}$. To state this result we have first demonstrated the following comparison result which makes it possible to compare tow solutions of $\mathscr{P}$ when we can compare their initial values and their values on the boundary Σ_a. Concretely, if (p_1, γ_1) and (p_2, γ_2) are two pairs satisfying

$$(p_i, \gamma_i) \in L^2(0, T, H^1(\Omega)) \times L^\infty(Q), \quad p_i \text{ is } 2\pi\text{-periodic}, \tag{78}$$

$$p_i \geq 0,\ \gamma_i \in H(p_i) \quad a.e. \quad in \quad Q, \tag{79}$$

$$-\int_Q h\gamma_i\xi_t + \int_Q h^3\nabla p_i \nabla\xi = \int_Q h\gamma_i\xi_x, \quad \forall\xi \in V, \tag{80}$$

$$p_i|_{\Sigma_a} = p_a^i,\ p_i|_{\Sigma_0} = 0 \tag{81}$$

and

$$p_a^1 \leq p_a^2 \text{ on } \Sigma_a \text{ and } \gamma_1(.,0) \leq \gamma_2(.,0) \text{ a.e. } \Omega. \tag{82}$$

Theorem 7 *Let $(p_i, \gamma_i)(i = 1, 2)$ be two pairs such that (78)–(82) hold, then for all $\xi \in \mathscr{D}^+(0, 1)$ we have*

$$\int_Q h^3 \frac{\partial(p_1 - p_2)^+}{\partial y}\xi' dxdydt \leq 0$$

and then

$$p_1 \leq p_2 \quad a.e. \quad in \quad Q.$$

Theorem 8 *Let $(p_i, \gamma_i)(i = 1, 2)$ be two pairs such that (78)–(82) hold, then*

$$\int_Q h(\gamma_1 - \gamma_2)^+(\xi_t + \xi_x)dxdydt = 0$$

for all $\xi \in H^1(Q)$ 2π-periodic and $\xi(x, y, 0) = \xi(x, y, T) = 0$ a.e. in Ω.

From theses results we obtain finally the uniqueness of the solution of problem $\mathscr{P}$:

Theorem 9 *There exist a unique solution (p, γ) of problem $\mathscr{P}$.*

See [25] for details.

4.2 A Semi-Discretised Euler Scheme

We applied in [27] an implicit Euler scheme to discretize time variable of problem $\mathscr{P}$. For a step $\tau = T/N$ we consider the family of discretized problems:

Problem $\mathscr{P}_n$

$$\text{Find } (p_n, \gamma_n) \in H^1(\Omega) \times L^\infty(\Omega) \text{ such that}$$

$$\int_\Omega h^3 \nabla p_n \nabla \xi + \frac{1}{\tau}\int_\Omega h\gamma_n \xi - \int_\Omega h\gamma_n \xi_x = \frac{1}{\tau}\int_\Omega h\gamma_{n-1}\xi, \ \forall \xi \in V_0, \tag{83}$$

$$p_n \in V_a, \quad p_n \geq 0 \tag{84}$$

and

$$\gamma_n \in H(p_n). \tag{85}$$

By applying the elliptic regularization techniques, we show the existence of solution for problem $\mathscr{P}_n$. It is then shown that the sequences of solutions $(p_n)_{n\geq 0}$ and $(\gamma_n)_{n\geq 0}$ are bounded in their spaces respectively. Then by the compactness arguments one obtains the weak convergences of the two sequences. Nevertheless, these weak convergences are insufficient to reach the limit in the nonlinear term of (83). A monotonic property of these sequences has been established for this purpose:

Let be (p_1, γ_1) and (p_2, γ_2) such that

$$(p_i, \gamma_i) \in H^1(\Omega) \times L^\infty(\Omega), \tag{86}$$

$$\frac{1}{\tau}\int_\Omega h\gamma_i \xi + \int_\Omega h^3 \nabla p_i \nabla \xi - \int_\Omega h\gamma_i \xi_x = \frac{1}{\tau}\int_\Omega h f_i \xi, \quad \forall \xi \in V_0, \tag{87}$$

$$p_i \geq 0, \ H(p_i) \leq \gamma_i \leq 1, \quad a.e. \quad in \quad \Omega \tag{88}$$

and

$$f_i \in L^\infty(\Omega)(i = 1, 2). \tag{89}$$

Then we have

Theorem 10 *For all $\xi \in \mathscr{D}(0, 1)$ such that $\xi \geq 0$ a.e. in $(0, 1)$ we have*

$$\int_\Omega h^3(x)(p_1 - p_2)^+ \xi'' dxdy \geq -1/\tau \int_\Omega h(f_1 - f_2)\chi([p_1 > p_2])\xi(y) dxdy$$

and then

$$f_1 \leq f_2 \text{ a.e. in } \Omega \Longrightarrow p_1 \leq p_2 \text{ a.e. in } \Omega$$

Theorem 11 *For all $n \in \mathbb{N}$ we have*

$$p_n \leq p_{n+1}$$

moreover if $\gamma_{n-1} \leq \gamma_{n-2}$ a.e. in Ω then

$$p_n = p_{n-1} \quad and \quad \gamma_n \leq \gamma_{n-1} a.e. \quad in \quad \Omega.$$

We get finally

Corollary 1 *If $\gamma_0 = 1$ a.e. in Ω then $p_n = p_1$, $\forall n \geq 1$ and the sequence (γ_n) converges to some limit (γ_*)in $L^\infty(\Omega)$.*

The condition $\gamma_0 = 1$ means that the bearing is full and there is no cavitated region. It seems natural to take this as an initial value. By other hand, convergence of the sequences $(p_n)_{n\geq 0}$ and $(\gamma_n)_{n\geq 0}$ does not depend on the parameter τ. This property is very important for the numerical approach of the problem.

For the numerical simulation of the problem $\mathscr{P}_n$ we apply a duality method introduced in [9] in the context of variational inequalities. To do this, we introduce the multiplier $\beta_n = \gamma_n - \omega p_n$ and the problem is written as:

Find $(p_n, \beta_n) \in H^1(\Omega) \times L^\infty(\Omega)$ such that

$$\int_\Omega h^3 \nabla p_n \nabla \xi + \frac{\omega}{\tau} \int_\Omega h p_n \xi - \omega \int_\Omega h p_n \xi_x = \frac{1}{\tau} \int_\Omega h \beta_n \xi$$

$$+ \int_\Omega h \beta_n \xi_x + \frac{1}{\tau} \int_\Omega h \gamma_{n-1} \xi, \tag{90}$$

$$p_n \in V_a,\, p_n \geq 0 \tag{91}$$

and

$$\beta_n \in H(p_n) - \omega p_n. \tag{92}$$

In [9], the following equivalence is stated:

$$\beta_n \in H(p_n) - \omega p_n \equiv \beta_n = H_\lambda^\omega(p_n + \lambda \beta_n), \tag{93}$$

where H_ω^λ is the Yosida approximation of operator $H - \omega I$.

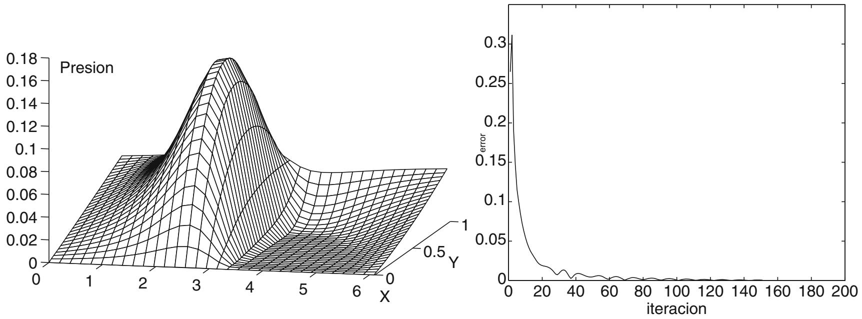


Fig. 12 Pressure p and relative error for $\alpha = 0.5$

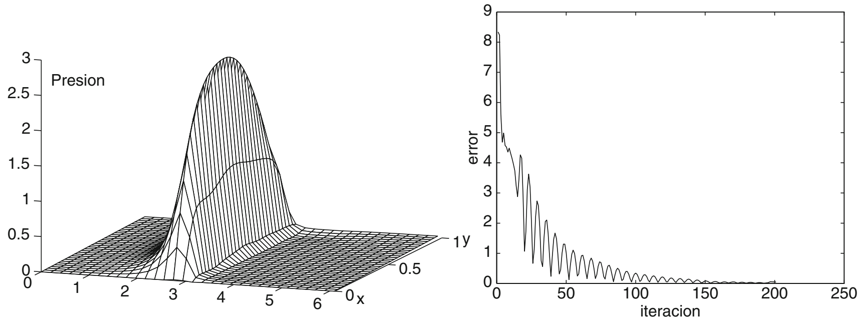


Fig. 13 Pressure p and relative error for $\alpha = 0.9$

We start the iterations with $p_n^0 = p_{n-1}$ and $\beta_n^0 = \beta_{n-1} - \omega p_{n-1}$. At each step

- The multiplier is updated by $\beta_n^j = H_\lambda^\omega(p_n^{j-1} + \lambda\beta_n^{j-1})$
- We solve the equation

$$\int_\Omega h^3 \nabla p_n^j \nabla \xi + \frac{\omega}{\tau}\int_\Omega h p_n^j \xi - \omega \int_\Omega h p_n^j \xi_x = \frac{1}{\tau}\int_\Omega h \beta_n^j \xi + \int_\Omega h \beta_n^j \xi_x + \frac{1}{\tau}\int_\Omega h \gamma_{n-1}\xi. \tag{94}$$

At each step, the error $|\beta_n^j - \beta_n^{j-1}|_\infty$ is given (Figs. 12 and 13).

References

1. S. Alvarez, Problemas de Frontera Libre en Teoria de Lubricación. Tesis Doctoral, Universidad Complutense de Madrid (1986)
2. S. Alvarez, Qualitative properties of the free boundary of the Reynolds equation in lubrication. Publ. Math. **33**, 235–251 (1989)

3. S. Alvarez, J. Carrillo, A free boundary problem in theory of lubrication. Commun. Partial Differ. Equ. **19**(11–12), 1743–1761 (1994)
4. I. Babuska, W. Rheinboldt, Error estimates for adaptive finite element computations. SIAM J. Numer. Anal. **15**, 736–754 (1978)
5. G. Bayada, M. Chambat, Nonlinear variational formulation for a cavitation problem in lubrication. J. Math. Anal. Appl. **90**, 286–298 (1982)
6. G. Bayada, M. Chambat, Sur quelques modélisations de la zone de cavitation en lubrification hydrodynamique. J. Theor. Appl. Mech. **5**, 701–729 (1986)
7. G. Bayada, M. Chambat, The transition between the Stokes equation and the Reynolds equation: a mathematical proof. Appl. Math. Optim. **14**, 73–93 (1986)
8. G. Bayada, J. Durany, C. Vazquez, Existence of a solution for a lubrication problem in elastic journal-bearing devices with thin bearing. Math. Methods Appl. Sci. **18**, 255–266 (1995)
9. A. Bermúdez, C. Moreno, Duality methods for solving variational inequalities. Comput. Math. Appl. **7**, 43–58 (1981)
10. A. Bermudez, C. Moreno, Duality Methods for solving variational inequalities. Comput. Methods Appl. Mech. Eng. **75**, 455–466 (1989)
11. H. Brezis, *Operateurs Maximaux Monotones et Semigroups de Contractions dans les Espaces de Hilbert* (North-Holland, Amsterdam, 1973)
12. A. Cameron, *Basic Lubrication Theory*. Ellis Horwood Series (Ellis Horwood, Chichester, 1981)
13. O. Chau, D. Goeleven, R. Oujja, An Adaptive finite element method for solving a free boundary problem with periodic boundary condition in lubrication theory. *Mathematical Analysis, Approximation Theory, and Their Applications*, ed. by Th.M. Rassias, V. Gupta (Springer, New York, 2016), pp. 107–120
14. G. Cimatti, How the Reynolds equation is related to the Stokes equation. Appl. Math. Optim. **10**, 223–248 (1983)
15. P. Clement, Approximation by finite element functions using local regularization. RAIRO Anal. Numer. **9**, 77–84 (1975)
16. W. Dorfler, A convergent adaptive algorithm for Poisson's equation. SIAM J. Numer. Anal. **33**(3), 1106–1124 (1996)
17. J. Dowson, C.M. Taylor, Cavitation in bearings. Annu. Rev. Fluid Mech. **1**, 35–66 (1979)
18. A. El Alaoui, G. Bayada, Une méthode de type caractéristiques pour la résolution d'un probleme de lubrification hydrodynamique. Math. Modell. Numer. Anal. **25**(4), 395–423 (1991)
19. R. Glowinski, O. Pironneau, Numerical methods for the first biharmonic equation and the two-dimensional Stokes problem. SIAM Rev. **21**, 167–212 (1979)
20. D. Goeleven, V.H. Nguyen, On the one-dimensional nonlinear elasto-hydrodynamic lubrication. Bull. Aust. Math. Soc. **50**, 253–372 (1994)
21. J.S. Guo, A variational inequality associated with a lubrication problem. Nonlinear Anal. **16**, 13–14 (1991)
22. D. Kinderlehrer, G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications* (Academic, New York, 1980)
23. H. Lewy, G. Stampacchia, On the regularity of the solution of a variational inequality. Commun. Pure Appl. Math. **22**, 153–188 (1969)
24. R. Oujja, A new method for cavitations approximation in some general lubrication devices. Appl. Math Comput. **181**, 1645–1656 (2006)
25. R. Oujja, S. Alvarez, A monotonicity result in a moving free boundary problem related to lubrication with cavitation. Adv. Math. Sci. Appl. **11**(1), 161–185 (2001)
26. R. Oujja, S. Alvarez, An iterative method for solving a free boundary problem for an infinite journal bearing. Appl. Math. Comput. **122**(1), 15–26 (2001)
27. R. Oujja, S. Alvarez, On the uniqueness of the solution of an evolution free boundary problem in theory of lubrication. Nonlinear Anal. Theory Methods Appl. **54**(5), 845–872 (2003)
28. R. Oujja, S. Alvarez, A new numerical approach for a lubrication free boundary problem. Appl. Math. Comput. **148**, 393–405 (2004)

29. A. Pazy, Semigroups of non-linear contractions in Hilbert Spaces, in *Problems in Nonlinear Analysis*, C.I.M.E. (Ed. Cremonese, Roma, 1971)
30. O. Reynolds, On the theory of lubrication and its applications to M. Beauchamp Toaer's experiments. Philos. Trans. R. Soc. Lond. **A11**(7), 157–234 (1886)
31. M. Rodriguez, A. Liñan, Cavitation in short bearing. Trans. ASME **107**, 142–144 (1985)
32. M.D. Savage, Cavitation in lubrication. Part: 1 on boundary conditions and cavity-fluid interfaces. J. Fluid Mech. **80**(4), 743–755 (1977)
33. C. Vázquez, Existence and uniqueness of solution for a lubrication problem with cavitation in a journal bearing with axial supply. Adv. Math. Sci. Appl. **4**(2), 313–331 (1994)

On the Spectrum of a Nonlinear Two Parameter Matrix Eigenvalue Problem

Michael Gil'

1 Introduction and Statement of the Main Result

The present paper is concerned with the problem

$$T_1 v_1 = (\lambda_1 A_{11} + \lambda_2 A_{12} + \lambda_1 \lambda_2 A_{13}) v_1, \tag{1.1}$$

$$T_2 v_2 = (\lambda_1 A_{21} + \lambda_2 A_{22} + \lambda_1 \lambda_2 A_{23}) v_2, \tag{1.2}$$

where $\lambda_1, \lambda_2 \in \mathbf{C}$; $v_p \in \mathbf{C}^{n_p}$; $T_p, A_{pj} \in \mathbf{C}^{n_p \times n_p}$ $(p = 1, 2;\ \ j = 1, 2, 3)$.

Denote problem (1.1), (1.2) by Λ. If for some λ_1, λ_2 problem Λ has a solution $v_1 \neq 0$ and $v_2 \neq 0$, then the pair $\lambda = (\lambda_1, \lambda_2)$ and (v_1, v_2) is called the eigenvalue of Λ and eigenvector corresponding to λ, respectively. Besides, λ_1 and λ_2 are the first coordinate and second one of λ. The set of all the eigenvalues of Λ is the spectrum and is denoted by $\Sigma(\Lambda)$; the set of all the p-th coordinates $(p = 1, 2)$ is denoted by $\sigma_p(\Lambda)$. So $\Sigma(\Lambda)$ is the pair $(\sigma_1(\Lambda), \sigma_2(\Lambda))$. In the general case, $\Sigma(\Lambda)$ can be an infinite set.

Multiparameter eigenvalue problems (linear and nonlinear) arise in numerous applications, cf. [8, 9, 14, 17]. The classical results on that problem can be found in the books [1, 16]. For some recent presentations of multiparameter spectral theory problems we refer the interested reader to [13, 14, 18, 19]. Problem (1.1), (1.2) has been deeply investigated in the paper [17] in connection with stability theory of delay-differential equations. In [17] it was also shown that a wide class of polynomial two parameter problems can be reduced to problem (1.1), (1.2).

M. Gil' (✉)
Department of Mathematics, Ben Gurion University of the Negev, Beer-Sheva, Israel
e-mail: gilmi@bezeqint.net

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization
and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_13

In the present paper we estimate the quantity $r_s^{(p)}(\Lambda) = \max\{|s| : s \in \sigma_p(\Lambda)\}$, which will be called *the spectral radius, corresponding to* $\sigma_p(\Lambda)$. To the best of our knowledge, bounds for the spectral radius have been obtained only in the case of linear two parameter eigenvalue problems, cf. [6, 7, 12, 24]. We also investigate perturbations of the considered problem and extend the Gershgorin type bounds for spectra to problem (1.1), (1.2).

Although problem (1.1), (1.2) can be reduced to a linear problem, cf. [14, 22] and references given therein, in appropriate situations it is preferable do not linearize it, since the linearization leads to large-dimensional matrices. In addition, nonlinear problems with "good" matrices, such as such as selfadjoint, unitary or diagonally dominant ones, are reduced to linear problems which do not have these properties. Because of this we do not linearize the considered problem.

Put

$$K_p = A_{13} \otimes A_{2p} - A_{1p} \otimes A_{23} \quad (p = 1, 2),$$

$$Z_{11} = T_1 \otimes A_{23} - A_{13} \otimes T_2 - A_{11} \otimes A_{22} + A_{12} \otimes A_{21},\ Z_{12} = T_1 \otimes A_{22} - A_{12} \otimes T_2, \tag{1.3}$$

where $\otimes$ means the Kronecker product, cf. [15]. Similarly,

$$Z_{21} = T_1 \otimes A_{23} - A_{13} \otimes T_2 - A_{12} \otimes A_{21} + A_{11} \otimes A_{22},\ Z_{22} = T_1 \otimes A_{21} - A_{11} \otimes T_2. \tag{1.4}$$

Our main tool in this paper is the norm estimate for the operator inverse to K_p.

Introduce the notations. Let $\mathbf{C}^n$ be the complex n-dimensional Euclidean space with a scalar product $(.,.)$, the Euclidean norm $\|.\| = \sqrt{(.,.)}$ and the unit matrix I. For a linear operator A in $\mathbf{C}^n$ (matrix), $\|A\| = \max_{x \in \mathbf{C}^n} \|Ax\|/\|x\|$ is the spectral (operator) norm, A^* is the adjoint operator, $\|A\|_F$ is the Frobenius norm: $\|A\|_F^2 = \mathrm{tr}(A^*A)$, $\sigma(A)$ denotes the spectrum, $\lambda_k(A)$ $(k = 1, \dots, n)$ are the eigenvalues with their multiplicities, A^{-1} is the inverse operator, and $R_\lambda(A) = (A - \lambda I)^{-1}$ $(\lambda \notin \sigma(A))$ is the resolvent, $r_s(A)$ denotes the upper spectral radius and $r_{low}(A)$ denotes the lower spectral radius: $r_{\mathrm{low}}(A) := \min_k |\lambda_k(A)|$.

By Schur's theorem [21, Section I.4.10.2], for any operator A in $\mathbf{C}^n$, there is an orthogonal normal basis (Schur's basis) $\{e_k\}_{k=1}^n$ in which A is represented by a triangular matrix:

$$Ae_k = \sum_{j=1}^{k} a_{jk} e_j \text{ with } a_{jk} = (Ae_k, e_j) \quad (j = 1, \dots, n),$$

and $a_{jj} = \lambda_j(A)$. So $A = D_A + V_A$ $(\sigma(A) = \sigma(D_A))$ with a normal (diagonal) matrix D_A defined by $D_A e_j = \lambda_j(A) e_j$ $(j = 1, \dots, n)$ and a nilpotent (strictly upper-triangular) matrix V_A defined by

$$V_A e_k = \sum_{j=1}^{k-1} a_{jk} e_j \quad (k = 2, \dots, n),\ V_A e_1 = 0.$$

As it is well-known the Schur basis is not unique. Let $|A|$ mean the operator, whose entries in some its Schur basis $\{e_k\}$ are the absolute values of the entries of operator A in that basis. That is,

$$|A|e_k = \sum_{j=1}^{k} |a_{jk}|e_j \quad (j = 1, \dots, n).$$

We will call $|A|$ the absolute value of A with respect to its Schur basis $\{e_k\}$. The smallest integer $\nu_A \leq n$, such that $|V_A|^{\nu_A} = 0$ will be called *the nilpotency index of A*.

Denote

$$g_0(K_p) := \frac{1}{\sqrt{2}} \|K_p^* - K_p\|_F \quad (p = 1, 2).$$

It is simple to see that

$$\|K_1^* - K_1\|_F \leq \|A_{13} - A_{13}^*\|_F \|A_{21}\|_F + \|A_{23} - A_{23}^*\|_F \|A_{11}\|_F$$

$$+\|A_{13}\|_F \|A_{21} - A_{21}^*\|_F + \|A_{23}\|_F \|A_{11} - A_{11}^*\|_F.$$

Similarly $\|K_2^* - K_2\|_F$ can be estimated. In addition, put

$$\gamma_0(K_p) := \sum_{j=0}^{\nu_{K_p}-1} \frac{g_0^j(K_p)}{\sqrt{j!}\, r_{\text{low}}^{j+1}(K_p)}.$$

In the next section we show that $\gamma_0(K_p)$ gives us the bound for the norm of K_p^{-1}. Obviously, $\nu_{K_p} < n_1 n_2$. So we can write

$$\gamma_0(K_p) \leq \sum_{j=0}^{n_1 n_2 - 1} \frac{g_0^j(K_p)}{\sqrt{j!}\, r_{\text{low}}^{j+1}(K_p)}.$$

Below we show that under additional conditions that inequality can be improved. If all the operators A_{jk} are Hermitian, then $g_0(K_p) = 0$ $(p = 1, 2)$ and therefore, $\gamma_0(K_p) = 1/r_{\text{low}}(K_p)$.

Now we are in a position to formulate the main result of this paper.

Theorem 1 *Let* $r_{\text{low}}(K_p) > 0$. *Then*

$$r_s^{(p)}(\Lambda) \leq \frac{1}{2}\gamma_0(K_p)\|Z_{p1}\| + \left(\frac{1}{4}\gamma_0^2(K_p)\|Z_{p1}\|^2 + \gamma_0(K_p)\|Z_{p2}\|\right)^{1/2} \quad (p = 1, 2). \tag{1.5}$$

This theorem is proved in the next section.

If all the operators A_{jk} are Hermitian, then as it was above mentioned, $\gamma_0(K_p) = 1/r_{\mathrm{low}}(K_p)$. In this case

$$r_s^{(p)}(\Lambda) \leq \frac{\|Z_{p1}\|}{2r_{\mathrm{low}}(K_p)} + \left(\frac{\|Z_{p1}\|^2}{4r_{\mathrm{low}}^2(K_p)} + \frac{\|Z_{p2}\|}{r_{\mathrm{low}}(K_p)}\right)^{1/2}. \tag{1.6}$$

Below we discuss the sharpness of Theorem 1.

2 Proof of Theorem 1

For an $A \in \mathbf{C}^{n\times n}$ put

$$g(A) := \left(\|A\|_F^2 - \sum_{k=1}^n |\lambda_k(A)|^2\right)^{1/2}.$$

Usually $g(A)$ is called the measure of nonnormality. It was introduced by Henrici in 1962, cf. [2, p. 102].

The following relations are checked in [10, Section 2.1]:

$$g^2(A) \leq \|A\|_F^2 - |\mathrm{tr}\ (A^2)|, \tag{2.1}$$

$$g(A) \leq \frac{1}{\sqrt{2}}\|A - A^*\|_F. \tag{2.2}$$

If matrices A_1 and A_2 have a joint Schur's basis, then $g(A_1+A_2) \leq g(A_1)+g(A_2)$. In addition, by the inequality between the geometric and arithmetic mean values,

$$(\frac{1}{n}\sum_{k=1}^n |\lambda_k(A)|^2)^n \geq (\prod_{k=1}^n |\lambda_k(A)|)^2.$$

Hence,

$$g^2(A) \leq \|A\|_F^2 - n(\det(A))^{2/n}. \tag{2.3}$$

Lemma 1 *For any $A \in \mathbf{C}^{n\times n}$ one has*

$$\|(A - \lambda I)^{-1}\| \leq \sum_{j=0}^{\nu_A-1} \frac{g^j(A)}{\sqrt{j!}\rho^{j+1}(A,\lambda)} \ \ (\lambda \notin \sigma(A)),$$

where $\rho(A,\lambda) = \min_k |\lambda - \lambda_k(A)|$.

For the proof see [12, Lemma 2.2]. The later lemma implies

$$\|K_p^{-1}\| \le \gamma(K_p), \tag{2.4}$$

where

$$\gamma(K_p) := \sum_{j=0}^{\nu_{K_p}-1} \frac{g^j(K_p)}{\sqrt{j!}r_{\text{low}}^{j+1}(K_p)}.$$

Due to (2.2) $g(K_p) \le \frac{1}{\sqrt{2}}\|K_p - K_p^*\|_F = g_0(K_p)$, and, consequently,

$$\|K_p^{-1}\| \le \gamma_0(K_p). \tag{2.5}$$

Furthermore, we need the following result proved in [17, Theorem 3].

Theorem 2 *If* (λ_1, λ_2) *is a solution of* (1.1), (1.2) *with corresponding eigenvector* (v_1, v_2)*, then:* λ_1 *is an eigenvalue of the problem*

$$\begin{gathered}[\lambda_1^2(A_{13} \otimes A_{21} - A_{11} \otimes A_{23}) + \lambda_1(T_1 \otimes A_{23} - A_{13} \otimes T_2 - A_{11} \otimes A_{22} + A_{12} \otimes A_{21}) \\ +T_1 \otimes A_{22} - A_{12} \otimes T_2](v_1 \otimes v_2) = 0\end{gathered} \tag{2.6}$$

and λ_2 *is an eigenvalue with of the problem*

$$\begin{gathered}[\lambda_2^2(A_{13} \otimes A_{22} - A_{12} \otimes A_{23}) + \lambda_2(T_1 \otimes A_{23} - A_{13} \otimes T_2 - A_{12} \otimes A_{21} + A_{11} \otimes A_{22}) \\ +T_1 \otimes A_{21} - A_{11} \otimes T_2](v_1 \otimes v_2) = 0.\end{gathered} \tag{2.7}$$

To finish the proof of Theorem 1 rewrite Eqs. (2.6), (2.7) as

$$[\lambda_p^2 K_p + \lambda_p Z_{p1} + Z_{p2}](v_1 \otimes v_2) = 0 \;\; (p = 1, 2). \tag{2.8}$$

Hence it follows

$$|\lambda_p|^2 \le |\lambda_p| \|K_p^{-1} Z_{11}\| + \|K_p^{-1} Z_{p2}\|$$

and therefore

$$|\lambda_p| \le \frac{1}{2}\|K_p^{-1} Z_{p1}\| + (\frac{1}{4}\|K_p^{-1} Z_{p1}\|^2 + \|K_p^{-1} Z_{p2}\|)^{1/2}. \tag{2.9}$$

Now (2.5) implies the required result. Q.E.D.

Remark 1 Consider the problem

$$T_1 v_1 = \lambda_1 \sum_{k=0}^{m} \lambda_2^k B_{1k} v_1 + \sum_{k=1}^{m} \lambda_2^k C_{1k} v_1, \tag{2.10}$$

$$T_2 v_2 = \lambda_1 \sum_{k=0}^{m} \lambda_2^k B_{2k} v_1 + \sum_{k=1}^{m} \lambda_2^k C_{2k} v_2, \tag{2.11}$$

where $B_{pj}, C_{pj} \in \mathbf{C}^{n_p \times n_p}$ $(p = 1, 2; \; j = 1, 2, 3)$. That problem plays an essential role in the theory of differential equations with several delays; as it is shown in [17, Theorem 4], by Theorem 2 problem (2.10), (2.11) can be reduced to the problem

$$[T_1 \otimes B_{20} - B_{10} \otimes T_2 + \sum_{k=1}^{m} \lambda_2^k (T_1 \otimes B_{2k} - B_{1k} \otimes T_2 - C_{1k} \otimes B_{20} + B_{10} \otimes C_{2k})$$

$$+ \sum_{j,k=1}^{m} \lambda_2^{k+j} (B_{1k} \otimes C_{2j} - C_{1k} \otimes B_{2j})](v_1 \otimes v_2).$$

3 Matrices with Joint Schur Basis

We will say that two operators are simultaneously triangularizable if they can be reduced to the triangular form by the same unitary operator. That is, they have a joint Schur basis. In this section we considerably improve Theorem 1 in the case when the matrices simultaneously triangularizable. In appropriate situations given problems can be considered as perturbations of problems with triangularizable matrices.

Throughout this section it is supposed that

$$\begin{aligned} &A_{1p} \text{ and } A_{13} \text{ are simultaneously triangularizable, as well as} \\ &A_{2p} \text{ and } A_{23} \text{ are simultaneously triangularizable } (p = 1, 2). \end{aligned} \tag{3.1}$$

An important example of K_p under condition (3.1) is the Kronecker sum $A_1 \otimes I + I \otimes A_2$. It should be noted that the Kronecker sum plays an essential role in the theory of matrix equations. For more details on matrix and operator equations see [3] and references given therein.

Simple calculations show that the eigenvalues of K_p are

$$\lambda_{st}(K_p) = \lambda_s(A_{13})\lambda_t(A_{2p}) - \lambda_s(A_{1p})\lambda_t(A_{23}) \;\; (s = 1, \ldots, n_1; \; t = 1, \ldots, n_2), \tag{3.2}$$

provided condition (3.1) holds. Again it is assumed that

$$r_{\text{low}}(K_p) = \min_{t,s} |\lambda_{st}(K_p)| > 0 \quad (p = 1, 2). \tag{3.3}$$

Put

$$\hat{g}(K_p) := (\tau_{2p} + g(A_{2p}))g(A_{13}) + \tau(A_{13})g(A_{2p}) + (\tau(A_{23}) + g(A_{23}))g(A_{1p}) \\ +\tau(A_{1p})g(A_{23}),$$

where

$$\tau(A) := (\sum_{k=1}^{n} |\lambda_k(A)|^2)^{1/2} \quad (A \in \mathbf{C}^{n\times n}).$$

It is clear that

$$\tau(A) \le r_s(A)\sqrt{n} \text{ and } \tau(A) \le \|A\|_F. \tag{3.4}$$

Theorem 3 *Let conditions* (3.1) *and* (3.3) *hold. Then*

$$\nu_{K_p} \le n_1 + n_2 + \min\{n_1, n_2\} - 2 \tag{3.5}$$

and $\|K_p^{-1}\| \le \hat{\gamma}(K_p)$, *where*

$$\hat{\gamma}(K_p) := \sum_{j=0}^{\nu_{K_p}-1} \frac{\hat{g}^j(K_p)}{\sqrt{j!}r_{\text{low}}^{j+1}(K_p)}. \tag{3.6}$$

This result is a direct application of Theorem 3 from [12]. Theorems 1 and 3 imply

Corollary 1 *Let conditions* (3.1) *and* (3.3) *hold. Then*

$$r_s^{(p)}(\Lambda) \le \frac{1}{2}\hat{\gamma}(K_p)\|Z_{p1}\| + \left(\frac{1}{4}\hat{\gamma}^2(K_p)\|Z_{p1}\|^2 + \hat{\gamma}(K_p)\|Z_{p2}\|\right)^{1/2}.$$

Recall that $g(A) = 0$ if A is normal. In particular cases, inequality (3.5) can be improved. Namely, the following two lemma are valid.

Lemma 2 *Under condition* (3.1), *let one of the following conditions hold: either operators* A_{13} *and* A_{2p} *are normal, or operators* A_{1p} *and* A_{23} $(p = 1, 2)$ *are normal. Then* $\nu_{K_p} \le n_1 + n_2 - 1$.

This result is due to Lemma 3.5 from [12].

Lemma 3 *Let condition* (3.1) *hold. Assume, in addition, that* A_{13} *and* A_{1p} $(p = 1, 2)$ *are normal. Then* $\nu_{K_p} \le n_2$.

Similarly, if under condition (3.1) A_{2p} *and* A_{23} *are normal, then* $\nu_{K_p} \le n_1$.

This result is due to Lemma 3.6 from [12].

4 Bounds Via Determinants

In this section we suggest a bound for the spectral radius via the corresponding determinant. In appropriate situations it can be more convenient than Theorem 1.

We need the following simple result: let A be an invertible $n \times n$-matrix. Then

$$\|A^{-1} \det A\| \le \frac{\|A\|_F^{n-1}}{(n-1)^{(n-1)/2}} \tag{4.1}$$

and

$$\|A^{-1} \det A\| \le \|A\|^{n-1}. \tag{4.2}$$

For the details see Corollary 3.2 from [11].

From (4.1) it follows

$$\|K_p^{-1}\| \le \theta_F(K_p) \tag{4.3}$$

where

$$\theta_F(K_p) := \frac{\|K_p\|_F^{n_1 n_2 - 1}}{(n_1 n_2 - 1)^{(n_1 n_2 - 1)/2} |\det K_p|}.$$

Note that

$$\|K_1\|_F \le \|A_{13}\|_F \|A_{21}\|_F + \|A_{11}\|_F \|A_{23}\|_F.$$

Similarly $\|K_2\|_F$ can be estimated. Moreover, (4.2) implies

$$\|K_p^{-1}\| \le \theta_2(K_p), \tag{4.4}$$

where

$$\theta_2(K_p) := \frac{\|K_p\|^{n_1 n_2 - 1}}{|\det K_p|}.$$

Making use inequality (2.9) from (4.3) and (4.4), we get

Lemma 4 *Let* $\det(K_p) \ne 0$ $(p = 1, 2)$. *Then*

$$r_s^{(p)}(\Lambda) \le \frac{1}{2}\theta_F(K_p)\|Z_{p1}\| + \left(\frac{1}{4}\theta_F^2(K_p)\|Z_{p1}\|^2 + \theta_F(K_p)\|Z_{p2}\|\right)^{1/2} \tag{4.5}$$

and

$$r_s^{(p)}(\Lambda) \le \frac{1}{2}\theta_2(K_p)\|Z_{p1}\| + \left(\frac{1}{4}\theta_2^2(K_p)\|Z_{p1}\|^2 + \theta_2(K_p)\|Z_{p2}\|\right)^{1/2}. \tag{4.6}$$

This lemma enables us to consider perturbations of problem (1.1), (1.2).

5 Perturbation of Problem (1.1), (1.2)

Together with problem Λ defined by (1.1), (1.2) consider the problem

$$\tilde{T}_1 x = (\tilde{\lambda}_1 \tilde{A}_{11} + \tilde{\lambda}_2 \tilde{A}_{12} + \tilde{\lambda}_1 \tilde{\lambda}_2 A_{13})\tilde{v}_1, \tag{5.1}$$

$$\tilde{T}_2 y = (\tilde{\lambda}_1 A_{21} + \tilde{\lambda}_2 A_{22} + \tilde{\lambda}_1 \tilde{\lambda}_2 A_{23})\tilde{v}_2, \tag{5.2}$$

where $\tilde{\lambda}_1, \tilde{\lambda}_2 \in \mathbf{C}$; $\tilde{v}_1 \in \mathbf{C}^{n_1}$, $\tilde{v}_2 \in \mathbf{C}^{n_2}$; $\tilde{T}_1, \tilde{A}_{1j} \in \mathbf{C}^{n_1\times n_1}$ and $\tilde{T}_2, \tilde{A}_{2j} \in \mathbf{C}^{n_2\times n_2}$ $(j = 1, 2, 3)$. Put

$$\tilde{K}_p = \tilde{A}_{13} \otimes \tilde{A}_{2p} - \tilde{A}_{1p} \otimes \tilde{A}_{23} \;\; (p = 1, 2),$$

and define $\tilde{Z}_{pk}$ by (1.3), (1.4) replacing T_p, A_{pk} by $\tilde{T}_p$, $\tilde{A}_{pk}$.

Let us estimate $\theta_F(\tilde{K}_p)$ and $\theta_2(\tilde{K}_p)$. To this end recall the following well-known inequality [2, p. 107]:

$$|\det\ A - \det\ \tilde{A}| \le n M_2^{n-1}(A, \tilde{A})\|A - \tilde{A}\| \;\; (A, \tilde{A} \in \mathbf{C}^{n\times n}), \tag{5.3}$$

where $M_2(A, \tilde{A}) := \max\{\|A\|, \|\tilde{A}\|\}$. Consequently,

$$|\det\ \tilde{A}| \ge |\det\ A| - n M_2^{n-1}(A, \tilde{A})\|A - \tilde{A}\|.$$

Hence,

$$|\det\ \tilde{K}_p| \ge |\det\ K_p| - n_1 n_2 M_2^{n_1 n_2 - 1}(K_p, \tilde{K}_p)\|K_p - \tilde{K}_p\| > 0, \tag{5.4}$$

provided

$$|\det\ K_p| > n_1 n_2 M_2^{n_1 n_2 - 1}(K_p, \tilde{K}_p)\|K_p - \tilde{K}_p\|. \tag{5.5}$$

Thus,

$$\theta_2(\tilde{K}_p) = \frac{\|\tilde{K}_p\|^{n_1 n_2 - 1}}{|\det\ \tilde{K}_p|} \le \hat{\theta}_2(\tilde{K}_p, K_p),$$

where

$$\hat{\theta}_2(\tilde{K}_p, K_p) := \frac{\|\tilde{K}_p\|^{n_1 n_2 - 1}}{|\det\ K_p| - n_1 n_2 M_2^{n_1 n_2 - 1}(K_p, \tilde{K}_p)\|K_p - \tilde{K}_p\|}.$$

Now (4.6) yields

Corollary 2 *Let condition* (5.5) *hold. Then*

$$r_s^{(p)}(\tilde{\Lambda}) \le \frac{1}{2}\hat{\theta}_2(\tilde{K}, K_p)\|\tilde{Z}_{p1}\| + \left(\frac{1}{4}\hat{\theta}_2^2(\tilde{K}_p, K_p)\|\tilde{Z}_{p1}\|^2 + \hat{\theta}_2(\tilde{K}_p, K_p)\|\tilde{Z}_{p2}\|\right)^{1/2}.$$

This corollary gives us a bound for the spectral radius of the perturbed problem, provided the spectral norm of $\tilde{K}_p - K_p$ is sufficiently small with respect to $|\det\ K_p|$. It can be more convenient for applications to the perturbed problem, than Theorem 1, if we know $r_s^{(p)}(\Lambda)$.

The spectral norm is unitarily invariant, but often it is not easy to compute that norm. To get the perturbation result in the the Frobenius norm we use the inequality

$$|\det\ A - \det\ \tilde{A}| \le \Delta_n(A, \tilde{A}) \quad (A, \tilde{A} \in \mathbf{C}^{n\times n}),$$

where

$$\Delta_n(A, \tilde{A}) := \frac{n^n}{2^{n-1} n^{n/2}(n-1)^{n-1}}\|A - \tilde{A}\|_F\ (\|A - \tilde{A}\|_F + \|A + \tilde{A}\|_F)^{n-1},$$

cf. Corollary 3.4 from [11]. So we have

$$|\det\ K_p - \det\ \tilde{K}_p| \le \Delta_{n_1 n_2}(K_p, \tilde{K}_p) > 0,$$

and therefore

$$|\det\ K_p - \det\ \tilde{K}_p| \ge |\det\ K_p| - \Delta_{n_1 n_2}(K_p, \tilde{K}_p) > 0,$$

provided

$$|\det\ K_p| > \Delta_{n_1 n_2}(K_p, \tilde{K}_p). \tag{5.6}$$

Thus,

$$\theta_F(\tilde{K}_p) = \frac{\|\tilde{K}_p\|_F^{n_1 n_2 - 1}}{|\det\ \tilde{K}_p|} \le \hat{\theta}_F(\tilde{K}_p, K_p),$$

where

$$\hat{\theta}_F(\tilde{K}_p, K_p) := \frac{\|\tilde{K}_p\|_F^{n_1n_2-1}}{|\det\ K_p| - \Delta_{n_1n_2}(K_p, \tilde{K}_p)}.$$

Now (4.5) yields

Corollary 3 *Let condition* (5.6) *hold. Then*

$$r_s^{(p)}(\tilde{\Lambda}) \le \frac{1}{2}\hat{\theta}_F(\tilde{K}_p, K_p)\|\tilde{Z}_{p1}\| + \left(\frac{1}{4}\hat{\theta}_F^2(\tilde{K}_p, K_p)\|\tilde{Z}_{p1}\|^2 + \hat{\theta}_F(\tilde{K}_p, K_p)\|\tilde{Z}_{p2}\|\right)^{1/2}.$$

This corollary, as the previous one, provides a bound for $r_s^{(p)}(\tilde{\Lambda})$ of the perturbed problem, provided the Frobenius norm of $\tilde{K}_p - K_p$ is sufficiently small with respect to the absolute value of the determinant of K_p. It also can be more convenient for application to the perturbed problem than Theorem 1, if we know $r_s^{(p)}(\Lambda)$.

6 Gerschgorin Type Bounds for Spectra

In this section we derive a simple lower bound for $r_{\text{low}}(K_l)$ and an upper bound for $r_s^{(p)}(\Lambda)$ in the case of diagonally dominant matrices. Our reasonings in this section are similar to the ones from [20, 23] (see also the references given therein).

Let $\{\omega_{lk}\}_{k=1}^{n_l}$ $(l = 1, 2)$ be an orthonormal basis in $\mathbf{C}^{n_l}$ and $H = \mathbf{C}^{n_1} \otimes \mathbf{C}^{n_2}$. Put $d_{jk} = \omega_{1j} \otimes \omega_{2k}$. So

$$x = \sum_{1\le j\le n_1, 1\le k\le n_2} x_{jk}d_{jk} \quad (x \in H)$$

with $x_{jk} = (x, d_{jk})_H$. Let M be a bilinear operator in H defined by

$$Md_{jk} = \sum_{t,s} m_{tjsk}d_{ts}. \tag{6.1}$$

Then

$$Mx = \sum_{t,s}\sum_{j,k} m_{tjsk}x_{jk}d_{ts}.$$

Application of the Gerschgorin circle theorem [15] in the considered case implies the following result.

Lemma 5 *All the eigenvalues of operator M defined by* (4.1) *lie in the union of the sets*

$$\{z \in \mathbf{C} : |m_{ttss} - z| \le \sum_{j\ne t, k\ne s} |m_{tjsk}| \;\; (1 \le j, t \le n_1, 1 \le k, s \le n_2)\}.$$

We say that M is diagonally-dominant, if the inequality

$$|m_{ttss}| > \sum_{j\ne t, k\ne s} |m_{tjsk}|$$

holds for all $t = 1, \ldots, n_1$; $s = 1, \ldots, n_2$.

Now let $c^{(p)}_{tjsk}$ $(1 \le t, j \le n_1, 1 \le s, k \le n_2;\; p = 1, 2)$ be the entries of K_p in an orthonormal basis. Assume that K_p is diagonally-dominant. Then due to the previous lemma we can write

$$r_{\text{low}}(K_p) \ge \min_{1\le t\le n_1, 1\le s\le n_2} (|c^{(p)}_{ttss}| - \sum_{j\ne t, k\ne s} |c^{(p)}_{tjsk}|).$$

Now let us estimate $r_s^{(p)}(\Lambda)$. To this end we need the following result.

Lemma 6 *Let M be defined by* (6.1)*, and P and Q be bilinear operator defined on H by*

$$Pd_{jk} = \sum_{t=1}^{n_1}\sum_{s=1}^{n_2} p_{tjsk} d_{ts}$$

and

$$Qd_{jk} = \sum_{t=1}^{n_1}\sum_{s=1}^{n_2} q_{tjsk} d_{ts}.$$

Then all the characteristic values of the pencil $M - zP - z^2Q$ lie in the union of the sets

$$\{z \in \mathbf{C} : |m_{ttss} - zp_{ttss} - z^2 q_{ttss}| \le \sum_{j\ne t, k\ne s} |m_{tjsk} - zp_{tjsk} - z^2 q_{tjsk}|\}$$

$$(1 \le t \le n_1, 1 \le s \le n_2).$$

Proof Let x be an eigenvector of the pencil $M - zP - z^2Q$ corresponding to a characteristic value λ: $Mx - \lambda Px - \lambda^2 Qx = 0$ $(x = (x_{jk}) \in H)$. We can write

$$\sum_{1\le j\le n_1, 1\le k\le n_2} (m_{tjsk} - \lambda p_{tjsk} - \lambda^2 q_{tjsk})x_{jk} = 0 \;\; (1 \le t \le n_1, 1 \le s \le n_2). \tag{6.2}$$

Hence

$$|m_{ttss} - \lambda p_{ttss} - \lambda^2 q_{ttss}||x_{ts}| \le \sum_{j\ne t, k\ne s} |m_{tjsk} - \lambda p_{tjsk} - \lambda^2 q_{ttsk}||x_{jk}|.$$

Deleting by the largest $|x_{ts}|$ we have

$$\begin{aligned} &|m_{ttss} - \lambda p_{ttss} - \lambda^2 q_{ttss}| \\ &\quad\le \sum_{j\ne t, k\ne s} |m_{tjsk} - \lambda p_{tjsk} - \lambda^2 q_{tjsk}| \;\; (1 \le t \le n_1, 1 \le s \le n_2), \end{aligned} \tag{6.3}$$

as claimed. Q.E.D.

Assume that

$$\xi_{ts}(Q) := \min_{t,s}(|q_{ttss}| - \sum_{j\ne t, k\ne s} |q_{tjsk}|) > 0. \tag{6.4}$$

That is Q is diagonally dominant. From (6.3), for any characteristic value λ of the pencil $Mx - \lambda Px - \lambda^2 Q$ with $r = |\lambda|$ we have

$$r^2\xi_{ts}(Q) \le r\chi_{ts}(P) + \chi_{ts}(M) \;\; (1 \le t \le n_1, 1 \le s \le n_2),$$

where

$$\chi_{ts}(M) := \sum_{1\le j\le n_1, 1\le k\le n_2} |m_{tjsk}|, \; \chi_{ts}(P) := \sum_{1\le j\le n_1, 1\le k\le n_2} |p_{tjsk}|$$

and therefore,

$$r \le \max_{t,s} \frac{\chi_{t,s}(P)}{2\xi_{t,s}(Q)} + \left(\frac{\chi^2_{t,s}(P)}{4\xi^2_{t,s}(Q)} + \frac{\chi_{t,s}(M)}{\xi_{t,s}(Q)}\right)^{1/2} \tag{6.5}$$

provided Q is diagonally-dominant. Assume that K_p is diagonally-dominant. Then due to (6.5) we can write

$$r_s^{(p)}(\varLambda) \le \max_{t,s} \frac{\chi_{t,s}(Z_{p1})}{2\xi_{t,s}(K_p)} + \left(\frac{\chi^2_{t,s}(Z_{p1})}{4\xi^2_{t,s}(K_p)} + \frac{\chi_{t,s}(Z_{p2})}{\xi_{t,s}(K_p)}\right)^{1/2} \;\; (p = 1, 2). \tag{6.6}$$

7 Sharpness of Theorem 1

Recall that problems (1.1), (1.2) and (2.8) are equivalent. For some $p = 1, 2$, let K_p,Z_{p1} and Z_{p2} be mutually commuting Hermitian matrices. Then (2.8) implies

$$\lambda_{pk}^2(\Lambda)\mu_k(K_p) + \lambda_{pk}(\Lambda)\mu_k(Z_{p1}) + \mu_k(Z_{p2}) = 0 \quad (k = 1, \dots, n_1 n_2),$$

where $\mu_k(K_p)$ and $\mu_k(Z_{pj})$ are the eigenvalues of K_p, and Z_{pj}, respectively; $\lambda_{pk}(\Lambda)$ are the eigenvalues of problem Λ corresponding to $\mu_k(K_p)$ and $\mu_k(Z_{pj})$. Then the spectrum of Λ consists of

$$\lambda_{pk+}(\Lambda) = -\frac{\mu_k(Z_{p1})}{2\mu_k(K_p)} + \left(\frac{\mu_k^2(Z_{p1})}{4\mu_k^2(K_p)} + \frac{\mu_k(Z_{p2})}{\mu_k(K_p)}\right)^{1/2}$$

and

$$\lambda_{pk-}(\Lambda) = -\frac{\mu_k(Z_{p1})}{2\mu_k(K_p)} - \left(\frac{\mu_k^2(Z_{p1})}{4\mu_k^2(K_p)} + \frac{\mu_k(Z_{p2})}{\mu_k(K_p)}\right)^{1/2}.$$

Assume that K_p and Z_{p2} are positive definite, and Z_{p1} is negative definite. In addition, for some index j, let

$$\mu_j(K_p) = \min_k \mu_k(K_p) = r_{\text{low}}(K_p) \ (> 0), |\mu_j(Z_{p1})| = \max_k |\mu_k(Z_{p1})| = \|Z_{p1}\|,$$

and $\mu_j(Z_{p2}) = \max_k \mu_k(Z_{p2}) = \|Z_{p2}\|$. Then

$$r_s^{(p)}(\Lambda) = \frac{\|Z_{p1}\|}{2r_{\text{low}}(K_p)} + \left(\frac{\|Z_{p1}\|^2}{4r_{\text{low}}^2(K_p)} + \frac{\|Z_{p2}\|}{r_{\text{low}}(K_p)}\right)^{1/2}.$$

So in the considered case inequality (1.6) is attained and therefore, *Theorem 1 is sharp*.

Example 1 Let $n_1 = n_2 = 2$,

$$A_{13} = \begin{pmatrix} 9 & 0.5 \\ 0.5 & 9 \end{pmatrix}, A_{23} = \begin{pmatrix} 20 & 1 \\ 1 & 20 \end{pmatrix}, \quad A_{11} = A_{22} = A_{21} = T_2 = T_1 = I.$$

So

$$\lambda_{1,2}(A_{13}) = 9 \pm 0.5, \lambda_{1,2}(A_{23}) = 20 \pm 1.$$

We have $K_1 = A_{13} \otimes I - I \otimes A_{23}$ and therefore

$$\lambda_{jk}(K_1) = \lambda_j(A_{13}) - \lambda_k(A_{23}) \ \ (j, k = 1, 2).$$

Consequently, $r_{low}(K_1) = 19 - 9.5 = 9.5$. Since the matrices are selfadjoint, we have $g_0(K_1) = 0$. In addition,

$$Z_{11} = I \otimes A_{23} - A_{13} \otimes I - I \otimes I + I \otimes I = I \otimes A_{23} - A_{13} \otimes I$$

and $Z_{12} = 0$ Besides,

$$\|Z_{11}\| = \max_{jk} |\lambda_j(A_{13}) - \lambda_k(A_{23})| = 12.5.$$

Thus Theorem 1 implies,

$$r_s^{(1)}(\Lambda) \leq \|Z_{11}\|/r_{low} \leq 12.5/9.5 \approx 1.315.$$

According to (2.6) the direct calculations show that $r_s^{(1)}(\Lambda) = 1$.

8 Conclusion

The present paper has been inspired by the paper [17]. The spectral radius is a maximal absolute values of the eigenvalues and thus it gives us the radius of the disc in which all the eigenvalues of the considered problem lie.

In Sects. 1 and 2 we have derived a bound for the spectral radius in the general case. In Sect. 3 the results of Sects. 1 and 2 have been improved in the cases when some matrices of the problem have a joint Schur basis. To investigate perturbations of problem (1.1), (1.2) in Sect. 4 we suggest additional bounds for the spectral radius by virtue of determinants and apply them in Sect. 5 to derive a bound for the spectral radius of the perturbed problem.

In Sect. 6 we investigate the problem with diagonally dominant matrices. In that case by the Geschgorin approach we obtained a simple bound for the spectral radius. In Sect. 7 it was shown that Theorem 1 is sharp, namely inequality (1.6) is attained under the conditions pointed in that section. In addition, Sect. 7 contains an illustrative example.

Recently Bindel and Hood [4, 5] have proved localization theorems for the pseudospectra of nonlinear eigenvalue problems. Besides, they generalized the Gershgorin and Bauer-Fike theorems. Our Sect. 6 is connected with the results of Bindel and Hood but we investigate the spectrum, not the pseudospectrum.

References

1. F.V. Atkinson, *Multiparameter Eigenvalue Problems* (Academic, New York, 1972)
2. R. Bhatia, *Perturbation Bounds for Matrix Eigenvalues*. Classics in Applied Mathematics, vol. 53 (SIAM, Philadelphia, 2007)
3. R. Bhatia, M. Uchiyama, The operator equation $\sum_{i=1}^{n} A^{n-i} X B^{i} = Y$. Expo. Math. **27**, 251–255 (2009)
4. D. Bindel, A. Hood, Localization theorems for nonlinear eigenvalue problems. SIAM J. Matrix Anal. Appl. **34**(4), 1728–1749 (2013)
5. D. Bindel, A. Hood, Localization theorems for nonlinear eigenvalue problems. SIAM Rev. **57**(4), 585–607 (2015)
6. P. Binding, P.J. Browne, A variational approach to multiparameter eigenvalue problems for matrices. SIAM J. Math. Anal. **8**, 763–777 (1977)
7. P. Binding, P.J. Browne, A variational approach to multiparameter eigenvalue problems in Hilbert space. SIAM J. Math. Anal. **9**, 1054–1067 (1978)
8. N. Cottin, Dynamic model updating a multiparameter eigenvalue problem. Mech. Syst. Signal Process. **15**, 649–665 (2001)
9. M. Faierman, *Two-Parameter Eigenvalue Problems in Ordinary Differential Equations*. Pitman Research Notes in Mathematics Series, vol. 205 (Longman Scientific and Technical, Harlow, 1991)
10. M.I. Gil', *Operator Functions and Localization of Spectra*. Lecture Notes in Mathematics, vol. 1830 (Springer, Berlin, 2003)
11. M.I. Gil', On spectral variation of two-parameter matrix eigenvalue problem. Publ. Math. Debrecen **87**(3–4), 269–278 (2015)
12. M.I. Gil', Bounds for the spectrum of a two parameter matrix eigenvalue problem. Linear Algebra Appl. **498**, 201–218 (2016)
13. M.E. Hochstenbach, B. Plestenjak, Harmonic Rayleigh-Ritz extraction for the multiparameter eigenvalue problem. Electron. Trans. Numer. Anal. **29**, 81–96 (2007/2008)
14. M.E. Hochstenbach, A. Muhič, B. Plestenjak, On linearizations of the quadratic two-parameter eigenvalue problem. Linear Algebra Appl. **436**(8), 2725–2743 (2012)
15. R.A. Horn, C.R. Johnson, *Topics in Matrix Analysis* (Cambridge University Press, Cambridge, 1991)
16. G.A. Isaev, Lectures on Multiparameter Spectral Theory, Dept. of Mathematics and Statistics. Univ. of Calgary (1985)
17. E. Jarlebring, M.E. Hochstenbach, Polynomial two-parameter eigenvalue problems and matrix pencil methods for stability of delay-differential equations. Algebra Appl. **431**, 369–380 (2009)
18. V.B. Khazanov, To solving spectral problems for multiparameter polynomial matrices. J. Math. Sci. **141**, 1690–1700 (2007)
19. T. Košir, Finite dimensional multiparameter spectral theory: the nonderogatory case. Algebra Appl. **212/213**, 45–70 (1994)
20. C.Q. Li, Y.T. Li, X. Kong, New eigenvalue inclusion sets for tensors. Numer. Linear Algebra Appl. **21**, 39–50 (2014)
21. M. Marcus, H. Minc, *A Survey of Matrix Theory and Matrix Inequalities* (Allyn and Bacon, Boston, 1964)
22. A. Muhič, B. Plestenjak, On the quadratic two-parameter eigenvalue problem and its linearization. Linear Algebra Appl. **432**, 2529–2542 (2010)
23. L.Q. Qi, Eigenvalues of a real supersymmetric tensor. J. Symb. Comput. **40**, 1302–1324 (2005)
24. H. Volkmer, On the minimal eigenvalue of a positive definite operator determinant. Proc. R. Soc. Edinb. Sect. A **103**, 201–208 (1986)

On the Properties of a Nonlocal Nonlinear Schrödinger Model and Its Soliton Solutions

Theodoros P. Horikis and Dimitrios J. Frantzeskakis

1 Introduction

Many physically different subjects can be brought together through their common modeling and mathematical description. Perhaps the most common (and rather unlike) example is water waves and nonlinear optics. Two systems are inextricably linked with both subjects: the universal Korteweg-de Vries (KdV) and nonlinear Schrödinger (NLS) equations [1]. Remarkable as these systems may be, for several physically relevant contexts their standard form turns out to be an oversimplified description as it cannot model, for example, higher dimensionality; for instance, the Kadomtsev-Petviashvilli (KP) equation is used as a generalization of the KdV to two spatial dimensions. Furthermore, these systems can be reduced from one to the other [74]. Importantly, all the above models are completely integrable by means of the Inverse Scattering Transform (IST) [2] and support soliton solutions, namely robust localized waves that have always been a central element in numerous studies in physics [17], applied mathematics [2] and engineering [24]. A unique property of solitons is that they feature a particle-like character, which enables them to interact elastically, preserving their shapes and velocities after colliding with each other.

In general, the theory of nonlinear waves, is governed by problems where several different temporal and/or spatial scales are present. Thus, asymptotic multiscale expansion methods are usually applied to derive nonlinear evolution equations more manageable to the problem at hand [30]. These asymptotic methods are used to establish connections between different systems, as mentioned above, which allows

T. P. Horikis (✉)
Department of Mathematics, University of Ioannina, Ioannina, Greece
e-mail: horikis@uoi.gr

D. J. Frantzeskakis
Department of Physics, National and Kapodistrian University of Athens, Athens, Greece
e-mail: dfrantz@phys.uoa.gr

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_14

for the construction of approximate solutions of the original models by means of the exact solutions of the reduced models. A prominent example is the connection of the defocusing nonlinear Schrödinger (NLS) equation with the KdV equation, which allowed for the description of shallow dark solitons of the former in terms of KdV solitons. Relevant studies started in the early 1970s [71] and continue to date [27]; importantly, they have also been extended to the case of nonintegrable systems, providing extremely useful information on the existence, stability and dynamics of solutions in various physical settings, such as nonlinear optics [40] and Bose-Einstein condensates [22, 37].

Multiscale expansion methods become even more useful when the original system is not integrable, without known solutions in explicit form. For several physically relevant contexts the standard NLS equation turns out to be an oversimplified description as it cannot model, for example, loss and gain which are inevitable in any real system [4]. Hence, in order to model important classes of physical systems in a relevant way, it is necessary to go beyond the standard NLS description. For instance, nematic liquid crystals [9, 16], atomic vapors [69] and other thermal nonlinear optical media [45, 65], as well as plasmas [48, 73] and dipolar bosonic quantum gases [60], constitute a class of systems that display nonlocal nonlinear response. The effect of the nonlocality on the NLS equation is rather profound. The integrable nature of the equation is generally lost and while soliton solutions may also be found, they lack the freedom of various parameters describing the soliton's properties (amplitude, velocity, etc). Nevertheless, nonlocal nonlinear systems may feature quite interesting properties. In particular, in the case of focusing nonlocal nonlinearities, collapse is arrested in higher-dimensions [45, 72], which results in stable solitons, as observed in experiments of, say [65, 69], even in the $(3 + 1)$-dimensional setting [55]—see, e.g., reviews [45, 54] and references therein. On the other hand, in the case of defocusing nonlocal nonlinearities, dark solitons that are supported in such settings [19, 26, 36, 63], may exhibit an attractive interaction [19], rather than a repulsive one, as is the situation in the case of a local nonlinearity—cf. the reviews [22, 37, 40] and references therein. Furthermore, dark solitons which are known to be prone to the transverse (or "snaking") instability in higher-dimensional settings [37, 41, 47, 62], can be stabilized due to the presence of the nonlocal nonlinearity [8].

Here, motivated by the above—and particularly by the fact that nonlocal nonlinearities have a profound effect on the form and stability properties of nonlinear excitations—we study a physically relevant nonlocal NLS model. Different versions of this model are studied: a $(1 + 1)$-dimensional scalar system, its vector generalization, as well as a scalar, fully $(3 + 1)$-dimensional model, in both the focusing and the defocusing regimes. We first present the simplest nontrivial solution, namely the continuous-wave (cw), and study its stability. We show that in the focusing (defocusing) version of the model the cw is modulationally unstable (stable). Then, we study soliton solutions of the model and present bright soliton solutions in a closed analytical form, for both the scalar and the vector versions of the nonlocal system in the focusing regime. In the defocusing regime, we use multiscale expansion methods to derive effective nonlinear evolution equations that describe a variety of approximate soliton solutions of the original model.

For the $(1 + 1)$-dimensional scalar model, we derive an effective KdV equation, which describes either dark or antidark solitons of the nonlocal NLS. In the vector version of the model, our analysis is performed for different boundary conditions: for nonvanishing conditions for both fields, and nonvanishing-vanishing conditions for each field. In the former case, we derive a KdV model, which describes dark-dark solitons solutions of the original problem, while in the second case, we derive a Mel'nikov system—namely a KdV equation with a self-consistent source satisfying a time-independent Schrödinger equation; this system describes dark-antidark solitons of the original nonlocal NLS.

For the fully $(3 + 1)$-dimensional scalar version of the nonlocal NLS model in the defocusing regime, we use again multiple scale expansion techniques, both in Cartesian and cylindrical geometries. This way, first we derive, at an intermediate stage of the asymptotic analysis, a 3D Boussinesq equation. Then, we consider cases corresponding to spatial or temporal structures and, upon introducing relevant scales and asymptotic expansions, we reduce the Boussinesq model to KP-type equations for right- and left-propagating waves. These models include various integrable and non-integrable equations at different dimensions and geometries, such as the KdV and the cKdV equation, the KP-I and KP-II equations, Johnson's equation, as well as the CI and CII equations. We thus predict the existence and stability of various solitary waves, namely spatial—planar or cylindrical (ring-shaped)—and dark or anti-dark (for weak or strong nonlocality, respectively). Furthermore, our analysis suggests the existence of temporal solitary waves, which become unstable in higher dimensions. Regarding approximate two-dimensional solitary wave solutions, it is found that they may exist in the form of weakly localized (i.e., algebraically decaying) dark "lumps", which satisfy effective KP-I models; such structures may be either of the spatial or temporal type and are supported in the weak nonlocality regime.

Moreover, we use direct numerical simulations to confirm our analytical findings. Indeed, our predictions are found to be in very good agreement with the numerical results: using the analytical forms of the spatial soliton solutions as initial conditions for the direct numerical integration of the original problem(s), we find that the solitons propagate undistorted. This holds at least for relatively small and intermediate propagation distances, while, even for longer propagation distances, instabilities are not observed in our simulations. This suggests that the solitons found have a good chance to be observed in experiments.

2 The 1D Scalar Nonlocal System

2.1 Modulation Instability and Bright Solitons

We consider the following system, expressed in normalized form, that governs propagation in nonlocal media [9, 39]:

$$i\frac{\partial u}{\partial z}+d\frac{\partial^2 u}{\partial x^2}+2g\theta u=0, \tag{1a}$$

$$\nu\frac{\partial^2\theta}{\partial x^2}-2q\theta=-2|u|^2. \tag{1b}$$

Depending on the physical situation, the system and its coefficients correspond to different physical quantities. For example, in the context of nematic liquid crystals, u is the complex valued, slowly-varying envelope of the optical electric field and θ is the optically induced deviation of the director angle. Note that u and θ depend on the propagation distance z (which is the evolution variable in this setting) and transverse coordinate x; thus solitons of system (1) are so-called *spatial solitons*, i.e., nondiffracting beams [39]. Furthermore, in Eq. (1), diffraction is represented by d and nonlinear coupling by g. The effect of nonlocality ν measures the strength of the response of the nematic in space, with a highly nonlocal response when $\nu(>0)$ is large. The parameter $q>0$ is related to the square of the applied static field which pre-tilts the nematic dielectric [59]. In this context, d, g, q are $O(1)$ while ν is large, i.e., $\nu=O(10^2)$ [9, 59].

In order to investigate the stability properties of system (1) consider its simplest nontrivial solution, i.e., a continuous-wave (cw) of the form:

$$u(z)=u_0e^{2ig\theta_0 z},\quad \theta_0=\frac{1}{q}u_0^2,$$

where u_0 is a real constant. To investigate if this cw solution is subject to modulational instability (MI), consider a small perturbation $u_1(x,z)$ (with $\max|u_1|\ll u_0$). Then, substituting the ansatz:

$$u(x,z)=[u_0+u_1(x,z)]e^{2ig\theta_0 z},$$

into Eq. (1) and linearizing with respect to u_1, which is assumed to behave as $\exp[i(kx-\omega z)]$, we obtain the following dispersion relation for the frequency ω and wavenumber k of the perturbation:

$$\omega^2=\frac{dk^2\left(d\nu k^4+2dqk^2-8gu_0^2\right)}{\nu k^2+2q}. \tag{2}$$

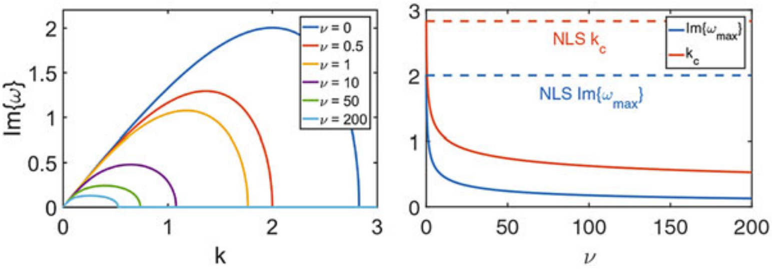


Fig. 1 Left: Growth rates for different values of the nonlocal parameter ν. Right: The change of critical values $\mathrm{Im}\{\omega_{\max}\}$ and k_c with the nonlocality ν

It is clear that in the self-focusing case with $dg > 0$ the system is unstable, whereas in the self-defocusing case with $dg < 0$ the system is stable (i.e., $\omega \in \mathbb{R}\forall k$). Also, in the local case (i.e., for $\nu = 0$) the equation reduces to the dispersion relation of the respective NLS equation with local cubic nonlinearity, which features the same stability criteria (see, e.g., Ref. [39]). From this dispersion relation we can identify three critical values that characterize the instability, namely the maximum growth rate, $\mathrm{Im}\{\omega_{\max}\}$, its location $k_{\max}$, and the width of the instability region (alias "instability band"), k_c. The value $\mathrm{Im}\{\omega_{\max}\}$ is a measure of the propagation distance needed for the instability to occur (the larger its value the faster the instability occurs) and k_c, defines the range of possible wavenumbers that can yield instability; the larger its value, the more unstable the system is, as more wave numbers can lead to unstable propagation. By differentiating (2), with respect to k, we find that $k_{\max}$ is the solution of the algebraic equation:

$$d\left(\nu k^3 + 2qk\right)^2 - 8gqu_0^2 = 0,$$

while k_c satisfies

$$d\nu k^4 + 2dqk^2 - 8gu_0^2 = 0.$$

Both equations can be solved in closed form (they are bi-quadratics) to give the relative dependance of $\mathrm{Im}\{\omega_{\max}\}$ and k_c with the nonlocality ν. We illustrate this in Fig. 1. Hereafter, we fix $d = 1/2$ and $g = q = u_0 = 1$.

This figure agrees with the findings of Ref. [44]: nonlocality has an increasingly stabilizing effect on the system. Indeed, both critical values that characterize the instability, $\mathrm{Im}\{\omega_{\max}\}$ and k_c, decrease as ν increases. This means that the effect of MI will need more distance to be exhibited; and if ν is large enough, this distance can be larger than the experimentally relevant scales. Further, a smaller range of wavenumbers will cause an instability. Notice, again, that while both values decrease, the effect, in the focusing case, is always present, just more suppressed as ν increases. The limiting NLS system is, by these values, significantly more unstable.

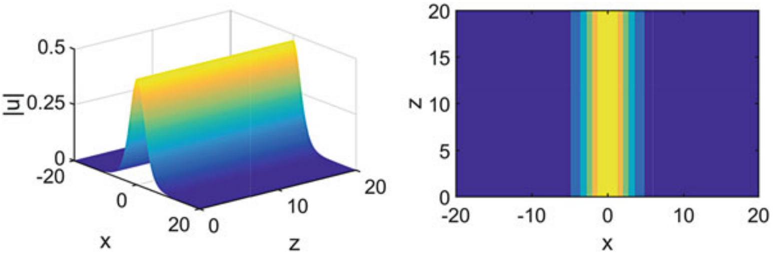
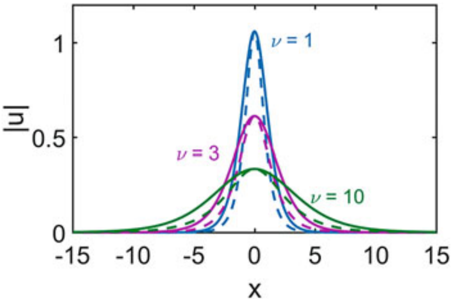


Fig. 2 The evolution of the single soliton of Eq. (1) with $\nu = 10$

Fig. 3 Typical nonlocal solitons for various values of ν. The dashed curves correspond to the relative NLS soliton of the same amplitude



It is now straightforward to find the soliton solution of Eq. (1), namely [49]:

$$u(x,z) = \frac{3q}{2}\sqrt{\frac{d}{g\nu}}\operatorname{sech}^2(\sqrt{q/2\nu}x)e^{2idq/\nu z}.$$

This soliton solution, while it obviously depends on ν, it has fixed amplitude (much like $\chi^{(2)}$ materials [13, 34]) which decays with ν. Notice that the relative ($\nu = 0$) single NLS soliton solution reads:

$$u_s(x,z) = u_0\operatorname{sech}(u_0\sqrt{g/dq}x)e^{iu_0^2gz/q}.$$

The immediate comparison reveals that the nonlocal solitons luck the freedom of various parameters describing the soliton's properties (amplitude, velocity, etc), but maintain the property of undistorted evolution, as shown in Fig. 2.

In fact, solutions with a free parameter for this system have been found but only in the defocusing case and under a small-amplitude approximation technique [26]; these solutions will be presented in the next section.

Finally, to illustrate how the nonlocality affects the corresponding solitons we shown in Fig. 3 below, the soliton solution for different values of ν as well as the relative NLS soliton for the same amplitude.

2.2 Dark and Anti-Dark Solitons

We now turn our attention to the defocusing system by setting $g = -1$ so that Eq. (1) become:

$$i\frac{\partial u}{\partial z} + d\frac{\partial^2 u}{\partial x^2} - 2\theta u = 0, \tag{3a}$$

$$\nu\frac{\partial^2 \theta}{\partial x^2} - 2q\theta = -2|u|^2. \tag{3b}$$

Both functions $u = u(z, x)$ and $\theta = \theta(z, x)$ are assumed to be non-zero at the boundaries (infinities). The MI analysis above indicates that the system is now stable and, as such, it is convenient to introduce the concept of a background function [3] such that the electric field and the angle can be written in the form:

$$u(z, x) = u_b(z)U(z, x), \quad \theta(z, x) = \theta_b(z)v(z, x), \tag{4}$$

where the subscript b denotes the background functions, which are clearly functions of z only. Note that at the boundaries ($|x| \to \infty$),

$$u = u(z) = u_b(z), \quad \theta = \theta(z) = v_b(z).$$

Taking the limit of Eq. (3), we find that these functions satisfy the following set of equations:

$$\left.\begin{array}{r} i\dfrac{du_b}{dz} - 2\theta_b u_b = 0 \\ -2q\theta_b = -2|u_b|^2 \end{array}\right\} \Leftrightarrow \left\{\begin{array}{l} i\dfrac{du_b}{dz} = 2\theta_b u_b \\ \theta_b = \dfrac{|u_b|^2}{q} \end{array}\right.$$

By writing $u_b = u_0 e^{i\varphi}$ one finally obtains the form of the background functions:

$$u_0(z) = u_0, \quad \varphi(z) = -\frac{2u_0^2}{q}z + \varphi(0), \quad \theta_b(z) = \frac{u_0^2}{q},$$

where $u_0 \in \mathbb{R}$ is a constant. Next substitute Eq. (4) into Eq. (3) to obtain the evolution equations for $U(z, x)$ and $v(z, x)$:

$$i\frac{\partial U}{\partial z} + \frac{1}{2}\frac{\partial^2 U}{\partial x^2} - 2\frac{u_0^2}{q}(v - 1)U = 0,$$

$$\nu\frac{\partial^2 v}{\partial x^2} - 2qv = -2q|U|^2.$$

It is trivial to check that these are also satisfied at the boundaries where

$$U = v = 1,$$

and any evolution of the boundary conditions has been absorbed by the background functions. The resulting equations have now fixed boundary conditions. We, then, employ the so-called Madelung transformation $U(z,x) = \rho(z,x)\exp[i\phi(z,x)]$ so that

$$\begin{gathered}
\frac{\partial\rho}{\partial z} + \frac{1}{2}\rho\frac{\partial^2\phi}{\partial x^2} + \frac{\partial\rho}{\partial x}\frac{\partial\phi}{\partial x} = 0, \\
\rho\frac{\partial\phi}{\partial z} + \frac{1}{2}\left[\rho\left(\frac{\partial\phi}{\partial x}\right)^2 - \frac{\partial^2\rho}{\partial x^2}\right] + \frac{2u_0^2}{q}(v-1)\rho = 0, \\
v\frac{\partial^2 v}{\partial x^2} - 2qv = -2q\rho^2,
\end{gathered}$$

since $v(z,x) \in \mathbb{R}$ and define new scales such that

$$Z = \varepsilon^3 z, \quad X = \varepsilon(x - Cz),$$

where C is a constant; this is actually the speed of sound, namely the velocity of small-amplitude and long-wavelength waves propagating along the background. Its value(s) will be determined—in a self-consistent manner—later in the analysis.

Additionally, expand amplitude and phase in powers of ε as follows:

$$\begin{aligned}
\rho &= \rho_0 + \varepsilon^2\rho_2 + \varepsilon^4\rho_4 + \cdots, \\
\phi &= \varepsilon\phi_1 + \varepsilon^3\phi_3 + \varepsilon^5\phi_5\cdots, \\
v &= v_0 + \varepsilon^2 v_2 + \varepsilon^4 v_4 + \cdots.
\end{aligned}$$

Substituting back to Eq. (3) we obtain at different order of ε:

$$\mathscr{O}(1): \begin{array}{c} \rho_0^2 = v_0, \\ v_0 = 1, \end{array}$$

$$\mathscr{O}(\varepsilon^2): \begin{array}{c} \dfrac{\partial\phi_1}{\partial X} = \dfrac{4u_0^2\rho_0}{Cq}\rho_2, \\ v_2 = 2\rho_0\rho_2, \end{array} \tag{5}$$

$$\mathscr{O}(\varepsilon^3): C\frac{\partial\rho_2}{\partial X} = \frac{\rho_0}{2}\frac{\partial^2\phi_1}{\partial X^2}. \tag{6}$$

The compatibility between Eqs. (5) and (6) suggests:

$$C^2 = \frac{2u_0^2\rho_0^2}{q} = \frac{2u_0^2}{q}.$$

The fact that C may have two signs implies that the KdV, cf. Eq. (9) below, may describe waveforms propagating either to the left or to the right. We will return to this with more comments below. Hereafter we will be using $\rho_0 = 1$. Also note that Eq. (5) is a simple equation connecting ϕ_1 and ρ_2 and will be used below to determine the phase ϕ_1.

At the next orders in ε we have:

$$\mathscr{O}(\varepsilon^4): \quad \begin{aligned} &\rho_0\frac{\partial\phi_1}{\partial Z} + \frac{2u_0^2\nu}{q^2}\frac{\partial^2\rho_2}{\partial X^2} - \frac{1}{2}\frac{\partial^2\rho_2}{\partial X^2} - C\frac{\partial\phi_3}{\partial X} + \frac{1}{2}\left(\frac{\partial\phi_1}{\partial X}\right)^2 \\ &-C\rho_2\frac{\partial\phi_1}{\partial X} + \frac{4u_0^2}{q}\rho_4 + \frac{6u_0^2}{q}\rho_2^2 = 0, \\ &v_4 = 2\rho_4 + \rho_2^2 + \frac{\nu}{q}\frac{\partial^2\rho_2}{\partial X^2}, \end{aligned} \tag{7}$$

$$\mathscr{O}(\varepsilon^5): \; \frac{\partial\rho_2}{\partial Z} + \frac{1}{2}\frac{\partial^2\phi_3}{\partial X^2} + \frac{1}{2}\rho_2\frac{\partial^2\phi_1}{\partial X^2} + \frac{\partial\rho_2}{\partial X}\frac{\partial\phi_1}{\partial X} - C\frac{\partial\rho_4}{\partial X} = 0. \tag{8}$$

Equations (7) and (8) are compatible iff:

$$\frac{\partial\rho_2}{\partial Z} + \frac{C(4u_0^2\nu - q^2)}{16u_0^2 q}\frac{\partial^3\rho_2}{\partial X^3} + \frac{6u_0^2}{Cq}\rho_2\frac{\partial\rho_2}{\partial X} = 0. \tag{9}$$

This compatibility condition is found upon differentiating Eq. (7) with respect to X (also using equations of lowest order in ε) and adding with Eq. (8).

Clearly Eq. (9) is a KdV equation and possesses soliton solutions which can be given in explicit form. Indeed, denoting

$$\alpha = \frac{C(4u_0^2\nu - q^2)}{16u_0^2 q}, \quad \beta = \frac{u_0^2}{Cq},$$

the soliton solution of Eq. (9) is given by:

$$\rho_2 = \frac{2\alpha}{\beta}\eta^2\,\mathrm{sech}^2(\eta X - 4\eta^3\alpha Z + X_0),$$

where η is a free parameter. The relative phase can be retrieved from Eq. (5):

$$\phi_1 = \frac{8u_0^2\alpha}{Cq\beta}\eta\tanh(\eta X - 4\eta^3\alpha Z + X_0),$$

while

$$v_2 = 2\rho_0\rho_2 = \frac{4\alpha}{\beta}\eta^2 \operatorname{sech}^2(\eta X - 4\eta^3\alpha Z + X_0).$$

Then, the amplitude of the solution of Eq. (3) up to $\mathscr{O}(\varepsilon^2)$ is

$$\rho = \rho_0 + \varepsilon^2\rho_2,$$

and since $\rho_0 = 1$ the sign of ρ_2 will determine whether the solution is an intensity dip (dark soliton) or an intensity hump (anti-dark soliton) off of a constant background. This is determined by the sign of the quantity

$$\frac{\alpha}{\beta} = \frac{C^2(4u_0^2\nu - q^2)}{16u_0^4}.$$

Then it is evident that when

$$\gamma = \frac{u_0^2\nu}{q^2} < \frac{1}{4},$$

the equation supports dark solitons (termed "dark nematicons" in the context of nematic liquid crystals), while when the inequality is reversed anti-dark humps are exhibited. It is also worth mentioning that these results do not depend on the sign of C. However, the sign of C affects the direction of propagation and the speed of the nematicon. Indeed, notice that the speed is dependent on the parameter α and as such the direction of propagation depends on the sign of C. While the ratio of the coefficient of the linear dispersion to the coefficient of the nonlinear term determines the type of nematicon its speed and direction are determined by the sign of the linear term. The speed of sound only affects the latter.

The resulting solutions obtained by the above method can be now written as:

$$u(z,x) = u_0(z)\left[1 + \varepsilon^2\frac{2\alpha}{\beta}\eta^2 \operatorname{sech}^2(\eta X - 4\eta^3\alpha Z + X_0)\right] e^{i[\theta(z)+\varepsilon\phi_1]},$$

$$\phi_1 = \frac{8u_0^2\alpha}{Cq\beta}\eta \tanh(\eta X - 4\eta^3\alpha Z + X_0),$$

with an additional higher order, $\mathscr{O}(\varepsilon^3)$, correction to the speed. Note that when $\nu = 0$ the amplitude ρ_2 is always negative as $\alpha/\beta = -q/8u_0^2 < 0$ and as such this is always a dip, i.e., a dark soliton and never a hump or anti-dark soliton.

We can summarize these results as follows:

- Dark nematicons are obtained when $\gamma < 1/4$ and propagate to the left if $C > 0$ or to the right if $C < 0$.

- Anti-dark nematicons are obtained when $\gamma > 1/4$ and propagate to the left if $C > 0$ or to the right if $C < 0$.

The critical point when $\gamma = 1/4$ marks the point where the linear dispersion of Eq. (9) changes sign. It has been shown that a soliton decays as it passes this point [50].

3 The 1D Vector Nonlocal System

We now turn our attention on another version of the nonlocal NLS system, namely to its vector counterpart in (1 + 1)-dimensions. In this case, physically speaking, the system is composed by three equations: two of them refer to polarised, coherent light beams of two different wavelengths propagating through a cell filled with a nematic liquid crystal; these equations are coupled to each other, and also coupled with a diffusion-type equation describing the evolution of the director angle (or the refractive index in the context of, e.g., media with thermal nonlinearities). This system is expressed in non-dimensional form as follows [7, 67]:

$$i\frac{\partial u}{\partial z} + \frac{d_1}{2}\frac{\partial^2 u}{\partial x^2} + 2g_1\theta u = 0, \tag{10}$$

$$i\frac{\partial v}{\partial z} + \frac{d_2}{2}\frac{\partial^2 v}{\partial x^2} + 2g_2\theta v = 0, \tag{11}$$

$$\nu\frac{\partial^2 \theta}{\partial x^2} - 2q\theta = -2(g_1|u|^2 + g_2|v|^2). \tag{12}$$

The variables u and v are the complex valued, slowly varying envelopes of the optical electric fields and θ is the optically induced deviation of the director angle, as before. Diffraction is represented by d_1, d_2 and nonlinearity by g_1, g_2. Importantly, these variables are allowed to vary their signs while all other constants are taken positive. When the signs of diffraction and nonlinearity are opposite, i.e., $d_1g_1,\ d_2g_2 < 0$, the system is termed defocusing and focusing otherwise. The location of the relative signs is not important: multiplying Eqs. (10) and (11) by -1 and changing $z \to -z$ moves the sign difference from one place to the other.

In order to investigate the MI of a pair of coupled waves we first consider the continuous-wave (cw) solution of Eqs. (10)–(12), i.e.,

$$u = u_0 e^{2ig_1\theta_0 z}, \quad v = v_0 e^{2ig_2\theta_0 z}, \quad \theta_0 = \frac{g_1u_0^2 + g_2v_0^2}{q},$$

where u_0 and v_0 are real constants. Now consider a small perturbation to this cw solution

$$u(z,x) = [u_0 + u_1(z,x)]e^{2ig_1\theta_0 z}, \quad v(z,x) = [v_0 + v_1(z,x)]e^{2ig_2\theta_0 z}$$

which we insert into the system (10)–(12). In order to simplify the analysis we first solve Eqs. (10) and (11) for θ and substitute in Eq. (12). While this eliminates one dependent variable it raises the overall order of the system and, as such, it only proves useful in the MI analysis where plane wave solutions are investigated (as solutions of a linear system). When solitons or exact solutions are the object of the analysis this is not recommended. The linearized equations (where terms of order u_1^2 and v_1^2 or higher have been neglected) for the small perturbing terms are found to be:

$$4iq\frac{\partial u_1}{\partial z} - 2i\nu\frac{\partial^3 u_1}{\partial z\partial x^2} - d_1\nu\frac{\partial^4 u_1}{\partial x^4} + 2d_1 q\frac{\partial^2 u_1}{\partial x^2} + 8g_1^2 u_0^2(u_1 + u_1^*) + 8g_1 g_2 u_0 v_0(v_1 + v_1^*) = 0, \tag{13}$$

$$4iq\frac{\partial v_1}{\partial z} - 2i\nu\frac{\partial^3 v_1}{\partial z\partial x^2} - d_2\nu\frac{\partial^4 v_1}{\partial x^4} + 2d_2 q\frac{\partial^2 v_1}{\partial x^2} + 8g_2^2 v_0^2(v_1 + v_1^*) + 8g_1 g_2 u_0 v_0(u_1 + u_1^*) = 0. \tag{14}$$

Notice that the terms involving ν, that induce the nonlocality, are higher order derivatives; without these terms the problem simply reduces to the linearized NLS problem. This means that their contribution is expected to be highly nontrivial, as they produce higher order polynomials in the dispersion relation; the effect of these terms will become more prominent below. Equations (13) and (14) admit solutions of the form

$$u_1(z,x) = c_1 e^{i(kx-\omega z)} + c_2 e^{-i(kx-\omega z)}, \quad v_1(z,x) = c_3 e^{i(kx-\omega z)} + c_4 e^{-i(kx-\omega z)}, \tag{15}$$

provided the dispersion relationship

$$p_1(k)\omega^4 + p_2(k)\omega^2 + p_3(k) = 0, \tag{16}$$

with

$$\begin{aligned} p_1(k) &= 16\left(k^2\nu + 2q\right), \\ p_2(k) &= -4\nu\left(d_1^2 + d_2^2\right)k^6 - 8q\left(d_1^2 + d_2^2\right)k^4 + 64\left(d_1 g_1^2 u_0^2 + d_2 g_2^2 v_0^2\right)k^2, \\ p_3(k) &= d_1^2 d_2^2 \nu k^{10} + 2d_1^2 d_2^2 q k^8 - 16 d_1 d_2\left(d_2 g_1^2 u_0^2 + d_1 g_2^2 v_0^2\right)k^6. \end{aligned}$$

Equation (16) is a bi-quadratic and can be solved analytically to produce $\omega = \omega(k)$ as:

$$\omega = \pm\sqrt{\frac{-p_2(k) \pm \sqrt{p_2^2(k) - 4p_1(k)p_3(k)}}{2p_1(k)}}. \tag{17}$$

The system is subject to MI as long as ω has complex solutions: the imaginary part, $\mathrm{Im}\{\omega\}$, will give the relative growth rate, as suggested by Eq. (15). To classify the nature of ω (real or complex) we need to solve a system of inequalities to ensure Eq. (16) only admits real solutions, thus avoiding any exponential growth. In particular, there are three polynomials in k which one needs to prove are positive. This will provide the appropriate conditions for stability. These are:

$$\begin{aligned}
\Delta(k) &= 65536k^{14}(2q + \nu k^2)[\nu^2(d_1^2 - d_2^2)^2k^8 + 4q\nu(d_1^2 - d_2^2)^2k^6 \\
&\quad + 4(d_1^2 - d_2^2)(d_1^2q^2 - 8d_1g_1^2\nu u_0^2 + d_2(-d_2q^2 + 8g_2^2v_0^2\nu)k^4 \\
&\quad - 64q(d_1^2 - d_2^2)(d_1g_1^2u_0^2 - d_2g_2^2v_0^2)k^2 + 256(d_1g_1^2u_0^2 + d_2g_2^2v_0^2)^2]^2Q_1(k), \\
P(k) &= 128(2q + \nu k^2)Q_2(k), \\
D(k) &= 64(2q + \nu k^2)[-4\nu(d_1^2 + d_2^2)k^6 - 8q(d_1^2 + d_2^2)k^4 \\
&\quad + 64(d_1g_1^2u_0^2 + d_2g_2^2v_0^2)k^2]^2Q_3(k).
\end{aligned}$$

Hence it is sufficient to show that the polynomials

$$\begin{aligned}
Q_1(k) &= (d_1^2d_2^2\nu)k^4 + (2d_1^2d_2^2q)k^2 - 16d_1d_2(d_2g_1^2u_0^2 + d_1g_2^2v_0^2), \\
Q_2(k) &= \nu(d_1^2 + d_2^2)k^4 + 2q(d_1^2 + d_2^2)k^2 - 16(d_1g_1^2u_0^2 + d_2g_2^2v_0^2), \\
Q_3(k) &= \nu^2(d_1^2 - d_2^2)^2k^8 + 4q\nu(d_1^2 - d_2^2)^2k^6 \\
&\quad + 4(d_1^2 - d_2^2)[d_1^2q^2 - 8d_1g_1^2u_0^2\nu + d_2(-d_2q^2 + 8\nu g_2^2v_0^2)]k^4 \\
&\quad - 64(d1^2 - d2^2)q(d_1g_1^2u_0^2 - d_2g_2^2v_0^2)k^2 + 256(d_1g_1^2u_0^2 - d_2g_2^2v_0^2)^2,
\end{aligned}$$

are always positive. This is obtained through Sturm's theorem [68]; since the coefficients of the highest order terms are positive and of even degree, it is sufficient to demand that these polynomials do not exhibit real roots. This happens when:

$$\text{(i)}\quad \frac{g_1^2u_0^2}{d_1} + \frac{g_2^2v_0^2}{d_2} < 0, \quad \text{and} \quad \text{(ii)}\quad d_1g_1^2u_0^2 + d_2g_2^2v_0^2 < 0.$$

It is now trivial to show that for both conditions to hold it is sufficient to pose that both d_1 and d_2 are negative. As such, the coupled system also follows the condition of stability of the single-component nonlocal equation, and stability is achieved iff the system is fully defocusing.

As mentioned above, the nonlocal term involving ν seems to have a stabilizing effect in the sense that growth rates, in the scalar case, are significantly smaller and

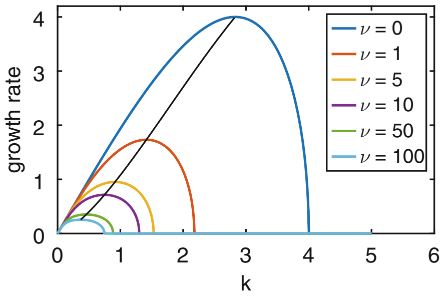


Fig. 4 The growth rate, $\text{Im}\{\omega\}$ of Eq. (17), of the focusing system for different values of the nonlocality factor ν. The black solid line corresponds to the maximum growth rate. All other parameters are kept equal to unity

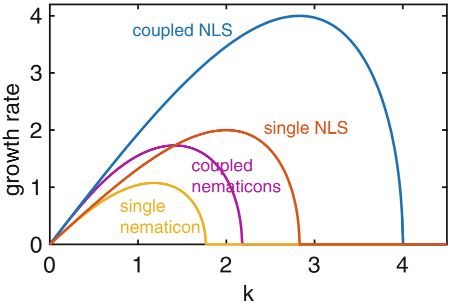


Fig. 5 Comparison of the growth rates to the relative NLS equations (Eqs. (10)–(12) with $\nu = 0$). All other parameters are kept equal to unity

MI will need more propagating distance to occur. This feature is preserved here as well, as seen in Fig. 4. In all figures, the growth rate is defined as $\text{Im}\{\omega\}$, while ω is obtained from Eq. (17). We return to this shortly.

The nonlocal term ν has a profound effect on the dynamics of plane waves. While the system is still unstable for large values of ν the range of wavenumbers that cause instability is significantly narrower and in addition the maximum growth rate is smaller. That means that for nematic crystals in particular, where $\nu = O(10^2)$, MI may be suppressed by increasing the value of the nonlocality or by choosing wavenumbers outside this narrow band that result in unstable propagation.

However, the coupling provides significantly higher growth rates as seen in Fig. 5 than the scalar system. As also expected, following this observation, the pure NLS system ($\nu = 0$) has the higher growth rates, so much so that even the single equation surpasses the coupled nonlocal system—cf. Fig. 5 (right). Furthermore, and contrary to the NLS, it has been shown, for the scalar case, that nonlocality of arbitrary shape can indeed eliminate collapse in all physical dimensions [12].

In the MI analysis above, one can identify some critical numbers that play a key role in the understanding of these results. First and foremost, we identify the so-called maximum growth rate. This value corresponds to the maxima of Fig. 4 and can be found by differentiating Eq. (17), solving the equation $\omega'(k) = 0$ for $k = k_{\max}$ and substituting back to $\omega_{\max} = \omega(k_{\max})$. The change of $\text{Im}\{\omega_{\max}\}$ with ν is shown both in Figs. 4 and 6. However, there is another value that may be interpreted in two ways. We define a critical wavenumber, k_c, which is essentially the greatest

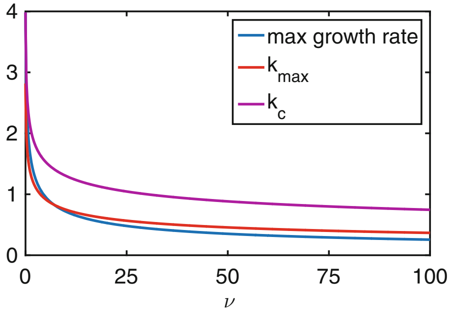


Fig. 6 Critical wavenumbers and growth rates and their dependence on the nonlocality

wavenumber for which instabilities can occur. To find this critical value one needs to solve the inequality below for k:

$$\frac{-p_2(k) \pm \sqrt{p_2^2(k) - 4p_1(k)p_3(k)}}{2p_1(k)} < 0.$$

Then the critical value can be identified as the solution of

$$d_1 d_2 \nu k^4 + 2d_1 d_2 q k^2 - 16(d_2 g_1^2 u_0^2 + d_1 g_2^2 v_0^2) = 0. \tag{18}$$

In Fig. 6, we show the dependence of these critical values with the nonlocality ν.

In particular, for the values of the figures above we find that

$$k_c = \sqrt{\frac{\sqrt{32\nu + 1}}{\nu} - \frac{1}{\nu}},$$

which also demonstrates the way this value is affected by the nonlocality. The relative $k_{\max}$ (and from that $\omega_{\max}$) value can be obtained from the solution of the equation $\nu^2 k^6 + 4\nu k^4 + 4k^2 - 32 = 0$.

Finally, and following the analysis of Ref. [14] one can seek the critical value of the nonlocality parameter that stabilizes a cw of particular wavenumber. This is retrieved again from Eq. (18) only now we solve for ν, i.e.,

$$\nu = 2\frac{-d_1 d_2 q k^2 + 8(d_2 g_1^2 u_0^2 + d_1 g_2^2 v_0^2)}{d_1 d_2 k^4}.$$

For example, for the values as above, the wave number $k = 1$ will always correspond to a stable cw iff $\nu \geq 30$, as also confirmed by Figs. 4 and 6. This critical value is not a general criterion for stability as it depends on the particular wavenumber. This means that while the particular critical value of ν may stabilize a specific wavenumber another value will render the system unstable, consistently with the analysis above. Only when the system is full defocusing MI is absent.

3.1 Vanishing Boundary Conditions

The dynamics of two-color nematicons propagation and interactions in the nonlocal limit is usually studied using variational method based on an appropriate trial function/anzatz whose parameters (amplitude, width, etc) are chosen so that the Lagrangian of the system is minimized [11, 66, 67]. However, since these are not exacts solutions they are expected to shed diffractive radiation much like the solutions of the regular NLS system. We intend to remedy this by finding exact solutions to Eqs. (10)–(12) and the conditions associated with these solutions.

As seen above the coupled system is, in the focusing case, unstable. This means that any initial condition is subject to instability. Thus, it is natural to seek conditions under which soliton solutions may exist that will not undergo the instability process. To find these solutions (if they exist), we assume that the stationary solutions of the system (10)–(12) take the form

$$u(z,x) = a_1\ \text{sech}^2(bx)e^{i\mu_1 z}, \quad v(z,x) = a_2\ \text{sech}^2(bx)e^{i\mu_2 z}, \quad \theta(x) = a_3\ \text{sech}^2(bx).$$

Substituting directly into Eqs. (10)–(12), we obtain the expressions for the soliton parameters:

$$\mu_1 = \frac{qd_1}{\nu}, \quad \mu_2 = \frac{qd_2}{\nu}, \quad b = \sqrt{\frac{q}{2\nu}}, \quad a_3 = \frac{3q\lambda}{4\nu}.$$

The solitons' amplitudes are related through

$$g_1 a_1^2 + g_2 a_2^2 = \lambda \frac{9q^2}{8\nu}, \tag{19}$$

subject to the condition:

$$\frac{d_1}{g_1} = \frac{d_2}{g_2} = \lambda.$$

Although the freedom of one free parameter (which also relates amplitude and velocity in the NLS case) is not redeemed, another property is obtained. We are now able to control the amplitude of one of the components through Eq. (19). However, this is again a fundamental difference with the NLS system. Indeed, in the NLS equations it is straightforward to obtain a variety of different cases and soliton types (bright and/or dark). Here, only the focusing case can produce bright solitons while we where not able to find dark solitons in this manner [15]. They may, however, be obtained, in the small amplitude limit, using the methods of Ref. [26]. One final comment is that this procedure may also be used for more than two equations relating the relative amplitudes through an equation of the form of Eq. (19).

The role of the nonlocal term ν is profound here as well. In particular, when $\nu \gg 1$, as is the case for liquid crystals, it may become (experimentally) more difficult to obtain soliton solutions and radiation free propagation. Indeed, from

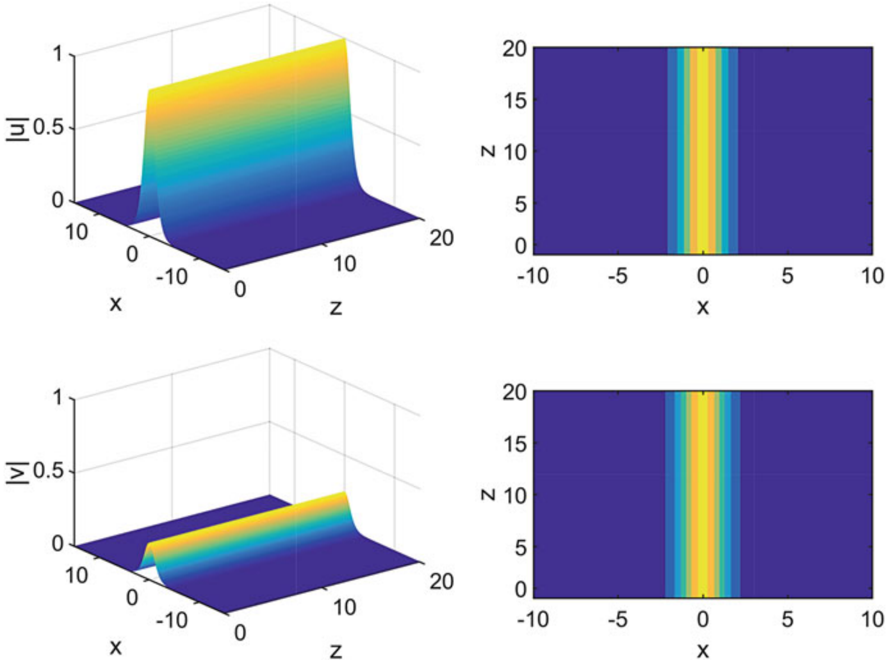


Fig. 7 A typical soliton evolution. Here we use $d_1 = g_1 = 1$, $d_2 = g_2 = 2$, $\nu = 1$, $q = 1$, $a_1 = 1$ and $a_2 = 1.5$ and a_2 is obtained from Eq. (19)

Eq. (19) the smaller the right hand side becomes the more difficult it becomes to obtain amplitudes for soliton propagation. In fact, when this term vanishes it results in $a_1 = a_2 = 0$, i.e. no soliton solutions exist. Furthermore, Eq. (19) suggests that

$$a_1^2 \leq \lambda \frac{9q^2}{8\nu g_1}, \quad a_2^2 \leq \lambda \frac{9q^2}{8\nu g_2},$$

meaning that even with the freedom to choose one of the amplitudes that cannot exceed this maximum value. With $\nu \gg 1$ it is further implied that solitons can only exist in the small amplitude limit, which we will explore further below. On the other hand if we choose initial conditions that obey the amplitude condition, Eq. (19), the result is a stable typical solitonic evolution as shown in Fig. 7.

3.2 Non-vanishing Boundary Conditions

Our analysis is now focused on soliton pairs that rely on the existence of a stable cw background and hence on the defocusing system where $d_1 g_1, d_2 g_2 < 0$. As such, we only consider Eq. (12). Write the solutions of this system in the form

$$E_1 = u_b(z)u(z,x), \tag{20a}$$

$$E_2 = v_b(z)v(z,x), \tag{20b}$$

$$\theta = \theta_b w(z,x), \tag{20c}$$

where the functions $u_b(z)$ and $v_b(z)$ correspond to the relative cw backgrounds so that

$$\left.\begin{array}{l} iu'_b - 2g_1\theta_b u_b = 0 \\ iv'_b - 2g_2\theta_b v_b = 0 \end{array}\right\} \Rightarrow \left\{\begin{array}{l} u_b(z) = u_0 e^{-2ig_1\theta_b z + ic_1} \\ v_b(z) = v_0 e^{-2ig_2\theta_b z + ic_2} \end{array}\right.$$

where $u_0, v_0, c_1, c_2 \in \mathbb{R}$ and $\theta_b = \frac{1}{q}(g_1 u_0^2 + g_2 v_0^2)$. Substituting back to Eq. (12) gives

$$i\frac{\partial u}{\partial z} + \frac{d_1}{2}\frac{\partial^2 u}{\partial x^2} - 2g_1\theta_b(w-1)u = 0, \tag{21a}$$

$$i\frac{\partial v}{\partial z} + \frac{d_2}{2}\frac{\partial^2 v}{\partial x^2} - 2g_2\theta_b(w-1)v = 0, \tag{21b}$$

$$\nu\frac{\partial^2 w}{\partial x^2} - 2qw = -\frac{2}{\theta_b}(g_1 u_0^2|E_1|^2 + g_2 v_0^2|E_2|^2), \tag{21c}$$

It is trivial to check that these are also satisfied at the boundaries where $u = v = w = 1$, and any evolution of the boundary conditions has been absorbed by the background functions. This way, the resulting equations have now fixed boundary conditions. Next, we employ the Madelung transformation:

$$u(x,z) = \rho_1(x,z)\exp[i\phi_1(x,z)],$$

$$v(x,z) = \rho_2(x,z)\exp[i\phi_2(x,z)],$$

so that:

$$d_j\frac{\partial^2 \rho_j}{\partial x^2} - 2\rho_j\frac{\partial \phi_j}{\partial z} - d_j\rho_j\left(\frac{\partial \phi_j}{\partial x}\right)^2 - 4g_j\theta_b\rho_j(w-1) = 0, \tag{22a}$$

$$\frac{\partial \rho_j}{\partial z} + \frac{1}{2}d_j\rho_j\frac{\partial^2 \phi_j}{\partial x^2} + d_j\frac{\partial \rho_j}{\partial x}\frac{\partial \phi_j}{\partial x} = 0, \tag{22b}$$

$$\nu\frac{\partial^2 w}{\partial x^2} - 2qw = -\frac{2}{\theta_b}(g_1 u_0^2\rho_1^2 + g_2 v_0^2\rho_2^2), \tag{22c}$$

where $j = 1, 2$, and recall that $w(z,x) \in \mathbb{R}$.

To analytically study system (22), and determine the unknown functions ρ_j, ϕ_j and w, we now employ the the reductive perturbation method [30]. We thus introduce the stretched variables:

$$Z = \varepsilon^3 z, \quad X = \varepsilon(x - Cz), \tag{23}$$

where C is the speed of sound (to be determined later in the analysis), namely the velocity of small-amplitude and long-wavelength waves propagating along the background. Additionally, we expand amplitudes and phases in powers of ε as follows:

$$\rho_j = \rho_{j0} + \varepsilon^2 \rho_{j2} + \varepsilon^4 \rho_{j4} + \cdots, \tag{24a}$$

$$\phi_j = \varepsilon \phi_{j1} + \varepsilon^3 \phi_{j3} + \varepsilon^5 \phi_{j5} + \cdots, \tag{24b}$$

$$w = 1 + \varepsilon^2 w_2 + \varepsilon^4 w_4 + \cdots, \tag{24c}$$

where $\rho_{j0} = 1$ and the rest of the unknown fields depend on the stretched variables (23). These values for ρ_{j0} is not only a result obtained from the perturbation analysis but is also anticipated from Eqs. (20) and (21). Recall, that the background has been removed, absorbed by the functions u_b and v_b, which, in general, are not equal.

Substituting back to Eq. (22) we obtain the following results. First, in the linear limit, i.e., at the lowest-order approximation in ε, we derive equations connecting the unknown fields, namely:

$$w_2 = \frac{2}{q\theta_b}(g_1 u_0^2 \rho_{21} + g_2 v_0^2 \rho_{22}), \quad \phi_{21} = \frac{g_2}{g_1}\phi_{11}, \tag{25a}$$

$$\rho_{22} = \frac{d_2 g_2}{d_1 g_1}\rho_{21}, \quad \frac{d_j}{2}\frac{\partial \phi_{j2}}{\partial X} = C\rho_{j2}, \tag{25b}$$

as well as the speed of sound

$$C^2 = \frac{2}{q}(d_1 g_1 u_0^2 + d_2 g_2 v_0^2). \tag{26}$$

Obviously, Eq. (25) suggest that only one equation for one of these fields will suffice to determine the rest of the unknown fields ρ_{j2}, ϕ_{j1} and w_2. This equation is derived to the next order of approximation, and turns out to be the following nonlinear equation for the field ρ_{12}:

$$\frac{\partial \rho_{12}}{\partial Z} + A_1 \frac{\partial^3 \rho_{12}}{\partial X^3} + 6A_2 \rho_{12} \frac{\partial \rho_{12}}{\partial X} = 0, \tag{27}$$

where coefficients A_1 and A_2 are given by:

$$A_1 = \frac{\nu C^4 - (d_1^3 g_1^2 u_0^2 + d_2^3 g_2^2 v_0^2)}{4C^2 q},$$

$$A_2 = \frac{d_1^2 g_1^3 u_0^2 + d_2^2 g_2^3 v_0^2}{C d_1 g_1 q}.$$

Equation (27) is the renowned KdV equation, which is completely integrable by means of the IST [2], and finds numerous applications in a variety of physical contexts [1, 17]. More recently, a KdV equation was derived from the single-component version of Eq. (12), and used to describe small-amplitude nematicons [26]; notice that the KdV model derived in [26] is identical with Eq. (27) when the coupling constants are set to zero. Notably, the same procedure can result in other integrable forms of the KdV in higher dimensions, such as the Kadomtsev-Petviashvilli (KP) equation, Johnson's equation, and others [27, 28].

These asymptotic reductions provide information on the type of the soliton solutions the original system may exhibit up to (and including) $O(\varepsilon^2)$. Indeed, first we note that the soliton solution of Eq. (27) takes the form (e.g., Ref. [1]),

$$\rho_{12}(Z, X) = \frac{2A_1}{A_2}\eta^2 \operatorname{sech}^2(\eta X - 4\eta^3 A_1 Z + X_0),$$

where η and X_0 are free parameters, setting the amplitude/width and initial position of the soliton, respectively. Then, it is straightforward to retrieve the pertinent phase,

$$\phi_{11} = -\frac{4A_1 C}{A_2 d_1}\eta \tanh(\eta X - 4\eta^3 A_1 Z + X_0),$$

so that, finally, the solutions for the two components may be written as:

$$E_1(z, x) \approx u_b(z)(1 + \varepsilon^2 \rho_{12}) \exp(i\varepsilon\phi_{12}), \tag{28}$$

$$E_2(z, x) \approx v_b(z)\left(1 + \varepsilon^2 \frac{d_2 g_2}{d_1 g_1}\rho_{12}\right) \exp\left(i\varepsilon \frac{g_2}{g_1}\phi_{12}\right). \tag{29}$$

It is now important to notice that the type of the solitons (28) and (29) depends crucially on the sign of the ratio A_1/A_2; this quantity changes sign according to the critical value ν_c, given by:

$$\nu_c = \frac{q^2\left(d_1^3 g_1^2 u_0^2 + d_2^3 g_2^2 v_0^2\right)}{4\left(d_1 g_1 u_0^2 + d_2 g_2 v_0^2\right)^2}.$$

Indeed, if the nonlocality parameter ν is such that $\nu < \nu_c$ (i.e., $A_1/A_2 > 0$), the solitons are dark, namely are intensity dips off of the cw background. On the other hand, if $\nu > \nu_c$ (i.e., $A_1/A_2 < 0$) the solitons are *antidark*, namely intensity elevations on top of the cw background. Notice that Eq. (19) suggest that the relative signs between the modes are the same and, as such, the only allowed pairs are solitons of the same kind. It should also be mentioned that if $A_1 = 0$, modification of the asymptotic analysis and inclusion of higher-order terms is needed. This has been addressed, to a certain extent, in Ref. [20], where, it was found that higher order dispersive terms can lead to resonant interactions with radiation, as expected, for the higher (fifth) order KdV equation.

To demonstrate the validity of our analysis, we perform direct numerical simulations we thus integrate Eq. (12) employing a high accuracy spectral integrator, and using initial conditions (at $z = 0$) taken from Eqs. (28) and (29), for both the dark and the antidark soliton pairs. The results are shown in Fig. 8, where a typical evolution of a dark soliton pair is depicted. Here, we choose parameter values $d_1 = d_2/1.5 = g_1 = g_2 = 1$, $u_0 = v_0 = 1$ and $q/5 = \nu = 1$. Similarly, in Fig. 9, we show a typical evolution of an antidark soliton pair; all parameters remain the same except $q = 1$. In both cases, it is clear that the solitons, not only exist, but also propagate undistorted on top of the cw background. It is also observed that the solitons propagate with constant speed, with the antidark soliton pair traveling faster than the dark one, as expected from Eq. (26).

3.3 Vanishing and Non-vanishing Boundary Conditions

Apart from soliton pairs of the same type, it is also possible to derive vector soliton solutions composed by different types of solitons. This can be done upon seeking solutions of the system of Eq. (12) such that one of the components decays to zero

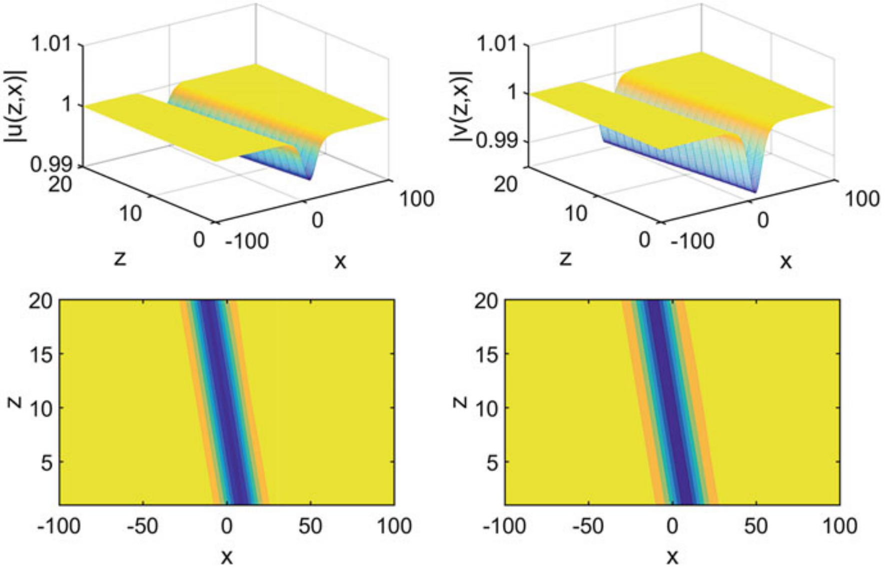


Fig. 8 The evolution of a typical dark soliton pair. Left and right columns depict the two components, while top and bottom panels show three-dimensional plots, and spatiotemporal contour plots, respectively

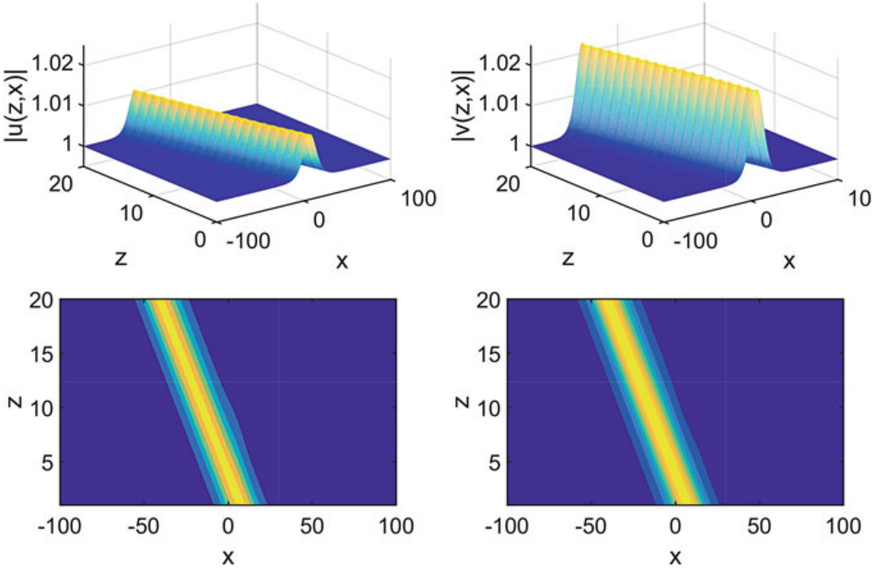


Fig. 9 Similar to Fig. 8, but now for a typical antidark soliton pair

at infinity, while the other tends to a constant, as before. In such a case, solutions of Eq. (12) are again taken to be of the form of Eq. (20), but now we assume that the background functions are given by:

$$u_b(x,z) = \exp\left[ikx - i(\omega - \varepsilon^2 \Omega)z\right],$$

$$\omega = \frac{1}{2q}(d_1 k^2 q + 4g_1 g_2 v_0^2),$$

$$v_b(z) = v_0 \exp(-2ig_2\theta_b z + i\psi_1), \;\; \theta_b = \frac{g_2 v_0^2}{q}.$$

Then, the system (12) is reduced to the form:

$$iu_z + \frac{d_1}{2}u_{xx} - 2g_1\theta_b(w-1)u - id_1 k u_x = 0,$$

$$iv_z + \frac{d_2}{2}v_{xx} - 2g_2\theta_b(w-1)v = 0,$$

$$\nu w_{xx} - 2qw = -\frac{2}{\theta_b}(g_1|u|^2 + g_2 v_0^2 |v|^2).$$

Then, using the stretched variables (23) and the asymptotic expansions (24), and following the procedure of the previous section, we obtain the following results.

First, at the leading order, $O(1)$, we get $\rho_{10} = 0$ and $\rho_{20} = w_0 = 1$, while in the linear limit, i.e., at the orders $O(\varepsilon^2)$ and $O(\varepsilon^3)$, we derive equations connecting the unknown fields, namely:

$$w_2 = 2\rho_{22}, \quad C\frac{\partial \phi_{21}}{\partial X} = \frac{4g_2^2 v_0^2}{q}\rho_{22},$$
$$\frac{d_2}{2}\frac{\partial^2 \phi_{21}}{\partial X^2} = C\frac{\partial \rho_{22}}{\partial X}, \quad k = \frac{C}{d_1}.$$

The above equations suggest that, now, the speed of sound is given by:

$$C^2 = \frac{2g_2^2 v_0^2 d_2}{q}.$$

Next, in the nonlinear regime, namely at $O(\varepsilon^4)$ and $O(\varepsilon^5)$, we obtain the following system for the fields ρ_{j2}:

$$\frac{8g_2^2 v_0^2}{Cq}\frac{\partial \rho_{22}}{\partial Z} - \frac{\left(d_2 q^2 - 4g_2^2 v_0^2 \nu\right)}{2q^2}\frac{\partial^3 \rho_{22}}{\partial X^3} + \frac{24g_2^2 v_0^2}{q}\rho_{22}\frac{\partial \rho_{22}}{\partial X}$$
$$+\frac{2g_1 g_2}{q}\frac{\partial}{\partial X}\left(\rho_{12}^2\right) = 0, \tag{30a}$$

$$\frac{d_1}{2}\frac{\partial^2 \rho_{12}}{\partial X^2} - \frac{4g_1 g_2 v_0^2}{q}\rho_{12}\rho_{22} = \varOmega\rho_{12}, \tag{30b}$$

as well as equations connecting fields that can be determined at a higher-order approximation. The system of Eqs. (30) is the so-called Mel'nikov system [51–53], and is apparently composed of a KdV equation with a self-consistent source, which satisfies a stationary Schrödinger equation. This system has been derived in earlier works to describe dark-bright solitons in nonlinear optical systems [21] and in Bose-Einstein condensates [5, 70]. The Mel'nikov system is completely integrable by the inverse scattering transform, and possesses a soliton solution of the form [52]:

$$\rho_{22}(Z, X) = -\frac{d_1 q}{4g_1 g_2 v_0^2}\eta^2 \mathrm{sech}^2(\eta X + bZ + X_0),$$
$$\rho_{12}(Z, X) = A\mathrm{sech}(\eta X + bZ + X_0),$$

where $\varOmega = (1/2)\eta^2 d_1$, while parameters η, A, and b are connected through the following equation:

$$Cd_1\left(4\nu g_2^2 v_0^2 - d_2 q^2\right)\eta^4 + 4qd_1 g_2^2 v_0^2 b\eta - 4Cg_1^2 g_2^2 v_0^2 A^2 = 0.$$

Using the above expressions, we can now express the relevant approximate [valid up to $O(\varepsilon^2)$] solutions of the original system for the two components $E_{1,2}$ as follows:

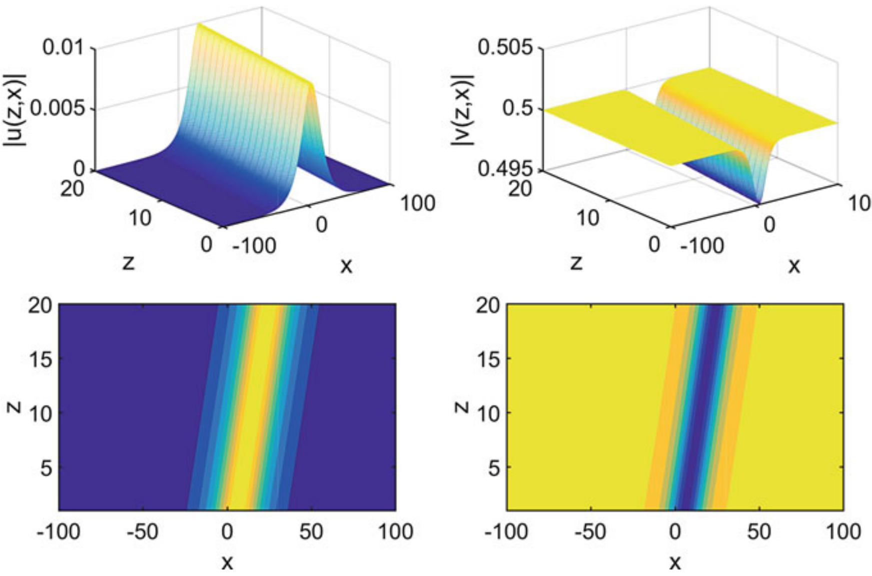


Fig. 10 Similar to Fig. 8, but now for a typical dark-bright soliton pair

$$E_1(z,x) \approx \varepsilon^2 u_b(z)\rho_{12} \exp(i\varepsilon\phi_{12}), \tag{31}$$

$$E_2(z,x) \approx \upsilon_b(z)\left(1+\varepsilon^2\rho_{22}\right)\exp\left(i\varepsilon\phi_{22}\right). \tag{32}$$

It is clear that the above solution represents a dark-bright soliton pair, for the components E_2 and E_1, respectively.

As in the case of the dark and antidark soliton pairs, we numerically integrate Eq. (12), using initial conditions (at $z = 0$) taken from Eqs. (31) and (32). The results are shown in Fig. 10, where a typical evolution of a dark-bright soliton pair is depicted. Here we choose all parameters equal to unity, except $\upsilon_0 = 1/2$. In this case too, the dark-bright soliton, not only exist, but also propagates undistorted with constant velocity, in excellent agreement with our analytical predictions.

4 The Fully 3D Scalar Nonlocal System

We now consider a natural, (3 + 1)-dimensional generalization of the nonlocal NLS model, which is again composed by a system of two coupled equations: one for the complex field amplitude u, and one for the nonlinear correction to the refractive index n (which is a real function). Here, in the present higher-dimensional setting, we focus on the defocusing nonlinearity, since the setting with the focusing nonlinearity is generally subject to collapse. The defocusing model applies to light propagation in nematic liquid crystals [9, 16], but also to thermal nonlinear optical media [45] (see also relevant theoretical and experimental work in [56]). The system under consideration is expressed in the following dimensionless form [54, 55]:

$$iu_z + \frac{1}{2}(\Delta u - Du_{tt}) - 2nu = 0, \tag{33}$$

$$d\Delta n - 2qn + 2|u|^2 = 0, \tag{34}$$

where subscripts denote partial derivatives, z is as before the evolution variable (propagation coordinate normalized to the diffraction length), t is retarded time, and Δ is the transverse Laplacian. Below, we consider both physically relevant geometries, Cartesian and cylindrical, for which the Laplacian respectively reads:

$$\Delta = \partial_x^2 + \partial_y^2, \qquad \Delta = \frac{1}{r}\partial_r(r\partial_r) + \frac{1}{r^2}\partial_\theta^2,$$

with transverse coordinates $\boldsymbol{r}_\perp = (x,\ y)$ or $\boldsymbol{r}_\perp = (r,\ \theta)$, respectively, normalized with respect to the beam width. Additionally, the real constants D, d, and q in Eqs. (33) and (34), which are assumed to be $\mathscr{O}(1)$ parameters in our analysis below, have the following physical significance. First, D represents the ratio of diffraction and dispersion lengths, with $D > 0$ ($D < 0$) corresponding to anomalous (normal) group-velocity dispersion (GVD). Second, in the context of nematic liquid crystals, and similarly to the 1D setting, the parameter q is related to the square of the applied static electric field that pre-tilts the nematic dielectric [6, 10, 59]. Third, parameter d measures the relative width of the response of the medium to the light field, and is connected to the nonlocality scale of the nonlinear response of the medium: in the limit $d \to 0$, the system of Eqs. (33) and (34) decouples and is reduced to a local $(3+1)$-dimensional NLS equation with a cubic defocusing nonlinearity.

To start our analysis, we use the Madelung transformation

$$u = u_0\sqrt{\rho}\exp(i\phi), \tag{35}$$

(u_0 being an arbitrary complex constant) to separate the real functions for the amplitude ρ and phase ϕ of u in Eq. (33), and derive from Eqs. (33) and (34) the following system:

$$\phi_z + 2n + \frac{1}{2}\left[(\nabla\phi)^2 - D\phi_t^2\right] - \frac{1}{2}\rho^{-1/2}\left[\Delta\rho^{1/2} - D\left(\rho^{1/2}\right)_{tt}\right] = 0, \tag{36}$$

$$\rho_z + \nabla\cdot(\rho\nabla\phi) - D(\rho\phi_t)_t = 0, \tag{37}$$

$$d\Delta n - 2qn + 2|u_0|^2\rho = 0, \tag{38}$$

where ∇ is the gradient operator, given by:

$$\nabla = (\partial_x,\ \partial_y), \qquad \nabla = \left(\partial_r,\ \frac{1}{r}\partial_\theta\right)$$

for the Cartesian or the cylindrical geometry, respectively.

It is readily observed that the above system possesses an exact steady-state solution of the form:

$$\phi = -\frac{2}{q}|u_0|^2 z, \qquad \rho = 1, \qquad n = \frac{1}{q}|u_0|^2,$$

which corresponds to the continuous-wave (cw) solution

$$u = u_0 \exp\left(-\frac{2i}{q}|u_0|^2 z\right), \quad n = \frac{1}{q}|u_0|^2, \tag{39}$$

of Eqs. (33) and (34). Note that the constant amplitude u_0 can be absorbed into ρ in the Madelung transformation; nevertheless, for convenience, we opt to use u_0 in Eq. (35) so that no extra free phase appears on the solutions that we present below, thus making presentation more clear. To further elaborate on this point, and also to underline the importance of this cw solution in our analysis (because the cw defines the background on top of which our solutions will propagate), let us write the solution of Eqs. (33) and (34) as:

$$u = U_0(z)\bar{u}, \quad n = n_0\bar{n},$$

where the background (cw) solution satisfies

$$\mathrm{i}\frac{dU_0}{dz} = 2n_0U_0, \quad n_0 = \frac{|U_0|}{q}.$$

Obviously, the solution of the above equations is given in Eq. (39), while $\bar{u}$ and $\bar{n}$ satisfy the system

$$i\bar{u}_z + \frac{1}{2}(\Delta\bar{u} - D\bar{u}_{tt}) - 2\frac{|u_0|^2}{q}(1-\bar{n})\bar{u} = 0,$$

$$\frac{d}{q}\Delta\bar{n} - 2\bar{n} + 2|\bar{u}|^2 = 0.$$

This system possesses a cw solution of unit amplitude, as would be the case if u_0 was not used in the Madelung transformation (i.e., in other words, u_0 will inevitably appear in the cw background solution).

Below we seek nonlinear excitations (e.g., solitary waves) which propagate on top of this cw background. It is, thus, relevant to investigate if this solution is subject to modulational instability (MI): evidently, nonlinear excitations corresponding to an unstable background do not have any physical purport. The stability of the cw solution can be investigated upon employing Eqs. (36)–(38) as follows. Let

$$\rho = 1 + \tilde{\rho}, \quad \phi = -\frac{2}{q}|u_0|^2 z + \tilde{\phi}, \quad n = \frac{1}{q}|u_0|^2 + \tilde{n},$$

where small perturbations $\tilde{\rho}$, $\tilde{\phi}$ and $\tilde{n}$ are assumed to be $\propto \exp[i(k_z z + \boldsymbol{k}_\perp \cdot \boldsymbol{r}_\perp - \omega t)]$. Here, we should recall that the evolution variable in our problem is the propagation distance z and, thus, the MI analysis is performed with respect to this variable; as such, k_z and its roots (real or imaginary) will determine the stability of the cw solution. To this end, substituting the above ansatz into Eqs. (36)–(38), we find that small-amplitude linear waves obey a dispersion relation of the following form,

$$k_z^2 = \frac{2|u_0|^2}{q}(\boldsymbol{k}_\perp^2 - D\omega^2)\left(1 + \frac{d\boldsymbol{k}_\perp^2}{2q}\right)^{-1} + \frac{1}{4}(\boldsymbol{k}_\perp^2 - D\omega^2)^2. \tag{40}$$

The results stemming from the above equation are as follows. First, Eq. (40) shows that the cw solution is always modulationally stable, i.e., $k_z \in \mathbb{R}\ \forall\, \boldsymbol{k}_\perp,\ \omega$, provided $D < 0$. Note, that in the (1+1)-dimensional case (corresponding to $k_y = 0$ and $D = 0$, or $\boldsymbol{k}_\perp = 0$ and $D = -1$), this result recovers the one presented in the previous section, also obtained in Ref. [45]. Second, if $D = |D| > 0$, the cw solution is unstable: in this case, perturbations grow exponentially, with the instability growth rate given by $\mathrm{Im}(k_z)$. Thus, hereafter, we focus on this case, and assume that $D = -|D|$, corresponding to the anomalous GVD regime. It is, therefore, clear that the asymptotic analysis and results that we present in the following sections are only valid in this regime; in the opposite case, of $D = |D| > 0$, since the cw background is unstable, any small perturbations on top of it will result in collapsing solutions.

Another physically relevant information stemming from Eq. (40) is that, in the long-wavelength and low-frequency limit (i.e., $|\boldsymbol{k}_\perp|,\ \omega \to 0$), small-amplitude spatial or temporal waves propagate on top of the cw background with "sound velocities" C^2 or V^2, respectively, which are given by:

$$C^2 = \frac{2|u_0|^2}{q}, \qquad V^2 = C^2|D|. \tag{41}$$

These characteristic velocities can also be determined, in a self-consistent manner, in the framework of the reductive perturbation method (see, e.g., previous section and Ref. [23]). It is also noted that in the unstable case of $D = |D| > 0$, the velocity V becomes imaginary, a fact that also indicates that perturbations of the cw solution grow exponentially in the propagation distance z.

4.1 The Boussinesq Equation

Observing that the dispersion relation (40) resembles the one of a Boussinesq equation [1, 32], we now derive from Eqs. (36)–(38) a Boussinesq equation, for either Cartesian or cylindrical geometry. We thus seek solutions of Eqs. (36)–(38) in the form of the following asymptotic expansions:

$$\phi = -\frac{2}{q}|u_0|^2 z + \varepsilon^{1/2}\Phi, \quad \rho = 1 + \varepsilon\rho_1 + \varepsilon^2\rho_2 + \cdots, \quad n = \frac{1}{q}|u_0|^2 + \varepsilon n_1 + \varepsilon^2 n_2 + \cdots, \tag{42}$$

where ε is a formal small parameter ($0 < \varepsilon \ll 1$), while unknown real functions Φ, ρ_j and n_j are assumed to depend on "slow variables". In particular, for the Cartesian geometry, Φ, ρ_j and n_j depend on $\{Z, X, Y, T\}$, while, for the cylindrical geometry, on $\{Z, R, \theta, T\}$ (the angular coordinate θ is assumed to remain unchanged); these slow variables are defined as:

$$Z = \varepsilon^{1/2}z, \quad X = \varepsilon^{1/2}x, \quad Y = \varepsilon^{1/2}y, \quad R = \varepsilon^{1/2}r, \quad T = \varepsilon^{1/2}t. \tag{43}$$

Substituting the expansions (42) into Eqs. (36)–(38), and using the variables in Eq. (43), we obtain the following results. First, Eq. (36) reads:

$$\Phi_Z + 2n_1 + \varepsilon\left\{\frac{1}{2}\left[(\tilde{\nabla}\Phi)^2 + |D|\Phi_T^2\right] - \frac{1}{4}\left(\tilde{\Delta}\rho_1 + |D|\rho_{1TT}\right) + 2n_2\right\} = \mathcal{O}(\varepsilon^2), \tag{44}$$

where

$$\tilde{\Delta} = \partial_X^2 + \partial_Y^2, \qquad \tilde{\nabla} = (\partial_X,\ \partial_Y),$$

for the Cartesian case, while

$$\tilde{\Delta} = \frac{1}{R}\partial_R(R\partial_R) + \frac{1}{R^2}\partial_\theta^2, \qquad \tilde{\nabla} = \left(\partial_R,\ \frac{1}{R}\partial_\theta\right),$$

for the cylindrical case. Second, Eq. (37) leads, at orders $\mathcal{O}(\varepsilon^{3/2})$ and $\mathcal{O}(\varepsilon^{5/2})$, to the following equations, respectively:

$$\rho_{1Z} + \tilde{\Delta}\Phi + |D|\Phi_{TT} = 0,$$

$$\rho_{2Z} + \tilde{\nabla}\cdot(\rho_1\tilde{\nabla}\Phi) + |D|(\rho_1\Phi_T)_T = 0.$$

Finally, Eq. (38), at orders $\mathcal{O}(\varepsilon)$ and $\mathcal{O}(\varepsilon^2)$, lead, respectively, to the equations:

$$-2qn_1 + 2|u_0|^2\rho_1 = 0, \tag{45}$$

$$d\tilde{\Delta}n_1 - 2qn_2 + 2|u_0|^2\rho_2 = 0. \tag{46}$$

The leading-order part of Eq. (44), together with Eq. (45), provides the following connection between functions Φ, n_1 and ρ_1:

$$\Phi_Z = -2n_1 = -C^2\rho_1. \tag{47}$$

Furthermore, from the system of Eqs. (44)–(46), it is possible to eliminate $\rho_{1,2}$ and $n_{1,2}$, and derive the following equation for Φ:

$$\Phi_{ZZ} - C^2\left(\tilde{\Delta}\Phi + |D|\Phi_{TT}\right) + \varepsilon\left\{\frac{1}{4C^2}\left(\alpha\tilde{\Delta}\Phi + |D|\Phi_{TT}\right)_{ZZ}\right.$$
$$\left. + \left[(\tilde{\nabla}\Phi)^2 + |D|\Phi_T^2\right]_Z + \Phi_Z\left(\tilde{\Delta}\Phi + |D|\Phi_{TT}\right)\right\} = \mathscr{O}(\varepsilon^2), \tag{48}$$

where the parameter α is given by:

$$\alpha = 1 - \frac{4d|u_0|^2}{q^2}. \tag{49}$$

It is clear that, to leading-order, Eq. (48) is a linear wave equation indicating that the velocities of spatial or temporal waves are indeed those given in Eq. (41). In addition, at order $\mathscr{O}(\varepsilon)$, Eq. (48) incorporates fourth-order dispersion terms and quadratic nonlinear terms. Obviously, Eq. (48) is a Boussinesq-type equation, either in Cartesian or cylindrical coordinates. The Boussinesq equation has been originally proposed for studies of waves in shallow water [1, 32, 35], but later it was also used in different contexts, ranging from ion-acoustic waves in plasmas [29, 35] to mechanical lattices and electrical transmission lines [64].

A Boussinesq equation, similar to that in Eq. (48), was derived from a (2 + 1)-dimensional NLS equation, in Cartesian coordinates, with a local defocusing nonlinearity [62]; analysis of the Boussinesq model [61] was used in [62] to investigate self-focusing and transverse instability of plane dark solitons (see also the review [41] and references therein). In fact, the Cartesian version of the Boussinesq model (48) is reduced to the one derived in [62] in the limit of $d \to 0$, corresponding to the local nonlinearity case.

4.2 Kadomtsev-Petviashvilli-Type Equations

We now proceed to derive the far-field equations stemming from the Boussinesq model (48), in the framework of multiscale asymptotic expansions. As is known, the far-field of the Boussinesq equation in (1 + 1)-dimensions is a pair of two KdV equations [1], while in (2 + 1)-dimensions, it is a pair of KP equations [62], for right- and left-going waves. Below we show that Eq. (48) gives rise to (3 + 1)-dimensional KP-type models for such waves. In addition, we will distinguish cases corresponding to two different types of solitary waves that may be supported either in Cartesian or cylindrical geometry: one, is oblique tube-shaped, oriented under a uniquely determined angle to the propagation axis, i.e., a *spatial solitary wave*; the other, is a constant-shape localized perturbation propagating along the z-axis, i.e., a *temporal solitary wave*.

4.2.1 Spatial Solitary Waves

First, we consider spatial solitary waves, which may have either the form of stripes propagating on the XZ plane (Cartesian geometry), or exhibit an annular shape, with the ring radius varying with the propagation distance (cylindrical geometry). We thus introduce the variables:

$$\chi = X - CZ, \quad \tilde{\chi} = X + CZ, \quad \mathscr{Z} = \varepsilon Z, \quad \mathscr{Y} = \varepsilon^{1/2} Y, \quad \mathscr{T} = \varepsilon^{1/2} T,$$

and

$$\rho = R - CZ, \quad \tilde{\rho} = R + CZ, \quad \mathscr{Z} = \varepsilon Z, \quad \Theta = \varepsilon^{-1/2} \theta, \quad \mathscr{T} = \varepsilon^{1/2} T,$$

for the two geometries, respectively. We also look for solutions of Eq. (48) in the form of the asymptotic expansion:

$$\Phi = \Phi_0 + \varepsilon \Phi_1 + \cdots . \tag{50}$$

Substituting Eq. (50) into Eq. (48), we obtain the following results. At leading-order, $\mathscr{O}(1)$:

$$4C^2 \Phi_{0\chi\tilde{\chi}} = 0,$$

for the Cartesian and cylindrical geometry, respectively. The above equations imply that, in each case, Φ_0 can be expressed as a superposition of a right-going wave, $\Phi_0^{(R)}$, depending on χ or ρ, and a left-going one, $\Phi_0^{(L)}$, depending on $\tilde{\chi}$ or $\tilde{\rho}$, namely:

$$\Phi_0 = \Phi_0^{(R)} + \Phi_0^{(L)}. \tag{51}$$

Second, at order $\mathscr{O}(\varepsilon)$:

$$\begin{gathered} 4C^2 \Phi_{1\chi\tilde{\chi}} = -C\left(\Phi_{0\chi\chi}^{(R)} \Phi_{0\tilde{\chi}}^{(L)} - \Phi_{0\chi}^{(R)} \Phi_{0\tilde{\chi}\tilde{\chi}}^{(L)}\right) \\ + \left[\left(-2C\Phi_{0\mathscr{Z}}^{(R)} + \frac{\alpha}{4}\Phi_{0\chi\chi\chi}^{(R)} - \frac{3C}{2}\Phi_{0\chi}^{(R)2}\right)_{\chi} - C^2\left(\Phi_{0\mathscr{Y}\mathscr{Y}}^{(R)} + |D|\Phi_{0\mathscr{T}\mathscr{T}}^{(R)}\right)\right] \\ + \left[\left(2C\Phi_{0\mathscr{Z}}^{(L)} + \frac{\alpha}{4}\Phi_{0\tilde{\chi}\tilde{\chi}\tilde{\chi}}^{(L)} + \frac{3C}{2}\Phi_{0\tilde{\chi}}^{(L)2}\right)_{\tilde{\chi}} - C^2\left(\Phi_{0\mathscr{Y}\mathscr{Y}}^{(L)} + |D|\Phi_{0\mathscr{T}\mathscr{T}}^{(L)}\right)\right], \end{gathered} \tag{52}$$

for the Cartesian geometry, and

$$4C^2\varPhi_{1\rho\tilde{\rho}} = -C\left(\varPhi_{0\rho\rho}^{(R)}\varPhi_{0\tilde{\rho}}^{(L)} - \varPhi_{0\rho}^{(R)}\varPhi_{0\tilde{\rho}\tilde{\rho}}^{(L)}\right)$$
$$+\left[\left(-2C\varPhi_{0\mathscr{X}}^{(R)} + \frac{\alpha}{4}\varPhi_{0\rho\rho\rho}^{(R)} - \frac{3C}{2}\varPhi_{0\rho}^{(R)2} - \frac{C}{\mathscr{X}}\varPhi_{0}^{(R)}\right)_{\rho} - \frac{1}{\mathscr{X}^2}\varPhi_{0\Theta\Theta}^{(R)} - V^2\varPhi_{0\mathscr{T}\mathscr{T}}^{(R)}\right]$$
$$+\left[\left(2C\varPhi_{0\mathscr{X}}^{(L)} + \frac{\alpha}{4}\varPhi_{0\tilde{\rho}\tilde{\rho}\tilde{\rho}}^{(L)} + \frac{3C}{2}\varPhi_{0\tilde{\rho}}^{(L)2} - \frac{C}{\mathscr{X}}\varPhi_{0}^{(L)}\right)_{\tilde{\rho}} - \frac{1}{\mathscr{X}^2}\varPhi_{0\Theta\Theta}^{(R)} - V^2\varPhi_{0\mathscr{T}\mathscr{T}}^{(L)}\right]. \tag{53}$$

for the cylindrical geometry. Upon integrating Eq. (52) in χ or $\tilde{\chi}$ [Eq. (53) in ρ or $\tilde{\rho}$], it is obvious that the terms in square brackets in the right-hand side are secular, because are functions of χ or $\tilde{\chi}$ (of ρ or $\tilde{\rho}$) alone. Removal of these terms leads to two uncoupled nonlinear evolution equations for $\varPhi_0^{(R)}$ and $\varPhi_0^{(L)}$. Furthermore, employing Eq. (47), it is straightforward to find that the amplitude ρ_1 can also be decomposed to a left- and a right-going wave, i.e., $\rho_1 = \rho_1^{(R)} + \rho_1^{(L)}$, with

$$\varPhi_{0\chi}^{(R)} = C\rho_1^{(R)}, \quad \varPhi_{0\tilde{\chi}}^{(L)} = -C\rho_1^{(L)}, \quad \text{and} \quad \varPhi_{0\rho}^{(R)} = C\rho_1^{(R)}, \quad \varPhi_{0\tilde{\rho}}^{(L)} = -C\rho_1^{(L)}.$$

Then, using the above expressions, the equations for $\varPhi_0^{(R)}$ and $\varPhi_0^{(L)}$ yield, in each geometry, two uncoupled equations for $\rho_1^{(R)}$ and $\rho_1^{(L)}$. In Cartesian geometry, these equations are:

$$\left(\rho_{1\mathscr{X}}^{(R)} - \frac{\alpha}{8C}\rho_{1\chi\chi\chi}^{(R)} + \frac{3C}{2}\rho_1^{(R)}\rho_{1\chi}^{(R)}\right)_{\chi} + \frac{C}{2}\left(\rho_{1\mathscr{Y}\mathscr{Y}}^{(R)} + |D|\rho_{1\mathscr{T}\mathscr{T}}^{(R)}\right) = 0, \tag{54}$$

$$\left(\rho_{1\mathscr{X}}^{(L)} + \frac{\alpha}{8C}\rho_{1\tilde{\chi}\tilde{\chi}\tilde{\chi}}^{(L)} - \frac{3C}{2}\rho_1^{(L)}\rho_{1\tilde{\chi}}^{(L)}\right)_{\tilde{\chi}} - \frac{C}{2}\left(\rho_{1\mathscr{Y}\mathscr{Y}}^{(L)} + |D|\rho_{1\mathscr{T}\mathscr{T}}^{(L)}\right) = 0. \tag{55}$$

On the other hand, equations for $\rho_1^{(R,L)}$ in cylindrical geometry, are:

$$\left(\rho_{1\mathscr{X}}^{(R)} - \frac{\alpha}{8C}\rho_{1\rho\rho\rho}^{(R)} + \frac{3C}{2}\rho_1^{(R)}\rho_{1\rho}^{(R)} + \frac{1}{2\mathscr{X}}\rho_1^{(R)}\right)_{\rho} + \frac{1}{2C}\left(\frac{1}{\mathscr{X}^2}\rho_{1\Theta\Theta}^{(R)} + V^2\rho_{1\mathscr{T}\mathscr{T}}^{(R)}\right) = 0, \tag{56}$$

$$\left(\rho_{1\mathscr{X}}^{(L)} + \frac{\alpha}{8C}\rho_{1\tilde{\rho}\tilde{\rho}\tilde{\rho}}^{(L)} - \frac{3C}{2}\rho_1^{(L)}\rho_{1\tilde{\rho}}^{(L)} - \frac{1}{2\mathscr{X}}\rho_1^{(L)}\right)_{\tilde{\rho}} - \frac{1}{2C}\left(\frac{1}{\mathscr{X}^2}\rho_{1\Theta\Theta}^{(L)} + V^2\rho_{1\mathscr{T}\mathscr{T}}^{(L)}\right) = 0. \tag{57}$$

4.2.2 Temporal Solitary Waves

We now proceed with the case of temporal solitary waves. First, introduce the variables:

$$\tau = T - VZ, \quad \tilde{\tau} = T + VZ, \quad \mathscr{Z} = \varepsilon Z,$$

as well as

$$\mathscr{X} = \varepsilon^{1/2}X, \quad \mathscr{Y} = \varepsilon^{1/2}Y, \quad \text{and} \quad \mathscr{R} = \varepsilon^{1/2}R, \quad \Theta = \varepsilon^{-1/2}\theta,$$

for the Cartesian and cylindrical geometry, respectively. Then, in each case, utilizing the above variables and the asymptotic expansion (50), we obtain from Eq. (48) the leading-order equation:

$$4V^2\Phi_{0\tau\tilde{\tau}} = 0,$$

which yields again Eq. (51). Furthermore, working as in the previous case, we obtain at order $\mathscr{O}(\varepsilon)$:

$$\begin{aligned} 4V^2\Phi_{1\tau\tilde{\tau}} = -V|D|\left(\Phi^{(R)}_{0\tau\tau}\Phi^{(L)}_{0\tilde{\tau}} - \Phi^{(R)}_{0\tau}\Phi^{(L)}_{0\tilde{\tau}\tilde{\tau}}\right) \\ +\left[\left(-2V\Phi^{(R)}_{0\mathscr{Z}} + \frac{D^2}{4}\Phi^{(R)}_{0\tau\tau\tau} - \frac{3|D|V}{2}\Phi^{(R)2}_{0\tau}\right)_\tau - C^2\hat{\Delta}\Phi^{(R)}_0\right], \\ +\left[\left(2V\Phi^{(L)}_{0\mathscr{Z}} + \frac{D^2}{4}\Phi^{(L)}_{0\tilde{\tau}\tilde{\tau}\tilde{\tau}} + \frac{3|D|V}{2}\Phi^{(L)2}_{0\tilde{\tau}}\right)_{\tilde{\tau}} - C^2\hat{\Delta}\Phi^{(L)}_0\right], \end{aligned} \tag{58}$$

where

$$\hat{\Delta} = \partial^2_{\mathscr{X}} + \partial^2_{\mathscr{Y}}, \quad \hat{\Delta} = \frac{1}{\mathscr{R}}\partial_{\mathscr{R}}(\mathscr{R}\partial_{\mathscr{R}}) + \frac{1}{\mathscr{R}^2}\partial^2_{\Theta},$$

for the two geometries, respectively. Then, employing Eq. (47), the amplitude ρ_1 is again expressed as $\rho_1 = \rho_1^{(R)} + \rho_1^{(L)}$, with

$$\Phi^{(R)}_{0\tau} = \frac{C^2}{V}\rho_1^{(R)}, \quad \Phi^{(L)}_{0\tilde{\tau}} = -\frac{C^2}{V}\rho_1^{(L)}. \tag{59}$$

Using Eqs. (59), we obtain from Eq. (58) the following equations for $\rho_1^{(R,L)}$:

$$\left(\rho^{(R)}_{1\mathscr{Z}} - \frac{D^2}{8V}\rho^{(R)}_{1\tau\tau\tau} + \frac{3V}{2}\rho_1^{(R)}\rho^{(R)}_{1\tau}\right)_\tau + \frac{V}{2|D|}\hat{\Delta}\rho_1^{(R)} = 0, \tag{60}$$

$$\left(\rho^{(L)}_{1\mathscr{X}} + \frac{D^2}{8V}\rho^{(L)}_{1\tilde{\tau}\tilde{\tau}\tilde{\tau}} - \frac{3V}{2}\rho^{(L)}_{1}\rho^{(L)}_{1\tilde{\tau}}\right)_{\tilde{\tau}} - \frac{V}{2|D|}\hat{\Delta}\rho^{(L)}_{1} = 0. \tag{61}$$

We conclude this section with the observation that all equations that were derived for $\rho_1^{(R,L)}$ are of the KP type, in both geometries. Below we elaborate more on these effective models, and focus on limiting cases corresponding to their lower-dimensional versions. For simplicity, we only consider the right-going waves, $\rho^{(R)}_{1\mathscr{X}}$, since $\rho_1^{(L)}(\mathscr{X}) = \rho_1^{(R)}(-\mathscr{X})$. In addition, we will present examples of solitary wave solutions of Eqs. (33) and (34) arising from these KP models.

5 Versions of the KP Equations and Solitary Waves

5.1 Classification of the Effective KP Models

First of all, it is convenient to further normalize the effective KP equations derived in the previous section in order to express them in their "standard" form [2, 29].

Consider, first, equations for spatial solitary waves, and introduce the transformations:

$$\mathscr{X} \to -\frac{\alpha}{8C}\mathscr{X}, \quad \mathscr{T} \to \sqrt{\frac{3|\alpha|}{4V^2}}\mathscr{T},$$

as well as

$$\mathscr{Y} \to \sqrt{\frac{3|\alpha|}{4C^2}}\mathscr{Y}, \quad \rho_1^{(R)} = -\frac{\alpha}{2C^2}U \quad \text{and} \quad \Theta \to \sqrt{\frac{3|\alpha|}{4}}\Theta, \quad \rho_1^{(R)} = -\frac{\alpha}{2C^2}W,$$

for the Cartesian and cylindrical geometry, respectively. This way, Eq. (54) is expressed as:

$$\left(U_{\mathscr{X}} + 6UU_{\chi} + U_{\chi\chi\chi}\right)_{\chi} + 3\sigma^2\left(U_{\mathscr{Y}\mathscr{Y}} + U_{\mathscr{T}\mathscr{T}}\right) = 0, \tag{62}$$

while Eq. (56) reads:

$$\left(W_{\mathscr{X}} + 6WW_{\rho} + W_{\rho\rho\rho} + \frac{1}{2\mathscr{X}}W\right)_{\rho} + 3\sigma^2\left(\frac{1}{\mathscr{X}^2}W_{\Theta\Theta} + W_{\mathscr{T}\mathscr{T}}\right) = 0. \tag{63}$$

In the above equations, parameter σ^2 is given by

$$\sigma^2 = -\text{sign}\{\alpha\},$$

and it is reminded that α is given by Eq. (49).

The $(1+1)$-dimensional versions of Eqs. (62) and (63), i.e., the ones referring to the $\mathscr{X}\chi$ and $\mathscr{X}\rho$ plane, have respectively the form of a KdV and a cylindrical KdV (cKdV) equation. Both models are completely integrable by means of the IST [2], and find numerous applications in a variety of physical contexts [1, 29, 32, 64]. The KdV and cKdV equations have been derived by means of multiscale expansion methods from local NLS models, with the aim to describe shallow planar dark solitons in Bose gases [71] and ring dark solitons in nonlinear optical media [42] (see also reviews [22, 40] and references therein). More recently, a KdV equation was derived from the $(1+1)$-dimensional version of Eqs. (33) and (34) for $D=0$, and used to describe small-amplitude nematicons [26]; in fact, the KdV model derived in [26] is identical with the $(1+1)$-dimensional version of Eq. (54) [or (62)].

Furthermore, there are two distinct $(2+1)$-dimensional versions of Eq. (62): a spatial one, in the $\mathscr{X}\chi\mathscr{Y}$ space, and a spatio-temporal one, in the $\mathscr{X}\chi\mathscr{T}$ space. These effective models can be used to describe either spatial optical solitons in nematic liquid crystals [9], or dispersion-induced dynamics of spatial solitons in thermal media [45]. Both these $(2+1)$-dimensional equations, are completely integrable by means of the IST [2].

Importantly, the $(2+1)$-dimensional versions, as well as the complete Eq. (62), include both versions of the KP equation, KP-I and KP-II [2]. Indeed, for $\sigma = 1$, i.e., $\alpha < 0 \Rightarrow d > (q/2|u_0|)^2$, Eq. (62) is a KP-II equation; on the other hand, for $\sigma = i$, i.e., $\alpha > 0 \Rightarrow d < (q/2|u_0|)^2$, Eq. (62) is a KP-I equation. Recalling that d is the degree of nonlocality of the system at hand (for $d \to 0$ nonlocal NLS Eqs. (33) and (34) become local), it is evident that relatively weak (strong) nonlocality, as defined by the above regimes of d, corresponds to a KP-I (KP-II) model. This fact has also important implications on the type and the stability of low-dimensional solitary waves that can be supported in the system (see below).

Similarly, we observe that there are two distinct $(2+1)$-dimensional versions of Eq. (63): a spatial one, in the $\mathscr{X}\rho\Theta$ space, and a spatio-temporal one, in the $\mathscr{X}\rho\mathscr{T}$ space, which find applications in the contexts discussed above in the Cartesian case. The spatial version of Eq. (63) is a cylindrical KP (cKP) equation, which is also known as the Johnson's equation [31], and describes nearly-concentric solitons in an ideal, inviscid fluid [32]. This model is, also, completely integrable by means of the IST [58]. On the other hand, in the $\mathscr{X}\rho\mathscr{T}$ space, Eq. (56) reduces to the so-called CI equation, which describes weak cylindrical ion-acoustic solitons in plasmas [29]. Unlike the Johnson's equation, the CI equation is not considered to be integrable, as it fails to pass the Painlevé test [2].

It is interesting to point out that there exist transformations mapping solutions of the KP and cKP equations [32]. Indeed, the map:

$$U(\mathscr{X},\ \chi,\ \mathscr{Y}) \to W(\mathscr{X},\ \rho,\ \Theta) := U\left(\mathscr{X},\ \rho - \frac{\mathscr{X}\Theta^2}{12\sigma^2},\ \mathscr{X}\Theta\right),$$

transforms any solution of the KP equation (62) into a solution of the cKP equation (63); conversely, the map:

$$W(\mathscr{X},\ \chi,\ \Theta) \to U(\mathscr{X},\ \chi,\ \mathscr{Y}) := W\left(\mathscr{X},\ \chi + \frac{\mathscr{Y}^2}{12\sigma^2\mathscr{X}},\ \frac{\mathscr{Y}}{\mathscr{X}}\right),$$

transforms any solution of the cKP equation (63) into a solution of the KP equation (62). Here, we should also note that the spatial (2+1)-dimensional versions of KP equation (62) and cKP equation (63) are also connected with another relevant model, the elliptic cKP (ecKP), that was recently presented and studied in [38]. In this work, it was shown that the ecKP model describes surface gravity waves of nearly elliptic fronts, and it is completely integrable. Based on the similarities of the hydrodynamic form (36)–(38) of the nonlocal NLS Eqs. (33) and (34) to the problem formulation of [38], we conjecture that, adopting an elliptic cylindrical coordinate system and following the lines of the analysis presented here, one could derive a $(3 + 1)$-dimensional version of the ecKP equation. Nevertheless, such a derivation is beyond the scope of the present work.

We now turn our attention to the KP models that describe temporal waves. As before, first we put Eq. (60) in the "standard" form. We thus introduce the transformations:

$$\mathscr{Z} \to -\frac{D^2}{8V}\mathscr{Z},\quad \{\mathscr{X},\ \mathscr{Y},\ \mathscr{R}\} \to \sqrt{\frac{3|D|^3}{4V^2}}\{\mathscr{X},\ \mathscr{Y},\ \mathscr{R}\},\quad \rho_1^{(R)} = -\frac{D^2}{2V^2}Q, \tag{64}$$

and obtain from Eq. (60) the models:

$$(Q_{\mathscr{Z}} + 6QQ_\tau + Q_{\tau\tau\tau})_\tau - 3\hat{\Delta}Q = 0, \tag{65}$$

and it is reminded that the Laplacian $\hat{\Delta}$ refers to either the Cartesian or the cylindrical geometry. Obviously, the $(1 + 1)$-dimensional version of Eq. (65) is the KdV equation. On the other hand, it is observed that, unlike the case of spatial solitary waves, the Cartesian version of Eq. (65) is solely of the KP-I type; in fact, in this case, transverse effects are not governed by the sign of parameter α. The $(2 + 1)$-dimensional version of Eq. (65) is completely integrable by means of the IST [2]. Finally, the cylindrical version of Eq. (65) is known as the CII equation, and describes cylindrical ion-acoustic solitons in plasmas [29].

5.2 *Solitary Wave Solutions*

The asymptotic reduction of the nonlocal NLS equations to the effective equations above, allows for the derivation of approximate solutions of Eqs. (33) and (34), valid up to—and including—order $\mathscr{O}(\varepsilon)$. Of particular interest are solitary wave solutions, which can be constructed from solutions of Eqs. (62), (63) and (65). These asymptotic reductions provide information on the type of the solitary wave, as well as on the stability of lower-dimensional solutions in higher-dimensional settings. Below, we showcase some characteristic examples along those lines.

Let us first consider the case of spatial solitary waves. The $(1+1)$-dimensional version of Eq. (62) is a KdV equation which possesses the commonly known soliton solution:

$$U = 2\kappa^2 \mathrm{sech}^2[\kappa(\chi - 4\kappa^2 \mathscr{Z} - \chi_0)], \tag{66}$$

where κ and χ_0 are constants. Using this solution, and reverting transformations for the independent variables and fields, we find the following approximate solution to Eqs. (33) and (34):

$$u \approx u_0 \left[1 - \frac{\varepsilon\kappa^2}{C^2}\alpha \mathrm{sech}^2(\xi)\right] \exp\left[-\frac{2i}{q}|u_0|^2 z - i\frac{\varepsilon^{1/2}\kappa}{C}\alpha \tanh(\xi)\right], \tag{67}$$

$$n \approx \frac{1}{q} + \frac{1}{2}\varepsilon\kappa^2\alpha \mathrm{sech}^2(\xi), \tag{68}$$

$$\xi \equiv \varepsilon^{1/2}\kappa(x - \upsilon_{\mathrm{s}} z - x_0), \quad \upsilon_{\mathrm{s}} \equiv C\left(1 - \frac{1}{2}\frac{\varepsilon\eta^2}{C^2}|\alpha|\right). \tag{69}$$

The solution for u has the form of either a density dip (for $\alpha > 0$) or a density hump (for $\alpha < 0$) on top of the cw background, with a tanh-shaped phase jump across the density minimum or maximum, respectively. It is thus either a dark soliton (for $\alpha > 0$) or an anti-dark soliton (for $\alpha < 0$); note that the soliton velocity υ_{s} is slightly below the speed of sound, as is the case of shallow dark solitons in local media [22, 40]. Note that if $d \to 0$, then $\alpha > 0$, which means that in the case of the local system the soliton is always dark. In other words, anti-dark solitons are only supported due to the presence of nonlocality, in accordance with the analysis of [26].

In Fig. 11, we depict the soliton solutions in Cartesian geometry according to Eq. (66). Here, solutions' profiles are plotted at $z = 0$; all parameter values are kept equal to unity, and we vary parameter q so that to obtain a dark and an anti-dark soliton, for $q = 1$ and $q = 5$, respectively. Furthermore, we use these profiles as initial conditions, and perform a direct numerical integration of Eqs. (33) and (34) to determine their evolution. For the simulations, we used a high accuracy spectral integrator in Cartesian coordinates. The results are shown in the contour plots of Fig. 12, where it is verified that these solutions maintain their stability—at least for relatively short propagation distances (see discussion below)—and propagating characteristics. Notice that, as expected from the analysis, the anti-dark soliton propagates at higher, though constant, velocity from its dark soliton counterpart.

The fact that Eq. (62) is either a KP-I (for $\alpha < 0$) or a KP-II (for $\alpha > 0$) equation, can be used to deduce stability of the approximate solitons in $(2+1)$-dimensions. Indeed, as is well known [2], line soliton solutions of KP-I are unstable, while those of KP-II are stable. This leads to the prediction that, in the context of the original problem, dark soliton stripes of the nonlocal problem will be unstable in the 2D setting, while anti-dark soliton stripes will be stable. Note that the instability in

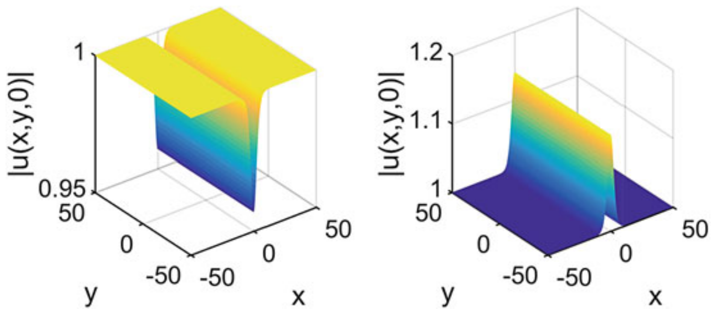


Fig. 11 Typical dark (left) and anti-dark soliton (right) profiles, at $z = 0$, in Cartesian geometry, for $q = 1$ and $q = 5$, respectively. All other parameter values are equal to unity

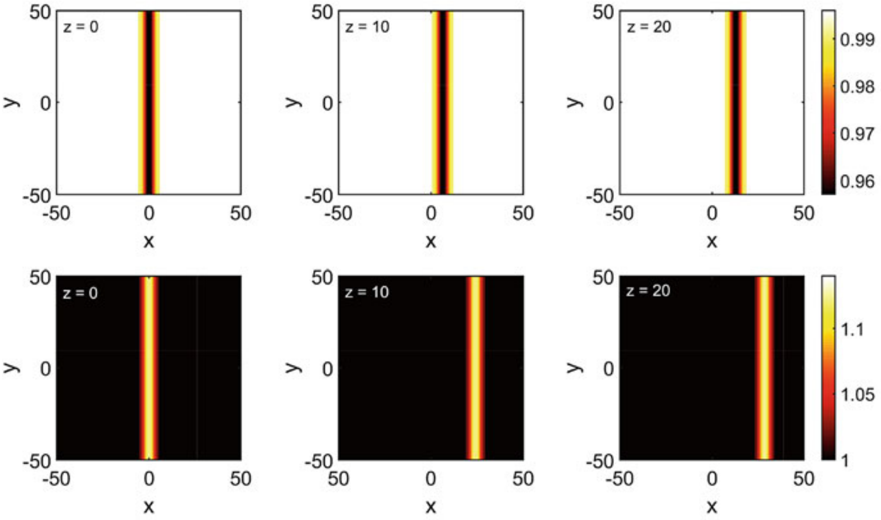


Fig. 12 Contour plots showing the evolution of the dark (top) and anti-dark (bottom) solitons of Fig. 11. Results have been obtained from direct numerical integration of Eqs. (33) and (34)

the context of the KP-I model was analyzed [61] and connected to the context of self-focusing and transverse instability of plane dark solitons in media with local defocusing nonlinearity [47, 62] (see also [41] for a review and references therein). It should also be mentioned that in the case of KP-I (for $\alpha < 0$), there exist "lump" solitons which are stable in the 2D setting [2]; these structures can be used to construct approximate solutions of the original problem which, in our case, will be 2D dark solitary waves, featuring an algebraic decay. These "lump" solitons, however (along with the planar ones discussed above), are unstable in the full $(3 + 1)$-dimensional setting [46].

We now turn to the case of the cylindrical geometry, and consider the $(1 + 1)$-dimensional version of Eq. (63), namely the cKdV equation. As mentioned above, this model is completely integrable by means of the IST. The solitary wave solution, which is expressed in terms of the Airy function [25], is composed of a primary wave and a shelf. An asymptotic analysis [33, 43] in the regime $|\mathscr{Z}| \gg |\rho|$ shows the following: to leading-order approximation, the primary wave $W(\rho, \mathscr{Z})$, that decays to zero at both upstream and downstream infinity, has a form similar to that of Eq. (66), with the obvious changes $\chi \to \rho$ and $\chi_0 \to \rho_0$, but with an important difference: κ now becomes a slowly-varying function of $\mathscr{Z}$, due to the presence of the term $W/(2\mathscr{Z})$. In fact, according to the analysis of Refs. [33, 43], and using the original coordinates, the following result can be obtained,

$$\kappa^2 = \kappa_0^2 \left(\frac{z_0}{z}\right)^{2/3}, \tag{70}$$

where κ_0^2 is a constant setting the solitary wave amplitude at $z = z_0$. Then, it is straightforward to express an approximate solution of Eqs. (33) and (34), but now for the cylindrical geometry, and for the primary solitary wave. This is of the form of Eqs. (67)–(69), but with the solitary wave amplitude and velocity varying as $z^{-2/3}$, and the width varying as $z^{1/3}$, as follows from Eqs. (66) and (70).

Obviously, this approximate solution is a ring-shaped solitary wave, on top of the cw background, which is either of the dark type (for $\alpha > 0$) or of the anti-dark type (for $\alpha < 0$). Note that ring dark solitons were predicted to occur in optical media exhibiting either Kerr [42] or non-Kerr [23] nonlinearities, and were later observed in experiments [18]. On the other hand, ring anti-dark solitons were only predicted to occur in non-Kerr—e.g., saturable media [23, 57]. This picture is complemented by our analysis, according to which a relatively strong [i.e., $d > (q/2|u_0|)^2$] nonlocal nonlinearity can also support ring anti-dark solitary waves.

In Fig. 13, typical ring dark and anti-dark soliton profiles, with parameter values identical to those used in the Cartesian case, are shown at $z = 0$; both solitons have an initial radius of $r_0 = 10$. In addition, in Fig. 14, contour plots depicting the evolution of the solitons' densities are shown; these results, as before, have been obtained via direct numerical integration of Eqs. (33) and (34). Much like the Cartesian case, the solitons propagate undistorted, i.e., the initial rings expand outwards, keeping their shapes during the evolution—at least for relatively short propagation distances (see below). It is also observed that the solitons expand (propagate) with constant speed, with the anti-dark soliton expanding faster than the dark soliton: indeed, the anti-dark soliton's radius is larger than that of the dark one, at the same propagation distance.

As in the Cartesian case, the effective equation (63) can also be used to predict (in)stability of the ring dark or anti-dark solitary waves in the $(2+1)$-dimensional setting. In particular, and similarly to the case of the planar solitons of Eq. (62), the case of $\alpha < 0$ ($\alpha > 0$), where ring dark (anti-dark) solitary waves exist, corresponds to a KP-I (KP-II) type model. It is, thus, clear that ring dark solitary waves are expected to be unstable, while ring anti-dark ones are predicted be stable.

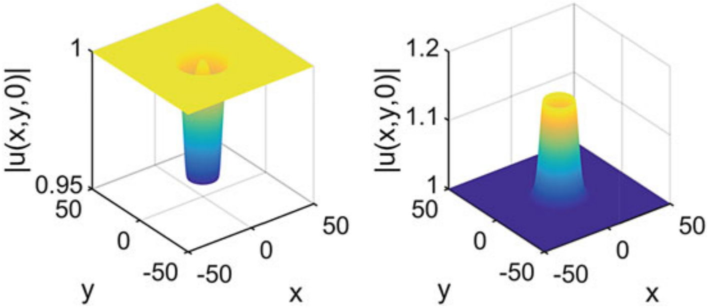


Fig. 13 Typical ring dark (left) and anti-dark soliton profiles, at $z = 0$. Both solitons have an initial radius $r_0 = 10$, while other parameter values are as in the Cartesian case

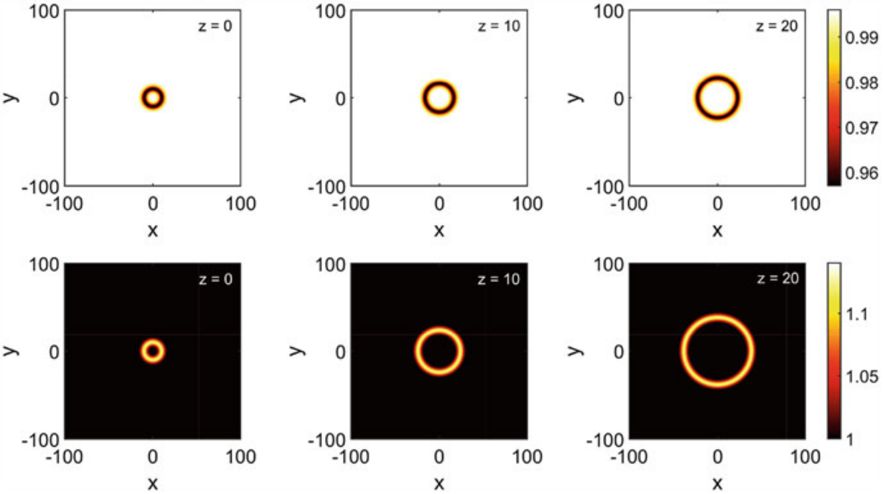


Fig. 14 Contour plots showing the evolution of the ring dark (top) and anti-dark (bottom) ring solitons of Fig. 13. Results have been obtained from direct numerical integration of Eqs. (33) and (34)

Finally, let us briefly discuss the case of temporal solitary waves described by Eq. (65). In the $(1 + 1)$-dimensional setting, the underlying KdV equation has a soliton solution similar to that in Eq. (66), with the obvious changes $\chi \rightarrow \tau$ and $\chi_0 \rightarrow \tau_0$. However, when expressed in terms of the original fields and variables of Eqs. (33) and (34), it is clear that the corresponding approximate solitary wave solution is solely of the dark type: this is due to the fact that parameter α is not involved in the normalization of the field Q [cf. Eq. (64)]. For the same reason, as was also mentioned in the previous section, the higher-dimensional versions of Eq. (65) are solely of the KP-I type. As a result, in the Cartesian $(2+1)$-dimensional setting corresponding to the usual KP-I model, one expects the existence of stable dark "lump" solitary wave solutions of the original model; these, however, are unstable in the full 3D setting [46]. Finally, regarding the cylindrical geometry, to the best of our knowledge, two-dimensional soliton solutions of the CII model are not known.

6 Summary and Conclusions

In this chapter we have studied the properties of a nonlocal nonlinear Schrödinger (NLS) model and focused, more specifically, on its soliton solutions. The considered nonlocal NLS is relevant to many physical contexts, as it describes the dynamics of optical beams in nematic liquid crystals, plasmas, and optical media exhibiting thermal nonlinearities. Generally, this nonlocal NLS consists of one (or more—in the vector version of the model) paraxial wave equation, describing the evolution of the optical field, coupled with a diffusion-type equation for the medium's effective refractive index.

We have studied various versions of this model: a $(1 + 1)$-dimensional scalar model, its vector generalization, as well as a scalar, fully $(3 + 1)$-dimensional model. We have also considered both the focusing and defocusing versions of the nonlocal NLS system. Our analysis started with the elementary solution—in the form of a continuous-wave (cw)—and the study of its stability. We have shown that in the focusing (defocusing) version of the model the cw is modulationally unstable (stable), and we have determined the relevant instability band and maximum growth rate. We found that the instability band and growth rate decrease due to nonlocality, which is a generic feature of nonlocal media.

Then, we have studied soliton solutions of the model. In the focusing regime, we have found bright soliton solutions in a closed analytical form, for both the scalar and the vector versions of the nonlocal system. In the defocusing regime, we have used multiscale expansion methods to derive effective nonlinear evolution equations that describe a variety of approximate soliton solutions of the original model. This is particularly important since, generally, nonlocal systems—and also the particular system considered in our work—do not possess solutions in closed analytical form. Our findings are summarized as follows.

For the $(1+1)$-dimensional scalar model, we derived an effective KdV equation, which describes either dark or antidark solitons of the nonlocal NLS. These are obtained, respectively, for relatively weak or strong nonlocality, with the relevant regimes discriminated by the sign of a physically relevant parameter. In the vector version of the model, our analysis was performed for different boundary conditions: for nonvanishing conditions for both fields, and nonvanishing-vanishing conditions for each field. In the former case, we derived a KdV model, which describes dark-dark solitons solutions of the original problem. In the second case, we derived a Mel'nikov system—namely a KdV equation with a self-consistent source satisfying a time-independent Schrödinger equation; this system describes dark-antidark solitons of the original nonlocal NLS.

We have also studied the fully $(3 + 1)$-dimensional scalar version of the nonlocal NLS model in the defocusing regime (the focusing one is generally subject to collapse). In this case too, it was found that the elementary cw solution is modulationally stable, which allowed us to find approximate soliton solutions on top of his stable background. Using again multiple scale expansion techniques, we have found various effective nonlinear evolution equations describing a wealth of approximate soliton solutions of the original problem, both in Cartesian and cylindrical geometries.

This way, first we derived, at an intermediate stage of the asymptotic analysis, a 3D Boussinesq equation. Then, we considered two cases, corresponding to spatial or temporal structures and, upon introducing relevant scales and asymptotic expansions, we reduced the Boussinesq model to KP-type equations that govern right- and left-propagating waves. These models include various integrable and non-integrable equations at different dimensionalities and geometries, such as the KdV and the cKdV equation, the KP-I and KP-II equations, Johnson's equation, as well as the CI and CII equations. Furthermore, useful results were deduced on the type and the stability of lower-dimensional solitary waves in higher-dimensional settings. In that regard, we identified parameter regimes, corresponding to relatively weak or strong nonlocality, for which we predicted the existence and stability of various solitary waves. Thus, we predicted the existence of spatial, planar or cylindrical (ring-shaped), dark or anti-dark solitary waves, for weak or strong nonlocality, respectively, and that dark (anti-dark) solitary waves are unstable (stable) in the $(1 + 1)$-dimensional setting. Furthermore, our analysis suggested the existence of temporal solitary waves, which become unstable in higher dimensions. Regarding approximate two-dimensional solitary wave solutions, it was found that they may exist in the form of algebraically decaying dark "lumps", which satisfy effective KP-I models; such structures may be either of the spatial or temporal type and are supported in the weak nonlocality regime.

Our analytical predictions were also corroborated by results of direct numerical simulations. Indeed, we have used the analytical forms of the spatial soliton profiles, in both the 1D and 2D (Cartesian and cylindrical) settings, and studied their evolution stemming from the direct numerical integration of the original nonlocal NLS model. We have thus found that all types of solitons propagate undistorted, as per the effective nonlinear evolution equation proper, at least for short propagation distances. Notice that, even for longer propagation distances, instabilities were not observed in our simulations, which suggests that the solitons presented here have a good chance to be observed in experiments.

Our analysis suggests various interesting directions for future studies. For instance, it would be relevant to extend our considerations to nonlocal models with a higher number of components in the higher-dimensional setting. In that case, it would be important to identify vector solitary wave structures and vortices in these models, extending previous studies in media with local nonlinearity [37].

References

1. M.J. Ablowitz, *Nonlinear Dispersive Waves: Asymptotic Analysis and Solitons* (Cambridge University Press, Cambridge, 2011)
2. M.J. Ablowitz, P.A. Clarkson, *Solitons, Nonlinear Evolution Equations and Inverse Scattering* (Cambridge University Press, Cambridge, 1991)
3. M.J. Ablowitz, T.P. Horikis, Interacting nonlinear wave envelopes and rogue wave formation in deep water. Phys. Fluids **27**, 012107 (2015)
4. G.P. Agrawal, *Nonlinear Fiber Optics* (Academic, Cambridge, 2013)
5. M. Aguero, D.J. Frantzeskakis, P.G. Kevrekidis, Asymptotic reductions of two coupled (2+1)-dimensional nonlinear Schrödinger equations: application to Bose-Einstein condensates. J. Phys. A Math. Gen. **39**, 7705–7718 (2006)
6. A. Alberucci, G. Assanto, Modeling nematicon propagation. Mol. Cryst. Liq. Cryst. **572**, 2–12 (2013)
7. A. Alberucci, M. Peccianti, G. Assanto, A. Dyadyusha, M. Kaczmarek, Two-color vector solitons in nonlocal media. Phys. Rev. Lett. **97**, 153903 (2006)
8. A. Armaroli, S. Trillo, Suppression of transverse instabilities of dark solitons and their dispersive shock waves. Phys. Rev. A **80**, 053803 (2009)
9. G. Assanto, *Nematicons: Spatial Optical Solitons in Nematic Liquid Crystals* (Wiley, Hoboken, 2012)
10. G. Assanto, A.A. Minzoni, N.F. Smyth, Light self-localization in nematic liquid crystals: modelling solitons in nonlocal reorientational media. J. Nonlinear Opt. Phys. Mater. **18**, 657–691 (2009)
11. G. Assanto, N.F. Smyth, A.L. Worthy, Two-color, nonlocal vector solitary waves with angular momentum in nematic liquid crystals. Phys. Rev. A **78**, 013832 (2008)
12. O. Bang, W. Krolikowski, J. Wyller, J.J. Rasmussen, Collapse arrest and soliton stabilization in nonlocal nonlinear media. Phys. Rev. E **66**, 046619 (2002)
13. A.V. Buryak, P. Di Trapani, D.V. Skryabin, S. Trillo, Optical solitons due to quadratic nonlinearities: from basic physics to futuristic applications. Phys. Rep. **370**, 63–235 (2002)
14. R.M. Caplan, R. Carretero-González, P.G. Kevrekidis, B.A. Malomed, Existence, stability, and scattering of bright vortices in the cubic-quintic nonlinear Schrödinger equation. Math. Comput. Simulat. **82**, 1150–1171 (2012)
15. E.G. Charalampidis, P.G. Kevrekidis, D.J. Frantzeskakis, B.A. Malomed, Dark-bright solitons in coupled nonlinear Schrödinger equations with unequal dispersion coefficients. Phys. Rev. E **91**, 012924 (2015)
16. C. Conti, M. Peccianti, G. Assanto, Route to nonlocality and observation of accessible solitons. Phys. Rev. Lett. **91**, 073901 (2003)
17. T. Dauxois, M. Peyrard, *Physics of Solitons* (Cambridge University Press, Cambridge, 2006)
18. A. Dreischuh, D.N. Neshev, G.G. Paulus, F. Grasbon, H. Walther, Ring dark solitary waves: experiment versus theory. Phys. Rev. E **66**, 066611 (2002)
19. A. Dreischuh, D.N. Neshev, D.E. Petersen, O. Bang, W. Krolikowski, Observation of attraction between dark solitons. Phys. Rev. Lett. **96**, 043901 (2006)
20. G. El, N.F. Smyth, Radiating dispersive shock waves in non-local optical media. Proc. Roy. Soc. Lond. A **472**, 20150633 (2016)
21. D.J. Frantzeskakis, Vector solitons supported by the third-order dispersion. Phys. Lett. A **285**, 363–367 (2001)
22. D.J. Frantzeskakis, Dark solitons in Bose-Einstein condensates: from theory to experiments. J. Phys. A Math. Theor. **43**, 213001 (2010)
23. D.J. Frantzeskakis, B.A. Malomed, Multiscale expansions for a generalized cylindrical nonlinear Schrödinger equation. Phys. Lett. A **264**, 179–185 (1999)
24. A. Hasegawa, Y. Kodama, *Solitons in Optical Communications* (Claredon Press, Oxford, 1995)
25. R. Hirota, Exact solutions to the equation describing "cylindrical solitons". Phys. Lett. A **71**, 393–394 (1979)

26. T.P. Horikis, Small-amplitude defocusing nematicons. J. Phys. A Math. Theor. **48**, 02FT01 (2015)
27. T.P. Horikis, D.J. Frantzeskakis, On the NLS to KdV connection. Rom. J. Phys. **59**, 195–203 (2014)
28. T.P. Horikis, D.J. Frantzeskakis, Asymptotic reductions and solitons of nonlocal nonlinear Schrödinger equations. J. Phys. A Math. Theor. **49**, 205202 (2016)
29. E. Infeld, G. Rowlands, *Nonlinear Waves, Solitons and Chaos* (Cambridge University Press, Cambridge, 1990)
30. A. Jeffrey, T. Kawahara, *Asymptotic Methods in Nonlinear Wave Theory* (Pitman Books, London, 1982)
31. R.S. Johnson, Water waves and Korteweg-de Vries equations. J. Fluid Mech. **97**, 701–719 (1980)
32. R.S. Johnson, *A Modern Introduction to the Mathematical Theory of Water Waves* (Cambridge University Press, Cambridge, 1997)
33. R.S. Johnson, A note on an asymptotic solution of the cylindrical Korteweg-de Vries equation. Wave Motion **30**, 1–16 (1999)
34. Y.N. Karamzin, A.P. Sukhorukov, Nonlinear interaction of diffracted light beams in a medium with quadratic nonlinearity: mutual focusing of beams and limitation on the efficiency of optical frequency converters. JETP Lett. **20**, 339–342 (1974)
35. V.I. Karpman, *Non-linear Waves in Dispersive Media* (Elsevier, Amsterdam, 1974)
36. Y.V. Kartashov, L. Torner, Gray spatial solitons in nonlocal nonlinear media. Opt. Lett. **32**, 946–948 (2007)
37. P.G. Kevrekidis, D.J. Frantzeskakis, R. Carretero-González, *The Defocusing Nonlinear Schrödinger Equation: From Dark Solitons to Vortices and Vortex Rings* (SIAM, Philadelphia, 2015)
38. K.R. Khusnutdinova, C. Klein, V.B. Matveev, A.O. Smirnov, On the integrable elliptic cylindrical Kadomtsev-Petviashvili equation. Chaos **23**, 013126 (2013)
39. Y.S. Kivshar, G.P. Agrawal, *Optical Solitons: From Fibers to Photonic Crystals* (Academic, Cambridge, 2003)
40. Y.S. Kivshar, B. Luther-Davies, Dark optical solitons: physics and applications. Phys. Rep. **298**, 81–197 (1998)
41. Y.S. Kivshar, D.E. Pelinovsky, Self-focusing and transverse instabilities of solitary waves. Phys. Rep. **331**, 117–195 (2000)
42. Y.S. Kivshar, X. Yang, Ring dark solitons. Phys. Rev. E **50**, R40–R43 (1994)
43. K. Ko, H.H. Kuehl, Cylindrical and spherical Korteweg-deVries solitary waves. Phys. Fluids **22**, 1343–1348 (1979)
44. W. Krolikowski, O. Bang, J.J. Rasmussen, J. Wyller, Modulational instability in nonlocal nonlinear kerr media. Phys. Rev. E **64**, 016612 (2001)
45. W. Krolikowski, O. Bang, N.I. Nikolov, D.N. Neshev, J. Wyller, J.J. Rasmussen, D. Edmundson, Modulational instability, solitons and beam propagation in spatially nonlocal nonlinear media. J. Opt. B Quantum Semiclassical Opt. **6**, S288–S294 (2004)
46. E.A. Kuznetsov, S.K. Turitsyn, Two- and three-dimensional solitons in weakly dispersive media. J. Exp. Theor. Phys. **55**, 844–847 (1982)
47. E.A. Kuznetsov, S.K. Turitsyn, Instability and collapse of solitons in media with a defocusing nonlinearity. J. Exp. Theor. Phys. **67**, 1583–1588 (1988)
48. A.G. Litvak, V.A. Mironov, G.M. Fraiman, A.D. Yunakovskii, Thermal self-effect of wave beams in a plasma with a nonlocal nonlinearity. Sov. J. Plasma Phys. **1**, 60–71 (1975)
49. J.M.L. MacNeil, N.F. Smyth, G. Assanto, Exact and approximate solutions for optical solitary waves in nematic liquid crystals. Phys. D **284**, 1–15 (2014)
50. B.A. Malomed, V.I. Shrira, Soliton caustics. Phys. D **53**, 1–12 (1991)
51. V.K. Mel'nikov, On equations for wave interactions. Lett. Math. Phys. **7**, 129–136 (1983)
52. V.K. Mel'nikov, Exact solutions of the korteweg-de vries equation with a self-consistent source. Phys. Lett. A **128**, 488–492 (1988)

53. V.K. Mel'nikov, Integration method of the korteweg-de vries equation with a self-consistent source. Phys. Lett. A **133**, 493–496 (1988)
54. D. Mihalache, Multidimensional solitons and vortices in nonlocal noninear optical media. Rom. Rep. Phys. **59**, 515–522 (2007)
55. D. Mihalache, D. Mazilu, F. Lederer, B.A. Malomed, Y.V. Kartashov, L.-C. Crasovan, L. Torner, Three-dimensional spatiotemporal optical solitons in nonlocal nonlinear media. Phys. Rev. E **73**, 025601(R) (2006)
56. A. Minovich, D.N. Neshev, A. Dreischuh, W. Krolikowski, Y.S. Kivshar, Experimental reconstruction of nonlocal response of thermal nonlinear optical media. Opt. Lett. **32**, 1599–1601 (2007)
57. H.E. Nistazakis, D.J. Frantzeskakis, B.A. Malomed, P.G. Kevrekidis, Head-on collisions of ring dark solitons. Phys. Lett. A **285**, 157–164 (2001)
58. W. Oevel, W.-H. Steeb, Painleve analysis for a time-dependent Kadomtsev-Petviashvili equation. Phys. Lett. A **103**, 239–242 (1984)
59. M. Peccianti, G. Assanto, Nematicons. Phys. Rep. **516**, 147–208 (2012)
60. P. Pedri, L. Santos, Two-dimensional bright solitons in dipolar Bose-Einstein condensates. Phys. Rev. Lett. **95**, 200404 (2005)
61. D.E. Pelinovsky, Y.A. Stepanyants, Solitary wave instability in the positive-dispersion media described by the two-dimensional Boussinesq equations. J. Exp. Theor. Phys. **79**, 105–112 (1993)
62. D.E. Pelinovsky, Y.A. Stepanyants, Y.S. Kivshar, Self-focusing of plane dark solitons in nonlinear defocusing media. Phys. Rev. E **51**, 5016–5026 (1995)
63. A. Piccardi, A. Alberucci, N. Tabiryan, G. Assanto, Dark nematicons. Opt. Lett. **36**, 1356–1358 (2011)
64. M. Remoissenet, *Waves Called Solitons* (Springer, Berlin, 1999)
65. C. Rotschild, T. Carmon, O. Cohen, O. Manela, M. Segev, Solitons in nonlinear media with an infinite range of nonlocality: first observation of coherent elliptic solitons and of vortex-ring solitons. Phys. Rev. Lett. **95**, 213904 (2005)
66. B.D. Skuse, N.F. Smyth, Two-color vector-soliton interactions in nematic liquid crystals in the local response regime. Phys. Rev. A **77**, 013817 (2008)
67. B.D. Skuse, N.F. Smyth, Interaction of two-color solitary waves in a liquid crystal in the nonlocal regime. Phys. Rev. A **79**, 063806 (2009)
68. J.C.F. Sturm, Mémoire sur la résolution des équations numériques. Bull. Sci. Férussac **11**, 419–425 (1829)
69. D. Suter, T. Blasberg, Stabilization of transverse solitary waves by a nonlocal response of the nonlinear medium. Phys. Rev. A **48**, 4583–4587 (1993)
70. F. Tsitoura, V. Achilleos, B.A. Malomed, D. Yan, P.G. Kevrekidis, D.J. Frantzeskakis. Matter-wave solitons in the counterflow of two immiscible superfluids. Phys. Rev. A **87**, 063624 (2013)
71. T. Tsuzuki, Nonlinear waves in the Pitaevskii-Gross equation. J. Low Temp. Phys. **4**, 441–457 (1971)
72. S.K. Turitsyn, Spatial dispersion of nonlinearity and stability of many dimensional solitons. Theor. Math. Phys. **64**, 797–801 (1985)
73. A.I. Yakimenko, Y.A. Zaliznyak, Y.S. Kivshar, Stable vortex solitons in nonlocal self-focusing nonlinear media. Phys. Rev. E **71**, 065603(R) (2005)
74. V.E. Zakharov, E.A. Kuznetsov, Multi-scale expansions in the theory of systems integrable by the inverse scattering transform. Phys. D **18**, 455–463 (1986)

Stability of a Cauchy-Jensen Additive Mapping in Various Normed Spaces

Hassan Azadi Kenary, Choonkil Park, Themistocles M. Rassias, and Jung Rye Lee

1 Introduction

A classical question in the theory of functional equations is the following: *When is it true that a function which approximately satisfies a functional equation must be close to an exact solution of the equation?* If the problem accepts a solution, we say that the equation is *stable*. The first stability problem concerning group homomorphisms was raised by Ulam [57] in 1940. In the next year, Hyers [22] gave a positive answer to the above question for additive groups under the assumption that the groups are Banach spaces. In 1978, Rassias [43] proved a generalization of Hyers's theorem for additive mappings.

Theorem 1 ([43]) *Let $f : E \to E'$ be a mapping from a normed vector space E into a Banach space E' subject to the inequality*

$$\|f(x+y) - f(x) - f(y)\| \leq \varepsilon(\|x\|^p + \|y\|^p)$$

H. Azadi Kenary
Department of Mathematics, College of Sciences, Yasouj University, Yasuj, Iran
e-mail: azadi@mail.yu.ac.ir

C. Park (✉)
Research Institute for Natural Sciences, Hanyang University, Seoul, South Korea
e-mail: baak@hanyang.ac.kr

Th. M. Rassias
Department of Mathematics, National Technical University of Athens, Athens, Greece
e-mail: trassias@math.ntua.gr

J. R. Lee
Department of Mathematics, Daejin University, Pocheon, South Korea
e-mail: jrlee@daejin.ac.kr

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_15

for all $x, y \in E$, where ε and p are constants with $\varepsilon > 0$ and $0 \leq p < 1$. Then the limit $L(x) = \lim_{n\to\infty} \frac{f(2^n x)}{2^n}$ exists for all $x \in E$ and $L : E \to E'$ is the unique linear mapping which satisfies

$$\|f(x) - L(x)\| \leq \frac{2\varepsilon}{2 - 2^p}\|x\|^p$$

for all $x \in E$. Also, if for each $x \in E$ the function $f(tx)$ is continuous in $t \in R$, then L is linear.

Furthermore, in 1994, a generalization of Rassias' theorem was obtained by Găvruta [20] by replacing the bound $\varepsilon(\|x\|^p + \|y\|^p)$ by a general control function $\varphi(x, y)$. In 1983, a Hyers-Ulam stability problem for the quadratic functional equation was proved by Skof [56] for mappings $f : X \to Y$, where X is a normed space and Y is a Banach space. In 1984, Cholewa [9] noticed that the theorem of Skof is still true if the relevant domain X is replaced by an Abelian group and, in 2002, Czerwik [11] proved the Hyers-Ulam stability of the quadratic functional equation. The reader is referred to [1–54] and references therein for detailed information on stability of functional equations.

In 1897, Hensel [21] introduced a normed space which does not have the Archimedean property. It turned out that non-Archimedean spaces have many nice applications (see [12, 27, 30, 31, 36]).

Katsaras [26] defined a fuzzy norm on a vector space to construct a fuzzy vector topological structure on the space. Some mathematicians have defined fuzzy norms on a vector space from various points of view (see [19, 32, 40]). In particular, Bag and Samanta [3], following Cheng and Mordeson [8], gave an idea of fuzzy norm in such a manner that the corresponding fuzzy metric is of Karmosil and Michalek type [25]. They established a decomposition theorem of a fuzzy norm into a family of crisp norms and investigated some properties of fuzzy normed spaces [4].

Definition 1 By a *non-Archimedean field* we mean a field K equipped with a function (valuation) $|\cdot| : K \to [0, \infty)$ such that, for all $r, s \in K$, the following conditions hold:

(a) $|r| = 0$ if and only if $r = 0$;
(b) $|rs| = |r||s|$;
(c) $|r + s| \leq \max\{|r|, |s|\}$.

Clearly, by (b), $|1| = |-1| = 1$ and so, by induction, it follows from (c) that $|n| \leq 1$ for all $n \geq 1$.

Definition 2 Let X be a vector space over a scalar field K with a non-Archimedean non-trivial valuation $|\cdot|$.

(1) A function $\|\cdot\| : X \to R$ is a *non-Archimedean norm* (valuation) if it satisfies the following conditions:

(a) $\|x\| = 0$ if and only if $x = 0$ for all $x \in X$;
(b) $\|rx\| = |r|\|x\|$ for all $r \in K$ and $x \in X$;

(*c*) the strong triangle inequality (ultra-metric) holds, that is,

$$\|x+y\| \leq \max\{\|x\|, \|y\|\}$$

for all $x, y \in X$.

(2) The space $(X, \|\cdot\|)$ is called a *non-Archimedean normed space*.

Note that $||x_n - x_m|| \leq max\{||x_{j+1} - x_j|| : m \leq j \leq n-1\}$ for all $m, n \in N$ with $n > m$.

Definition 3 Let $(X, \|\cdot\|)$ be a non-Archimedean normed space.

(*a*) A sequence $\{x_n\}$ is a *Cauchy sequence* in X if $\{x_{n+1} - x_n\}$ converges to zero in X.

(*b*) The non-Archimedean normed space $(X, \|\cdot\|)$ is said to be *complete* if every Cauchy sequence in X is convergent.

The most important examples of non-Archimedean spaces are p-adic numbers. A key property of p-adic numbers is that they do not satisfy the Archimedean axiom: for all $x, y > 0$, there exists a positive integer n such that $x < ny$.

Example 1 Fix a prime number p. For any nonzero rational number x, there exists a unique positive integer n_x such that $x = \frac{a}{b} p^{n_x}$, where a and b are positive integers not divisible by p. Then $|x|_p := p^{-n_x}$ defines a non-Archimedean norm on Q. The completion of Q with respect to the metric $d(x, y) = |x - y|_p$ is denoted by Q_p, which is called the *p-adic number field*. In fact, Q_p is the set of all formal series $x = \sum_{k \geq n_x}^{\infty} a_k p^k$, where $|a_k| \leq p-1$. The addition and multiplication between any two elements of Q_p are defined naturally. The norm $|\sum_{k \geq n_x}^{\infty} a_k p^k|_p = p^{-n_x}$ is a non-Archimedean norm on Q_p and Q_p is a locally compact filed.

In Sect. 3, we adopt the usual terminology, notions and conventions of the theory of random normed spaces as in [55].

Throughout this paper, let Δ^+ denote the set of all probability distribution functions $F : R \cup [-\infty, +\infty] \to [0, 1]$ such that F is left-continuous and nondecreasing on R and $F(0) = 0, F(+\infty) = 1$. It is clear that the set $D^+ = \{F \in \Delta^+ : l^- F(-\infty) = 1\}$, where $l^- F(x) = \lim_{t \to x^-} F(t)$, is a subset of Δ^+. The set Δ^+ is partially ordered by the usual point-wise ordering of functions, that is, $F \leq G$ if and only if $F(t) \leq G(t)$ for all $t \in R$. For any $a \geq 0$, the element $H_a(t)$ of D^+ is defined by

$$H_a(t) = \begin{cases} 0, \text{ if } t \leq a, \\ 1, \text{ if } t > a. \end{cases}$$

We can easily show that the maximal element in Δ^+ is the distribution function $H_0(t)$.

Definition 4 A function $T : [0,1]^2 \to [0,1]$ is a *continuous triangular norm* (briefly, a t-norm) if T satisfies the following conditions:

(a) T is commutative and associative;
(b) T is continuous;
(c) $T(x,1) = x$ for all $x \in [0,1]$;
(d) $T(x,y) \leq T(z,w)$ whenever $x \leq z$ and $y \leq w$ for all $x, y, z, w \in [0,1]$.

Three typical examples of continuous t-norms are as follows: $T(x,y) = xy$, $T(x,y) = \max\{a+b-1, 0\}$, $T(x,y) = \min(a,b)$. Recall that, if T is a t-norm and $\{x_n\}$ is a sequence in $[0,1]$, then $T_{i=1}^n x_i$ is defined recursively by $T_{i=1}^1 x_1 = x_1$ and $T_{i=1}^n x_i = T(T_{i=1}^{n-1} x_i, x_n)$ for all $n \geq 2$. $T_{i=n}^\infty x_i$ is defined by $T_{i=1}^\infty x_{n+i}$.

Definition 5 A *random normed space* (briefly, RN-space) is a triple (X, μ, T), where X is a vector space, T is a continuous t-norm and $\mu : X \to D^+$ is a mapping such that the following conditions hold:

(a) $\mu_x(t) = H_0(t)$ for all $t > 0$ if and only if $x = 0$;
(b) $\mu_{\alpha x}(t) = \mu_x\left(\frac{t}{|\alpha|}\right)$ for all $\alpha \in R$ with $\alpha \neq 0$, $x \in X$ and $t \geq 0$;
(c) $\mu_{x+y}(t+s) \geq T(\mu_x(t), \mu_y(s))$ for all $x, y \in X$ and $t, s \geq 0$.

Every normed space $(X, \|\cdot\|)$ defines a random normed space (X, μ, T_M), where $\mu_u(t) = \frac{t}{t+\|u\|}$ for all $t > 0$ and T_M is the minimum t-norm. This space X is called the *induced random normed space*.

If the t-norm T is such that $\sup_{0<a<1} T(a,a) = 1$, then every RN-space (X, μ, T) is a metrizable linear topological space with the topology τ (called the *μ-topology* or the *(ε, δ)-topology*, where $\varepsilon > 0$ and $\lambda \in (0,1)$) induced by the base $\{U(\varepsilon, \lambda)\}$ of neighborhoods of θ, where

$$U(\varepsilon, \lambda) = \{x \in X : \mu_x(\varepsilon) > 1 - \lambda\}.$$

Definition 6 Let (X, μ, T) be an RN-space.

(a) A sequence $\{x_n\}$ in X is said to be *convergent* to a point $x \in X$ (write $x_n \to x$ as $n \to \infty$) if

$$\lim_{n\to\infty} \mu_{x_n - x}(t) = 1$$

for all $t > 0$.

(b) A sequence $\{x_n\}$ in X is called a *Cauchy sequence* in X if

$$\lim_{n\to\infty} \mu_{x_n - x_m}(t) = 1$$

for all $t > 0$.

(c) The RN-space (X, μ, T) is said to be *complete* if every Cauchy sequence in X is convergent.

Theorem 2 *If (X, μ, T) is an RN-space and $\{x_n\}$ is a sequence such that $x_n \to x$, then* $\lim_{n\to\infty} \mu_{x_n}(t) = \mu_x(t)$.

Definition 7 Let X be a real vector space. A function $N : X \times R \to [0, 1]$ is called a fuzzy norm on X if for all $x, y \in X$ and all $s, t \in R$,

(*N1*) $N(x, t) = 0$ for $t \leq 0$;
(*N2*) $x = 0$ if and only if $N(x, t) = 1$ for all $t > 0$;
(*N3*) $N(cx, t) = N\left(x, \frac{t}{|c|}\right)$ if $c \neq 0$;
(*N4*) $N(x + y, c + t) \geq min\{N(x, s), N(y, t)\}$;
(*N5*) $N(x, .)$ is a non-decreasing function of R and $\lim_{t\to\infty} N(x, t) = 1$;
(*N6*) for $x \neq 0$, $N(x, .)$ is continuous on R.

The pair (X, N) is called a fuzzy normed vector space.

Example 2 Let $(X, \|.\|)$ be a normed linear space and $\alpha, \beta > 0$. Then

$$N(x, t) = \begin{cases} \frac{\alpha t}{\alpha t + \beta\|x\|} & t > 0, x \in X \\ 0 & t \leq 0, x \in X \end{cases}$$

is a fuzzy norm on X.

Definition 8 Let (X, N) be a fuzzy normed vector space. A sequence $\{x_n\}$ in X is said to be convergent or converge if there exists an $x \in X$ such that $\lim_{t\to\infty} N(x_n - x, t) = 1$ for all $t > 0$. In this case, x is called the limit of the sequence $\{x_n\}$ in X and we denote it by $N - \lim_{t\to\infty} x_n = x$.

Definition 9 Let (X, N) be a fuzzy normed vector space. A sequence $\{x_n\}$ in X is called Cauchy if for each $\varepsilon > 0$ and each $t > 0$ there exists an $n_0 \in N$ such that for all $n \geq n_0$ and all $p > 0$, we have $N(x_{n+p} - x_n, t) > 1 - \varepsilon$.

It is well known that every convergent sequence in a fuzzy normed vector space is Cauchy. If each Cauchy sequence is convergent, then the fuzzy norm is said to be complete and the fuzzy normed vector space is called a fuzzy Banach space.

We say that a mapping $f : X \to Y$ between fuzzy normed vector spaces X and Y is continuous at a point $x \in X$ if for each sequence $\{x_n\}$ converging to $x_0 \in X$, then the sequence $\{f(x_n)\}$ converges to $f(x_0)$. If $f : X \to Y$ is continuous at each $x \in X$, then $f : X \to Y$ is said to be continuous on X.

Definition 10 Let X be a set. A function $d : X \times X \to [0, \infty]$ is called a generalized metric on X if d satisfies the following conditions:

(a) $d(x, y) = 0$ if and only if $x = y$ for all $x, y \in X$;
(b) $d(x, y) = d(y, x)$ for all $x, y \in X$;
(c) $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in X$.

Theorem 3 *Let (X,d) be a complete generalized metric space and $J : X \to X$ be a strictly contractive mapping with Lipschitz constant $L < 1$. Then, for all $x \in X$, either $d(J^n x, J^{n+1}x) = \infty$ for all nonnegative integers n or there exists a positive integer n_0 such that*

(a) *$d(J^n x, J^{n+1}x) < \infty$ for all $n_0 \geq n_0$;*
(b) *the sequence $\{J^n x\}$ converges to a fixed point y^* of J;*
(c) *y^* is the unique fixed point of J in the set $Y = \{y \in X : d(J^{n_0}x, y) < \infty\}$;*
(d) *$d(y, y^*) \leq \frac{d(y,Jy)}{1-L}$ for all $y \in Y$.*

2 Non-Archimedean Stability of the Functional Equation (1)

In this section, we deal with the stability problem for the Cauchy-Jensen additive functional equation (1) in non-Archimedean normed spaces.

2.1 Fixed Point Method

Theorem 4 *Let X is a non-Archimedean normed space and that Y be a complete non-Archimedean space. Let $\varphi : X^3 \to [0, \infty)$ be a function such that there exists an $\alpha < 1$ with*

$$\varphi\left(\frac{x}{2}, \frac{y}{2}, \frac{z}{2}\right) \leq \frac{\alpha\varphi\,(x, y, z)}{|2|} \tag{1}$$

for all $x, y, z \in X$. Let $f : X \to Y$ be a mapping with $f(0) = 0$ satisfying

$$\left\| f\left(\frac{x+y+z}{2}\right) + f\left(\frac{x-y+z}{2}\right) - f(x) - f(z)\right\|_Y \leq \varphi(x, y, z) \tag{2}$$

for all $x, y, z \in X$. Then there exists a unique additive mapping $L : X \to Y$ such that

$$\|f(x) - L(x)\|_Y \leq \frac{\alpha\varphi(x, 2x, x)}{|2| - |2|\alpha} \tag{3}$$

for all $x \in X$.

Proof Putting $y = 2x$ and $z = x$ in (2), we get

$$\|f(2x) - 2f(x)\|_Y \leq \varphi(x, 2x, x) \tag{4}$$

for all $x \in X$. So

$$\left\| f(x) - 2f\left(\frac{x}{2}\right)\right\|_Y \leq \varphi\left(\frac{x}{2}, x, \frac{x}{2}\right) \leq \frac{\alpha\varphi(x, 2x, x)}{|2|} \tag{5}$$

for all $x \in X$. Consider the set $S := \{h : X \to Y\}$ and introduce the generalized metric on S:

$$d(g,h) = \inf\Big\{\mu \in (0,+\infty) : \|g(x) - h(x)\|_Y \le \mu\varphi(x,2x,x), \ \ \forall x \in X\Big\},$$

where, as usual, $\inf\phi = +\infty$. It is easy to show that (S,d) is complete (see [33]). Now we consider the linear mapping $J : S \to S$ such that

$$Jg(x) := 2g\left(\frac{x}{2}\right)$$

for all $x \in X$. Let $g, h \in S$ be given such that $d(g,h) = \varepsilon$. Then

$$\|g(x) - h(x)\|_Y \le \varepsilon\varphi(x,2x,x)$$

for all $x \in X$. Hence

$$\begin{aligned}\|Jg(x) - Jh(x)\|_Y &= \left\|2g\left(\frac{x}{2}\right) - 2h\left(\frac{x}{2}\right)\right\|_Y = |2|\left\|g\left(\frac{x}{2}\right) - h\left(\frac{x}{2}\right)\right\|_Y \\ &\le |2|\varphi\left(\frac{x}{2}, x, \frac{x}{2}\right) \le \alpha\cdot\varepsilon\varphi(x,2x,x)\end{aligned}$$

for all $x \in X$. So $d(g,h) = \varepsilon$ implies that $d(Jg,Jh) \le \alpha\varepsilon$. This means that $d(Jg,Jh) \le \alpha d(g,h)$ for all $g,h \in S$. It follows from (5) that $d(f,Jf) \le \frac{\alpha}{|2|}$.

By Theorem 3, there exists a mapping $L : X \to Y$ satisfying the following:

(1) L is a fixed point of J, i.e.,

$$\frac{L(x)}{2} = L\left(\frac{x}{2}\right) \tag{6}$$

for all $x \in X$. The mapping L is a unique fixed point of J in the set $M = \{g \in S : d(h,g) < \infty\}$. This implies that L is a unique mapping satisfying (6) such that there exists a $\mu \in (0,\infty)$ satisfying $\|f(x) - L(x)\|_Y \le \mu\varphi(x,2x,x)$ for all $x \in X$;

(2) $d(J^n f, L) \to 0$ as $n \to \infty$. This implies the equality

$$\lim_{n\to\infty} 2^n f\left(\frac{x}{2^n}\right) = L(x) \tag{7}$$

for all $x \in X$;

(3) $d(f,L) \le \frac{1}{1-\alpha}d(f,Jf)$, which implies the inequality $d(f,L) \le \frac{\alpha}{|2|-|2|\alpha}$. This implies that the inequalities (3) holds.

It follows from (2) and (6) that

$$\begin{aligned}&\left\|L\left(\frac{x+y+z}{2}\right)+L\left(\frac{x-y+z}{2}\right)-L(x)-L(z)\right\|_Y\\&=\lim_{n\to\infty}|2|^n\left\|f\left(\frac{x+y+z}{2^{n+1}}\right)+f\left(\frac{x-y+z}{2^{n+1}}\right)-f\left(\frac{x}{2^n}\right)-f\left(\frac{z}{2^n}\right)\right\|_Y\\&\le\lim_{n\to\infty}|2|^n\varphi\left(\frac{x}{2^n},\frac{y}{2^n},\frac{z}{2^n}\right)\le\lim_{n\to\infty}|2|^n.\frac{\alpha^n\varphi(x,y,z)}{|2|^n}=0\end{aligned}$$

for all $x, y, z \in X$. So

$$L\left(\frac{x+y+z}{2}\right)+L\left(\frac{x-y+z}{2}\right)=L(x)+L(z)$$

for all $x, y, z \in X$. Hence $L : X \to Y$ is a Cauchy-Jensen mapping. It follows from (1), (5) and (7) that

$$\begin{aligned}\left\|2L\left(\frac{x}{2}\right)-L(x)\right\|_Y&=\lim_{n\to\infty}|2|^n\left\|2f\left(\frac{x}{2^{n+1}}\right)-f\left(\frac{x}{2^n}\right)\right\|_Y\\&\le\lim_{n\to\infty}|2|^n\varphi\left(\frac{x}{2^{n+1}},\frac{x}{2^n},\frac{x}{2^{n+1}}\right)\le\lim_{n\to\infty}|2|^n.\frac{\alpha^n\varphi(x,2x,x)}{|2|^n}=0\end{aligned}$$

for all $x \in X$. So

$$2L\left(\frac{x}{2}\right)-L(x)=0$$

for all $x \in X$. Hence $L : X \to Y$ is additive and we get the desired result.

Corollary 1 *Let θ be a positive real number and r is a real number with $0 < r < 1$. Let $f : X \to Y$ be a mapping with $f(0) = 0$ satisfying*

$$\left\|f\left(\frac{x+y+z}{2}\right)+f\left(\frac{x-y+z}{2}\right)-f(x)-f(z)\right\|_Y\le\theta\left(\|x\|^r+\|y\|^r+\|z\|^r\right)$$

for all $x, y, z \in X$. Then there exists a unique additive mapping $L : X \to Y$ such that

$$\|f(x)-L(x)\|_Y\le\frac{|2|\theta(2+|2|^r)\|x\|^r}{|2|^{r+1}-|2|^2}$$

for all $x \in X$.

Proof The proof follows from Theorem 4 by taking

$$\varphi(x,y,z)=\left(\|x\|^r+\|y\|^r+\|z\|^r\right)$$

for all $x, y, z \in X$. Then we can choose $\alpha = |2|^{1-r}$ and we get the desired result.

Theorem 5 *Let X is a non-Archimedean normed space and that Y be a complete non-Archimedean space. Let $\varphi : X^3 \to [0, \infty)$ be a function such that there exists an $\alpha < 1$ with*

$$\varphi (x, y, z) \leq |2|\alpha\varphi \left(\frac{x}{2}, \frac{y}{2}, \frac{z}{2}\right)$$

for all $x, y, z \in X$. Let $f : X \to Y$ be a mapping with $f(0) = 0$ satisfying (2). Then there exists a unique additive mapping $L : X \to Y$ such that

$$\|f(x) - L(x)\|_Y \leq \frac{\varphi(x, 2x, x)}{|2| - |2|\alpha} \tag{8}$$

for all $x \in X$.

Proof Let (S, d) be the generalized metric space defined in the proof of Theorem 4. Now we consider the linear mapping $J : S \to S$ such that

$$Jg(x) := \frac{g(2x)}{2}$$

for all $x \in X$. Let $g, h \in S$ be given such that $d(g, h) = \varepsilon$. Then

$$\|g(x) - h(x)\|_Y \leq \varepsilon\varphi(x, 2x, x)$$

for all $x \in X$. Hence

$$\begin{aligned} \|Jg(x) - Jh(x)\|_Y &= \left\|\frac{g(2x)}{2} - \frac{h(2x)}{2}\right\|_Y = \frac{\|g(2x) - h(2x)\|_Y}{|2|} \\ &\leq \frac{\varphi (2x, 4x, 2x)}{|2|} \leq \frac{|2|\alpha \cdot \varepsilon\varphi(x, 2x, x)}{|2|} \end{aligned}$$

for all $x \in X$. So $d(g, h) = \varepsilon$ implies that $d(Jg, Jh) \leq \alpha\varepsilon$. This means that $d(Jg, Jh) \leq \alpha d(g, h)$ for all $g, h \in S$. It follows from (4) that $d(f, Jf) \leq \frac{1}{|2|}$.

By Theorem 3, there exists a mapping $L : X \to Y$ satisfying the following:

(1) L is a fixed point of J, i.e.,

$$L(2x) = 2L (x) \tag{9}$$

for all $x \in X$. The mapping L is a unique fixed point of J in the set $M = \{g \in S : d(h, g) < \infty\}$. This implies that L is a unique mapping satisfying (9) such that there exists a $\mu \in (0, \infty)$ satisfying $\|f(x) - L(x)\|_Y \leq \mu\varphi(x, 2x, x)$ for all $x \in X$;

(2) $d(J^n f, L) \to 0$ as $n \to \infty$. This implies the equality

$$\lim_{n\to\infty} \frac{f(2^n x)}{2^n} = L(x)$$

for all $x \in X$;

(3) $d(f, L) \leq \frac{1}{1-\alpha} d(f, Jf)$, which implies the inequality $d(f, L) \leq \frac{1}{|2|-|2|\alpha}$. This implies that (8) holds. The rest of the proof is similar to the proof of Theorem 4.

Corollary 2 *Let θ be a positive real number and r is a real number with $r > 1$. Let $f : X \to Y$ be a mapping with $f(0) = 0$ satisfying*

$$\left\| f\left(\frac{x+y+z}{2}\right) + f\left(\frac{x-y+z}{2}\right) - f(x) - f(z) \right\|_Y \leq \theta\left(\|x\|^r + \|y\|^r + \|z\|^r\right)$$

for all $x, y, z \in X$. Then there exists a unique Then there exists a unique additive mapping $L : X \to Y$ such that

$$\|f(x) - L(x)\|_Y \leq \frac{\theta(2 + |2|^r)\|x\|^r}{|2| - |2|^r}$$

for all $x \in X$.

Proof The proof follows from Theorem 5 by taking

$$\varphi(x, y, z) = \left(\|x\|^r + \|y\|^r + \|z\|^r\right)$$

for all $x, y, z \in X$. Then we can choose $\alpha = |2|^{r-1}$ and we get the desired result.

2.2 Direct Method

In this section, using direct method, we prove the Hyers-Ulam stability of the Cauchy-Jensen additive functional equation (1) in non-Archimedean spaces.

Theorem 6 *Let G is an additive semigroup and that X is a non-Archimedean Banach space. Assume that $\zeta : G^3 \to [0, +\infty)$ is a function such that*

$$\lim_{n\to\infty} |2|^n \zeta\left(\frac{x}{2^n}, \frac{y}{2^n}, \frac{z}{2^n}\right) = 0 \tag{10}$$

for all $x, y, z \in G$. Suppose that, for any $x \in G$, the limit

$$\Psi(x) = \lim_{n\to\infty} \max_{0\leq k<n} |2|^k \zeta\left(\frac{x}{2^{k+1}}, \frac{x}{2^k}, \frac{x}{2^{k+1}}\right) \tag{11}$$

exists and $f : G \to X$ is a mapping with $f(0) = 0$ satisfying

$$\left\| f\left(\frac{x+y+z}{2}\right) + f\left(\frac{x-y+z}{2}\right) - f(x) - f(z) \right\|_X \leq \zeta(x, y, z) \tag{12}$$

Then the limit $A(x) := \lim_{n\to\infty} 2^n f\left(\frac{x}{2^n}\right)$ exists for all $x \in G$ and defines an additive mapping $A : G \to X$ such that

$$\|f(x) - A(x)\| \leq \Psi(x). \tag{13}$$

Moreover, if

$$\lim_{j\to\infty} \lim_{n\to\infty} \max_{j\leq k<n+j} |2|^k \zeta\left(\frac{x}{2^{k+1}}, \frac{x}{2^k}, \frac{x}{2^{k+1}}\right) = 0$$

then A is the unique additive mapping satisfying (13).

Proof Putting $y = 2x$ and $z = x$ in (12), we get

$$\|f(2x) - 2f(x)\|_Y \leq \zeta(x, 2x, x) \tag{14}$$

for all $x \in G$. Replacing x by $\frac{x}{2^{n+1}}$ in (14), we obtain

$$\left\|2^{n+1} f\left(\frac{x}{2^{n+1}}\right) - 2^n f\left(\frac{x}{2^n}\right)\right\| \leq |2|^n \zeta\left(\frac{x}{2^{n+1}}, \frac{x}{2^n}, \frac{x}{2^{n+1}}\right). \tag{15}$$

Thus, it follows from (10) and (15) that the sequence $\left\{2^n f\left(\frac{x}{2^n}\right)\right\}_{n\geq 1}$ is a Cauchy sequence. Since X is complete, it follows that $\left\{2^n f\left(\frac{x}{2^n}\right)\right\}_{n\geq 1}$ is convergent. Set

$$A(x) := \lim_{n\to\infty} 2^n f\left(\frac{x}{2^n}\right). \tag{16}$$

By induction on n, one can show that

$$\left\|2^n f\left(\frac{x}{2^n}\right) - f(x)\right\| \leq \max\left\{|2|^k \zeta\left(\frac{x}{2^{k+1}}, \frac{x}{2^k}, \frac{x}{2^{k+1}}\right); 0 \leq k < n\right\} \tag{17}$$

for all $n \geq 1$ and $x \in G$. By taking $n \to \infty$ in (17) and using (11), one obtains (13). By (10), (12) and (16), we get

$$\begin{aligned}
&\left\|A\left(\frac{x+y+z}{2}\right) + A\left(\frac{x-y+z}{2}\right) - A(x) - A(z)\right\| \\
&= \lim_{n\to\infty} |2|^n \left\|f\left(\frac{x+y+z}{2^{n+1}}\right) + f\left(\frac{x-y+z}{2^{n+1}}\right) - f\left(\frac{x}{2^n}\right) - f\left(\frac{z}{2^n}\right)\right\| \\
&\leq \lim_{n\to\infty} |2|^n \zeta\left(\frac{x}{2^n}, \frac{y}{2^n}, \frac{z}{2^n}\right) = 0
\end{aligned}$$

for all $x, y, z \in X$. So

$$A\left(\frac{x+y+z}{2}\right)+A\left(\frac{x-y+z}{2}\right)=A(x)+A(z) \tag{18}$$

for all $x, y, z \in G$. Letting $y=0$ in (18), we get

$$2L\left(\frac{x+z}{2}\right)=L(x)+L(z) \tag{19}$$

for all $x, z \in G$. Since

$$L(0)=\lim_{n\to+\infty}2^n f\left(\frac{0}{2^n}\right)=\lim_{n\to+\infty}2^n f(0)=0,$$

by letting $y=2x$ and $z=x$ in (18), we get

$$A(2x)=2A(x)$$

for all $x \in G$. Replacing x by $2x$ and z by $2z$ in (19), we get

$$A(a+z)=A(x)+A(z)$$

for all $x, z \in G$. Hence $A : G \to X$ is additive.

To prove the uniqueness property of A, let L be another mapping satisfying (13). Then we have

$$\begin{aligned}&\left\|A(x)-L(x)\right\|_X\\&=\lim_{n\to\infty}|2|^n\left\|A\left(\frac{x}{2^n}\right)-L\left(\frac{x}{2^n}\right)\right\|_X\\&\le\lim_{k\to\infty}|2|^n\max\left\{\left\|A\left(\frac{x}{2^n}\right)-f\left(\frac{x}{2^n}\right)\right\|_X,\left\|f\left(\frac{x}{2^n}\right)-L\left(\frac{x}{2^n}\right)\right\|_X\right\}\\&\le\lim_{j\to\infty}\lim_{n\to\infty}\max_{j\le k<n+j}|2|^k\zeta\left(\frac{x}{2^{k+1}},\frac{x}{2^k},\frac{x}{2^{k+1}}\right)=0\end{aligned}$$

for all $x \in G$. Therefore, $A=L$. This completes the proof.

Corollary 3 *Let $\xi : [0,\infty) \to [0,\infty)$ be a function satisfying*

$$\xi\left(\frac{t}{|2|}\right)\le\xi\left(\frac{1}{|2|}\right)\xi(t),\quad \xi\left(\frac{1}{|2|}\right)<\frac{1}{|2|}$$

for all $t \ge 0$. Assume that $\kappa > 0$ and $f : G \to X$ is a mapping with $f(0)=0$ such that

$$\left\|f\left(\frac{x+y+z}{2}\right)+f\left(\frac{x-y+z}{2}\right)-f(x)-f(z)\right\|_Y\le\kappa\left(\xi(|x|)+\xi(|y|)+\xi(|z|)\right)$$

for all $x, y, z \in G$. Then there exists a unique additive mapping $A : G \to X$ such that

$$\|f(x) - A(x)\| \le \frac{(2 + |2|)\xi(|x|)}{|2|}$$

for all $x \in G$.

Proof If we define $\zeta : G^3 \to [0, \infty)$ by $\zeta(x, y, z) := \kappa\,(\xi(|x|) + \xi(|y|) + \xi(|z|))$, then we have

$$\lim_{n\to\infty} |2|^n \,\zeta\left(\frac{x}{2^n}, \frac{y}{2^n}, \frac{z}{2^n}\right) = 0$$

for all $x, y, z \in G$. On the other hand,

$$\Psi(x) = \zeta\left(\frac{x}{2}, x, \frac{x}{2}\right) = \frac{(2 + |2|)\xi(|x|)}{|2|}$$

exists for all $x \in G$. Also, we have

$$\lim_{j\to\infty} \lim_{n\to\infty} \max_{j\le k<n+j} |2|^k \,\zeta\left(\frac{x}{2^{k+1}}, \frac{x}{2^k}, \frac{x}{2^{k+1}}\right) = \lim_{j\to\infty} |2|^j \,\zeta\left(\frac{x}{2^{j+1}}, \frac{x}{2^j}, \frac{x}{2^{j+1}}\right) = 0.$$

Applying Theorem 6, we have the conclusion.

Theorem 7 *Let G is an additive semigroup and that X is a non-Archimedean Banach space. Assume that $\zeta : G^3 \to [0, +\infty)$ is a function such that*

$$\lim_{n\to\infty} \frac{\zeta\,(2^n x, 2^n y, 2^n z)}{2^n} = 0 \tag{20}$$

for all $x, y, z \in G$. Suppose that, for any $x \in G$, the limit

$$\Psi(x) = \lim_{n\to\infty} \max_{0\le k<n} \frac{\zeta\left(2^k x, 2^{k+1} x, 2^k x\right)}{|2|^k} \tag{21}$$

exists and $f : G \to X$ is a mapping with $f(0) = 0$ and satisfying (12). Then the limit $A(x) := \lim_{n\to\infty} \frac{f(2^n x)}{2^n}$ exists for all $x \in G$ and

$$\|f(x) - A(x)\| \le \frac{\Psi(x)}{|2|}. \tag{22}$$

for all $x \in G$. Moreover, if

$$\lim_{j\to\infty}\lim_{n\to\infty}\max_{j\le k<n+j}\frac{\zeta\left(2^kx,2^{k+1}x,2^kx\right)}{|2|^k}=0,$$

then A is the unique mapping satisfying (22).

Proof It follows from (14), we get

$$\left\|f(x)-\frac{f(2x)}{2}\right\|_X\le\frac{\zeta(x,2x,x)}{|2|} \tag{23}$$

for all $x \in G$. Replacing x by $2^n x$ in (23), we obtain

$$\left\|\frac{f(2^nx)}{2^n}-\frac{f(2^{n+1}x)}{2^{n+1}}\right\|_X\le\frac{\zeta\left(2^nx,2^{n+1}x,2^nx\right)}{|2|^{n+1}}. \tag{24}$$

Thus it follows from (20) and (24) that the sequence $\left\{\frac{f(2^nx)}{2^n}\right\}_{n\ge1}$ is convergent. Set

$$A(x):=\lim_{n\to\infty}\frac{f(2^nx)}{2^n}.$$

On the other hand, it follows from (24) that

$$\begin{aligned}\left\|\frac{f(2^px)}{2^p}-\frac{f(2^qx)}{2^q}\right\|=\left\|\sum_{k=p}^{q-1}\frac{f(2^{k+1}x)}{2^{k+1}}-\frac{f(2^kx)}{2^k}\right\|&\le\max_{p\le k<q}\left\{\left\|\frac{f(2^{k+1}x)}{2^{k+1}}-\frac{f(2^kx)}{2^k}\right\|\right\}\\&\le\frac{1}{|2|}\max_{p\le k<q}\frac{\zeta\left(2^kx,2^{k+1}x,2^kx\right)}{|2|^k}\end{aligned}$$

for all $x \in G$ and $p, q \ge 0$ with $q > p \ge 0$. Letting $p = 0$, taking $q \to \infty$ in the last inequality and using (21), we obtain (22).

The rest of the proof is similar to the proof of Theorem 6.

Corollary 4 *Let $\xi : [0,\infty) \to [0,\infty)$ be a function satisfying*

$$\xi(|2|\,t)\le\xi(|2|)\,\xi(t),\quad \xi(|2|)<|2|$$

for all $t \ge 0$. Let $\kappa > 0$ and $f : G \to X$ be a mapping with $f(0) = 0$ satisfying

$$\left\|f\left(\frac{x+y+z}{2}\right)+f\left(\frac{x-y+z}{2}\right)-f(x)-f(z)\right\|\le\kappa\left(\xi(|x|)\cdot\xi(|y|)\cdot\xi(|z|)\right)$$

for all $x, y, z \in G$. Then there exists a unique additive mapping $A : G \to X$ such that

$$\|f(x) - A(x)\| \le \kappa\xi(|x|)^3$$

for all $x \in G$.

Proof If we define $\zeta : G^3 \to [0, \infty)$ by

$$\zeta(x, y, z) := \kappa\,(\xi(|x|) \cdot \xi(|y|) \cdot \xi(|z|))$$

and apply Theorem 7, then we get the conclusion.

3 Random Stability of the Functional Equation (1)

In this section, using the fixed point and direct methods, we prove the Hyers-Ulam stability of the functional equation (1) in random normed spaces.

3.1 Direct Method

Theorem 8 *Let X be a real linear space, $(Z, \mu', \min)$ be an RN-space and $\varphi : X^3 \to Z$ be a function such that there exists $0 < \alpha < \frac{1}{2}$ such that*

$$\mu'_{\varphi(\frac{x}{2}, \frac{y}{2}, \frac{z}{2})}(t) \ge \mu'_{\varphi(x,y,z)}\left(\frac{t}{\alpha}\right) \tag{25}$$

for all $x, y, z \in X$ and $t > 0$ and $\lim_{n\to\infty} \mu'_{\varphi(\frac{x}{2^n}, \frac{y}{2^n}, \frac{z}{2^n})}\left(\frac{t}{2^n}\right) = 1$ for all $x, y, z \in X$ and $t > 0$. Let $(Y, \mu, \min)$ be a complete RN-space. If $f : X \to Y$ is a mapping with $f(0) = 0$ such that

$$\mu_{f\left(\frac{x+y+z}{2}\right)+f\left(\frac{x-y+z}{2}\right)-f(x)-f(z)}(t) \ge \mu'_{\varphi(x,y,z)}(t) \tag{26}$$

for all $x, y, z \in X$ and $t > 0$. Then the limit $A(x) = \lim_{n\to\infty} 2^n f\left(\frac{x}{2^n}\right)$ exists for all $x \in X$ and defines a unique additive mapping $A : X \to Y$ such that

$$\mu_{f(x)-A(x)}(t) \ge \mu'_{\varphi(x,2x,x)}\left(\frac{(1-2\alpha)t}{\alpha}\right). \tag{27}$$

for all $x \in X$ and $t > 0$.

Proof Putting $y = 2x$ and $z = x$ in (26), we see that

$$\mu_{f(2x)-2f(x)}(t) \geq \mu'_{\varphi(x,2x,x)}(t). \tag{28}$$

Replacing x by $\frac{x}{2}$ in (28), we obtain

$$\mu_{2f(\frac{x}{2})-f(x)}(t) \geq \mu'_{\varphi(\frac{x}{2},x,\frac{x}{2})}(t) \geq \mu'_{\varphi(x,2x,x)}\left(\frac{t}{\alpha}\right) \tag{29}$$

for all $x \in X$. Replacing x by $\frac{x}{2^n}$ in (29) and using (25), we obtain

$$\mu_{2^{n+1}f\left(\frac{x}{2^{n+1}}\right)-2^n f\left(\frac{x}{2^n}\right)}(t) \geq \mu'_{\varphi\left(\frac{x}{2^{n+1}},\frac{x}{2^n},\frac{x}{2^{n+1}}\right)}\left(\frac{t}{2^n}\right) \geq \mu'_{\varphi(x,2x,x)}\left(\frac{t}{2^n\alpha^{n+1}}\right)$$

and so

$$\begin{aligned}
\mu_{2^n f\left(\frac{x}{2^n}\right)-f(x)}\left(\sum_{k=0}^{n-1} 2^k\alpha^{k+1}t\right) &= \mu_{\sum_{k=0}^{n-1} 2^{k+1}f\left(\frac{x}{2^{k+1}}\right)-2^k f\left(\frac{x}{2^k}\right)}\left(\sum_{k=0}^{n-1} 2^k\alpha^{k+1}t\right) \\
&\geq T_{k=0}^{n-1}\left(\mu_{2^{k+1}f\left(\frac{x}{2^{k+1}}\right)-2^k f\left(\frac{x}{2^k}\right)}(2^k\alpha^{k+1}t)\right) \\
&\geq T_{k=0}^{n-1}\left(\mu'_{\varphi(x,2x,x)}(t)\right) \\
&= \mu'_{\varphi(x,2x,x)}(t).
\end{aligned}$$

This implies that

$$\mu_{2^n f\left(\frac{x}{2^n}\right)-f(x)}(t) \geq \mu'_{\varphi(x,2x,x)}\left(\frac{t}{\sum_{k=0}^{n-1} 2^k\alpha^{k+1}}\right). \tag{30}$$

Replacing x by $\frac{x}{2^p}$ in (30), we obtain

$$\mu_{2^{n+p}f\left(\frac{x}{2^{n+p}}\right)-2^p f\left(\frac{x}{2^p}\right)}(t) \geq \mu'_{\varphi(x,2x,x)}\left(\frac{t}{\sum_{k=p}^{n+p-1} 2^k\alpha^{k+1}}\right). \tag{31}$$

Since $\lim_{p,n\to\infty}\mu'_{\varphi(x,2x,x)}\left(\frac{t}{\sum_{k=p}^{n+p-1} 2^k\alpha^{k+1}}\right) = 1$, it follows that $\left\{2^n f(\frac{x}{2^n})\right\}_{n=1}^{\infty}$ is a Cauchy sequence in a complete RN-space $(Y, \mu, \min)$ and so there exists a point $A(x) \in Y$ such that $\lim_{n\to\infty} 2^n f\left(\frac{x}{2^n}\right) = A(x)$. Fix $x \in X$ and put $p = 0$ in (31). Then, for any $\varepsilon > 0$,

$$\begin{aligned}\mu_{A(x)-f(x)}(t+\varepsilon) &\geq T\left(\mu_{A(x)-2^n f\left(\frac{x}{2^n}\right)}(\varepsilon), \mu_{2^n f\left(\frac{x}{2^n}\right)-f(x)}(t)\right)\\ &\geq T\left(\mu_{A(x)-2^n f\left(\frac{x}{2^n}\right)}(\varepsilon), \mu'_{\varphi(x,2x,x)}\left(\frac{t}{\sum_{k=0}^{n-1} 2^k \alpha^{k+1}}\right)\right). \end{aligned} \tag{32}$$

Taking $n \to \infty$ in (32), we get

$$\mu_{A(x)-f(x)}(t+\varepsilon) \geq \mu'_{\varphi(x,2x,x)}\left(\frac{(1-2\alpha)t}{\alpha}\right). \tag{33}$$

Since ε is arbitrary, by taking $\varepsilon \to 0$ in (33), we get

$$\mu_{A(x)-f(x)}(t) \geq \mu'_{\varphi(x,2x,x)}\left(\frac{(1-2\alpha)t}{\alpha}\right).$$

Replacing x, y and z by $\frac{x}{2^n}$, $\frac{y}{2^n}$ and $\frac{z}{2^n}$ in (26), respectively, we get

$$\mu_{2^n f\left(\frac{x+y+z}{2^{n+1}}\right)+2^n f\left(\frac{x-y+z}{2^{n+1}}\right)-2^n f\left(\frac{x}{2^n}\right)-2^n f\left(\frac{z}{2^n}\right)}(t) \geq \mu'_{\varphi\left(\frac{x}{2^n},\frac{y}{2^n},\frac{z}{2^n}\right)}\left(\frac{t}{2^n}\right)$$

for all $x, y, z \in X$ and $t > 0$. Since $\lim_{n\to\infty} \mu'_{\varphi\left(\frac{x}{2^n},\frac{y}{2^n},\frac{z}{2^n}\right)}\left(\frac{t}{2^n}\right) = 1$, we conclude that A satisfies (1). On the other hand,

$$2A\left(\frac{x}{2}\right) - A(x) = \lim_{n\to\infty} 2^{n+1} f\left(\frac{x}{2^{n+1}}\right) - \lim_{n\to\infty} 2^n f\left(\frac{x}{2^n}\right) = 0.$$

This implies that $A : X \to Y$ is an additive mapping.

To prove the uniqueness of the additive mapping A, assume that there exists another additive mapping $L : X \to Y$ which satisfies (27). Then we have

$$\begin{aligned}\mu_{A(x)-L(x)}(t) &= \lim_{n\to\infty} \mu_{2^n A\left(\frac{x}{2^n}\right)-2^n L\left(\frac{x}{2^n}\right)}(t)\\ &\geq \lim_{n\to\infty} \min\left\{\mu_{2^n A\left(\frac{x}{2^n}\right)-2^n f\left(\frac{x}{2^n}\right)}\left(\frac{t}{2}\right), \mu_{2^n f\left(\frac{x}{2^n}\right)-2^n L\left(\frac{x}{2^n}\right)}\left(\frac{t}{2}\right)\right\}\\ &\geq \lim_{n\to\infty} \mu'_{\varphi\left(\frac{x}{2^n},\frac{2x}{2^n},\frac{x}{2^n}\right)}\left(\frac{(1-2\alpha)t}{2^n}\right) \geq \lim_{n\to\infty} \mu'_{\varphi(x,2x,x)}\left(\frac{(1-2\alpha)t}{2^n\alpha^n}\right).\end{aligned}$$

Since $\lim_{n\to\infty} \mu'_{\varphi(x,2x,x)}\left(\frac{(1-2\alpha)t}{2^n\alpha^n}\right) = 1$, $\mu_{A(x)-L(x)}(t) = 1$ for all $t > 0$ and so $A(x) = L(x)$. This completes the proof.

Corollary 5 *Let X be a real normed linear space,* $(Z, \mu', \min)$ *be an RN-space and* $(Y, \mu, \min)$ *be a complete RN-space. Let r be a positive real number with* $r > 1$, $z_0 \in Z$ *and* $f : X \to Y$ *be a mapping with* $f(0) = 0$ *satisfying*

$$\mu_{f\left(\frac{x+y+z}{2}\right)+f\left(\frac{x-y+z}{2}\right)-f(x)-f(z)}(t) \geq \mu'_{(\|x\|^r+\|y\|^r+\|z\|^r)z_0}(t) \tag{34}$$

for all $x, y \in X$ and $t > 0$. Then the limit $A(x) = \lim_{n\to\infty} 2^n f\left(\frac{x}{2^n}\right)$ exists for all $x \in X$ and defines a unique additive mapping $A : X \to Y$ such that and

$$\mu_{f(x)-A(x)}(t) \geq \mu'_{\|x\|^p z_0}\left(\frac{(2^r-2)t}{2^r+2}\right)$$

for all $x \in X$ and $t > 0$.

Proof Let $\alpha = 2^{-r}$ and $\varphi : X^3 \to Z$ be a mapping defined by $\varphi(x, y, z) = (\|x\|^r + \|y\|^r + \|z\|^r)z_0$. Then, from Theorem 8, the conclusion follows.

Theorem 9 *Let X be a real linear space, $(Z, \mu', \min)$ be an RN-space and $\varphi : X^3 \to Z$ be a function such that there exists $0 < \alpha < 2$ such that $\mu'_{\varphi(2x,2y,2z)}(t) \geq \mu'_{\alpha\varphi(x,y,z)}(t)$ for all $x \in X$ and $t > 0$ and*

$$\lim_{n\to\infty} \mu'_{\varphi(2^n x,2^n y,2^n z)}(2^n x) = 1$$

for all $x, y, z \in X$ and $t > 0$. Let $(Y, \mu, \min)$ be a complete RN-space. If $f : X \to Y$ is a mapping with $f(0) = 0$ satisfying (26). Then the limit $A(x) = \lim_{n\to\infty} \frac{f(2^n x)}{2^n}$ exists for all $x \in X$ and defines a unique additive mapping $A : X \to Y$ such that

$$\mu_{f(x)-A(x)}(t) \geq \mu'_{\varphi(x,2x,x)}((2-\alpha)t). \tag{35}$$

for all $x \in X$ and $t > 0$.

Proof It follows from (28) that

$$\mu_{\frac{f(2x)}{2}-f(x)}(t) \geq \mu'_{\varphi(x,2x,x)}(2t). \tag{36}$$

Replacing x by $2^n x$ in (36), we obtain that

$$\mu_{\frac{f(2^{n+1}x)}{2^{n+1}}-\frac{f(2^n x)}{2^n}}(t) \geq \mu'_{\varphi(2^n x,2^{n+1}x,2^n x)}(2^{n+1}t) \geq \mu_{\varphi(x,2x,x)}\left(\frac{2^{n+1}t}{\alpha^n}\right).$$

The rest of the proof is similar to the proof of Theorem 8.

Corollary 6 *Let X be a real normed linear space, $(Z, \mu', \min)$ be an RN-space and $(Y, \mu, \min)$ be a complete RN-space. Let r be a positive real number with $0 < r < 1$, $z_0 \in Z$ and $f : X \to Y$ be a mapping with $f(0) = 0$ satisfying (34). Then the limit $A(x) = \lim_{n\to\infty} \frac{f(2^n x)}{2^n}$ exists for all $x \in X$ and defines a unique additive mapping $A : X \to Y$ such that and*

$$\mu_{f(x)-A(x)}(t) \geq \mu'_{\|x\|^p z_0}\left(\frac{(2-2^r)t}{2^r+2}\right)$$

for all $x \in X$ and $t > 0$.

Proof Let $\alpha = 2^r$ and $\varphi : X^3 \to Z$ be a mapping defined by $\varphi(x, y, z) = (\|x\|^r + \|y\|^r + \|z\|^r)z_0$. Then, from Theorem 9, the conclusion follows.

3.2 Fixed Point Method

Theorem 10 *Let X be a linear space, (Y, μ, T_M) be a complete RN-space and Φ be a mapping from X^3 to D^+($\Phi(x, y, z)$ is denoted by $\Phi_{x,y.z}$) such that there exists $0 < \alpha < \frac{1}{2}$ such that*

$$\Phi_{2x,2y,2z}(t) \leq \Phi_{x,y,z}(\alpha t) \tag{37}$$

for all $x, y, z \in X$ and $t > 0$. Let $f : X \to Y$ be a mapping with $f(0) = 0$ satisfying

$$\mu_{f\left(\frac{x+y+z}{2}\right)+f\left(\frac{x-y+z}{2}\right)-f(x)-f(z)}(t) \geq \Phi_{x,y,z}(t) \tag{38}$$

for all $x, y, z \in X$ and $t > 0$. Then, for all $x \in X$, $A(x) := \lim_{n\to\infty} 2^n f\left(\frac{x}{2^n}\right)$ exists and $A : X \to Y$ is a unique additive mapping such that

$$\mu_{f(x)-A(x)}(t) \geq \Phi_{x,2x,x}\left(\frac{(1-2\alpha)t}{\alpha}\right) \tag{39}$$

for all $x \in X$ and $t > 0$.

Proof Putting $y = 2x$ and $z = x$ in (38), we have

$$\mu_{2f(\frac{x}{2})-f(x)}(t) \geq \Phi_{\frac{x}{2},x,\frac{x}{2}}(t) \geq \Phi_{x,2x,x}\left(\frac{t}{\alpha}\right) \tag{40}$$

for all $x \in X$ and $t > 0$. Consider the set $S := \{g : X \to Y\}$ and the generalized metric d in S defined by

$$d(f, g) = \inf_{u\in(0,\infty)} \left\{\mu_{g(x)-h(x)}(ut) \geq \Phi_{x,2x,x}(t),\ \forall x \in X,\ t > 0\right\}, \tag{41}$$

where $\inf \emptyset = +\infty$. It is easy to show that (S, d) is complete (see [33], Lemma 2.1). Now, we consider a linear mapping $J : (S, d) \to (S, d)$ such that

$$Jh(x) := 2h\left(\frac{x}{2}\right) \tag{42}$$

for all $x \in X$. First, we prove that J is a strictly contractive mapping with the Lipschitz constant 2α. In fact, let $g, h \in S$ be such that $d(g,h) < \varepsilon$. Then we have $\mu_{g(x)-h(x)}(\varepsilon t) \geq \Phi_{x,2x,x}(t)$ for all $x \in X$ and $t > 0$ and so

$$\begin{aligned}\mu_{Jg(x)-Jh(x)}(2\alpha\varepsilon t) = \mu_{2g(\frac{x}{2})-2h(\frac{x}{2})}(2\alpha\varepsilon t) &= \mu_{g(\frac{x}{2})-h(\frac{x}{2})}(\alpha\varepsilon t)\\ &\geq \Phi_{\frac{x}{2},x,\frac{x}{2}}(\alpha t)\\ &\geq \Phi_{x,2x,x}(t)\end{aligned}$$

for all $x \in X$ and $t > 0$. Thus $d(g,h) < \varepsilon$ implies that $d(Jg, Jh) < 2\alpha\varepsilon$. This means that $d(Jg, Jh) \leq 2\alpha d(g,h)$ for all $g, h \in S$. It follows from (40) that

$$d(f, Jf) \leq \alpha.$$

By Theorem 3, there exists a mapping $A : X \to Y$ satisfying the following:

(1) A is a fixed point of J, that is,

$$A\left(\frac{x}{2}\right) = \frac{1}{2}A(x) \tag{43}$$

for all $x \in X$. The mapping A is a unique fixed point of J in the set $\Omega = \{h \in S : d(g,h) < \infty\}$. This implies that A is a unique mapping satisfying (43) such that there exists $u \in (0,\infty)$ satisfying $\mu_{f(x)-A(x)}(ut) \geq \Phi_{x,2x,x}(t)$ for all $x \in X$ and $t > 0$.

(2) $d(J^n f, A) \to 0$ as $n \to \infty$. This implies the equality

$$\lim_{n\to\infty} 2^n f\left(\frac{x}{2^n}\right) = A(x)$$

for all $x \in X$.

(3) $d(f, A) \leq \frac{d(f,Jf)}{1-2\alpha}$ with $f \in \Omega$, which implies the inequality $d(f,A) \leq \frac{\alpha}{1-2\alpha}$ and so

$$\mu_{f(x)-A(x)}\left(\frac{\alpha t}{1-2\alpha}\right) \geq \Phi_{x,2x,x}(t)$$

for all $x \in X$ and $t > 0$. This implies that the inequality (39) holds. On the other hand

$$\mu_{2^n f\left(\frac{x+y+z}{2^{n+1}}\right)+2^n f\left(\frac{x-y+z}{2^{n+1}}\right)-2^n f\left(\frac{x}{2^n}\right)-2^n f\left(\frac{z}{2^n}\right)}(t) \geq \Phi_{\frac{x}{2^n},\frac{y}{2^n},\frac{z}{2^n}}\left(\frac{t}{2^n}\right)$$

for all $x, y, z \in X$, $t > 0$ and $n \geq 1$. By (37), we know that

$$\Phi_{\frac{x}{2^n}, \frac{y}{2^n}, \frac{z}{2^n}}\left(\frac{t}{2^n}\right) \geq \Phi_{x,y,z}\left(\frac{t}{(2\alpha)^n}\right).$$

Since $\lim_{n\to\infty} \Phi_{x,y,z}\left(\frac{t}{(2\alpha)^n}\right) = 1$ for all $x, y, z \in X$ and $t > 0$, we have

$$\mu_{A\left(\frac{x+y+z}{2}\right)+A\left(\frac{x-y+z}{2}\right)-A(x)-A(z)}(t) = 1$$

for all $x, y, z \in X$ and $t > 0$. Thus the mapping $A : X \to Y$ satisfying (1). Furthermore

$$\begin{aligned} A(2x) - 2A(x) &= \lim_{n\to\infty} 2^n f\left(\frac{x}{2^{n-1}}\right) - 2\lim_{n\to\infty} 2^n f\left(\frac{x}{2^n}\right) \\ &= 2\left[\lim_{n\to\infty} 2^{n-1} f\left(\frac{x}{2^{n-1}}\right) - \lim_{n\to\infty} 2^n f\left(\frac{x}{2^n}\right)\right] \\ &= 0. \end{aligned}$$

This completes the proof.

Corollary 7 *Let X be a real normed space, $\theta \geq 0$ and r be a real number with $r > 1$. Let $f : X \to Y$ be a mapping with $f(0) = 0$ satisfying*

$$\mu_{f\left(\frac{x+y+z}{2}\right)+f\left(\frac{x-y+z}{2}\right)-f(x)-f(z)}(t) \geq \frac{t}{t + \theta\left(\|x\|^r + \|y\|^r + \|z\|^r\right)} \tag{44}$$

for all $x, y, z \in X$ and $t > 0$. Then $A(x) = \lim_{n\to\infty} 2^n f\left(\frac{x}{2^n}\right)$ exists for all $x \in X$ and $A : X \to Y$ is a unique additive mapping such that

$$\mu_{f(x)-A(x)}(t) \geq \frac{(2^r - 2)t}{(2^r - 2)t + (2^r + 2)\theta\|x\|^r}$$

for all $x \in X$ and $t > 0$.

Proof The proof follows from Theorem 10 if we take

$$\Phi_{x,y,z}(t) = \frac{t}{t + \theta\left(\|x\|^r + \|y\|^r + \|z\|^r\right)}$$

for all $x, y, z \in X$ and $t > 0$. In fact, if we choose $\alpha = 2^{-r}$, then we get the desired result.

Theorem 11 *Let X be a linear space, (Y, μ, T_M) be a complete RN-space and Φ be a mapping from X^3 to D^+ ($\Phi(x, y, z)$ is denoted by $\Phi_{x,y,z}$) such that for some $0 < \alpha < 2$*

$$\Phi_{\frac{x}{2}, \frac{y}{2}, \frac{z}{2}}(t) \leq \Phi_{x,y,z}(\alpha t)$$

for all $x, y, z \in X$ and $t > 0$. Let $f : X \to Y$ be a mapping with $f(0) = 0$ satisfying (38). Then the limit $A(x) := \lim_{n\to\infty} \frac{f(2^n x)}{2^n}$ exists for all $x \in X$ and $A : X \to Y$ is a unique additive mapping such that

$$\mu_{f(x)-A(x)}(t) \geq \Phi_{x,2x,x}((2-\alpha)t) \tag{45}$$

for all $x \in X$ and $t > 0$.

Proof Putting $y = 2x$ and $z = x$ in (38), we have

$$\mu_{\frac{f(2x)}{2}-f(x)}(t) \geq \Phi_{x,2x,x}(2t) \tag{46}$$

for all $x \in X$ and $t > 0$. Let (S, d) be the generalized metric space defined in the proof of Theorem 8. Now, we consider a linear mapping $J : (S, d) \to (S, d)$ such that

$$Jh(x) := \frac{1}{2}h(2x) \tag{47}$$

for all $x \in X$. It follows from (46) that $d(f, Jf) \leq \frac{1}{2}$. By Theorem 3, there exists a mapping $A : X \to Y$ satisfying the following:

(1) A is a fixed point of J, that is,

$$A(2x) = 2A(x) \tag{48}$$

for all $x \in X$. The mapping A is a unique fixed point of J in the set $\Omega = \{h \in S : d(g, h) < \infty\}$. This implies that A is a unique mapping satisfying (48) such that there exists $u \in (0, \infty)$ satisfying $\mu_{f(x)-A(x)}(ut) \geq \Phi_{x,2x,x}(t)$ for all $x \in X$ and $t > 0$.

(2) $d(J^n f, A) \to 0$ as $n \to \infty$. This implies the equality

$$\lim_{n\to\infty} \frac{f(2^n x)}{2^n} = A(x)$$

for all $x \in X$.

(3) $d(f, A) \leq \frac{d(f,Jf)}{1-\frac{\alpha}{2}}$ with $f \in \Omega$, which implies the inequality

$$\mu_{f(x)-A(x)}\left(\frac{t}{2-\alpha}\right) \geq \Phi_{x,2x,x}(t)$$

for all $x \in X$ and $t > 0$. This implies that the inequality (45) holds.

The rest of the proof is similar to the proof of Theorem 10.

Corollary 8 *Let X be a real normed space, $\theta \geq 0$ and r be a real number with $0 < r < 1$. Let $f : X \to Y$ be a mapping with $f(0) = 0$ satisfying (44). Then* *the limit* $A(x) = \lim_{n\to\infty} \frac{f(2^n x)}{2^n}$ *exists for all* $x \in X$ *and* $A : X \to Y$ *is a unique additive mapping such that*

$$\mu_{f(x)-A(x)}(t) \geq \frac{(2-2^r)t}{(2-2^r)t + (2^r+2)\theta\|x\|^r}$$

for all $x \in X$ *and* $t > 0$.

Proof The proof follows from Theorem 11 if we take

$$\Phi_{x,y}(t) = \frac{t}{t + \theta(\|x\|^r + \|y\|^r + \|z\|^r)}$$

for all $x, y, z \in X$ and $t > 0$. In fact, if we choose $\alpha = 2^r$, then we get the desired result.

4 Fuzzy Stability of the Functional Equation (1)

Throughout this section, using the fixed point and direct methods, we prove the Hyers-Ulam stability of functional equation (1) in fuzzy normed spaces.

4.1 Direct Method

In this section, using the direct method, we prove the Hyers-Ulam stability of the functional equation (1) in fuzzy Banach spaces. Throughout this section, we assume that X is a linear space, (Y, N) is a fuzzy Banach space and (Z, N') is a fuzzy normed space. Moreover, we assume that $N(x, .)$ is a left continuous function on R.

Theorem 12 *Assume that a mapping* $f : X \to Y$ *with* $f(0) = 0$ *satisfies the inequality*

$$N\left(f\left(\frac{x+y+z}{2}\right) + f\left(\frac{x-y+z}{2}\right) - f(x) - f(z), t\right) \geq N'(\varphi(x,y,z), t) \tag{49}$$

for all $x, y, z \in X$, $t > 0$ *and* $\varphi : X^3 \to Z$ *is a mapping for which there is a constant* $r \in R$ *satisfying* $0 < |r| < \frac{1}{2}$ *such that*

$$N'\left(\varphi\left(\frac{x}{2}, \frac{y}{2}, \frac{z}{2}\right), t\right) \geq N'\left(\varphi(x,y,z), \frac{t}{|r|}\right) \tag{50}$$

for all $x, y, z \in X$ and all $t > 0$. Then we can find a unique additive mapping $A : X \to Y$ satisfying (1) and the inequality

$$N(f(x) - A(x), t) \geq N'\left(\frac{|r|\varphi(x, 2x, x)}{1 - 2|r|}, t\right) \tag{51}$$

for all $x \in X$ and all $t > 0$.

Proof It follows from (50) that

$$N'\left(\varphi\left(\frac{x}{2^j}, \frac{y}{2^j}, \frac{z}{2^j}\right), t\right) \geq N'\left(\varphi(x, y, z), \frac{t}{|r|^j}\right). \tag{52}$$

So $N'\left(\varphi\left(\frac{x}{2^j}, \frac{y}{2^j}, \frac{z}{2^j}\right), |r|^j t\right) \geq N'\left(\varphi(x, y, z), t\right)$ for all $x, y, z \in X$ and all $t > 0$.

Substituting $y = 2x$ and $z = x$ in (49), we obtain

$$N\left(f(2x) - 2f(x), t\right) \geq N'(\varphi(x, 2x, x), t) \tag{53}$$

So

$$N\left(f(x) - 2f\left(\frac{x}{2}\right), t\right) \geq N'\left(\varphi\left(\frac{x}{2}, x, \frac{x}{2}\right), t\right) \tag{54}$$

for all $x \in X$ and all $t > 0$. Replacing x by $\frac{x}{2^j}$ in (54), we have

$$\begin{aligned} N\left(2^{j+1} f\left(\frac{x}{2^{j+1}}\right) - 2^j f\left(\frac{x}{2^j}\right), 2^j t\right) &\geq N'\left(\varphi\left(\frac{x}{2^{j+1}}, \frac{x}{2^j}, \frac{x}{2^{j+1}}\right), t\right) \\ &\geq N'\left(\varphi\left(x, 2x, x\right), \frac{t}{|r|^{j+1}}\right) \end{aligned} \tag{55}$$

for all $x \in X$, all $t > 0$ and any integer $j \geq 0$. So

$$\begin{aligned} &N\left(f(x) - 2^n f\left(\frac{x}{2^n}\right), \sum_{j=0}^{n-1} 2^j |r|^{j+1} t\right) \\ &\quad = N\left(\sum_{j=0}^{n-1}\left[2^{j+1} f\left(\frac{x}{2^{j+1}}\right) - 2^j f\left(\frac{x}{2^j}\right)\right], \sum_{j=0}^{n-1} 2^j |r|^{j+1} t\right) \\ &\quad \geq \min_{0 \leq j \leq n-1}\left\{N\left(2^{j+1} f\left(\frac{x}{2^{j+1}}\right) - 2^j f\left(\frac{x}{2^j}\right), 2^j |r|^{j+1} t\right)\right\} \\ &\quad \geq N'(\varphi(x, 2x, x), t) \end{aligned}$$

which means

$$N\left(2^{n+p}f\left(\frac{x}{2^{n+p}}\right)-2^pf\left(\frac{x}{2^p}\right),\sum_{j=0}^{n-1}2^j|r|^{j+1}t\right)\geq N'\left(\varphi\left(\frac{x}{2^p},\frac{2x}{2^p},\frac{x}{2^p}\right),t\right)$$
$$\geq N'\left(\varphi(x,2x,x),\frac{t}{|r|^p}\right)$$

for all $x\in X$, $t>0$ and all integers $n>0$, $p\geq 0$. So

$$N\left(2^{n+p}f\left(\frac{x}{2^{n+p}}\right)-2^pf\left(\frac{x}{2^p}\right),\sum_{j=0}^{n-1}2^{j+p}|r|^{j+p+1}t\right)\geq N'(\varphi(x,2x,x),t)$$

for all $x\in X$, $t>0$ and all integers $n>0$, $p\geq 0$. Hence one obtains

$$N\left(2^{n+p}f\left(\frac{x}{2^{n+p}}\right)-2^pf\left(\frac{x}{2^p}\right),t\right)$$
$$\geq N'\left(\varphi(x,2x,x),\frac{t}{\sum_{j=0}^{n-1}2^{j+p}|r|^{j+p+1}}\right) \tag{56}$$

for all $x\in X$, $t>0$ and all integers $n>0$, $p\geq 0$. Since the series

$$\sum_{j=0}^{\infty}2^j|r|^j$$

is convergent series, we obtain by taking the limit $p\to\infty$ in the last inequality that a sequence $\left\{2^nf\left(\frac{x}{2^n}\right)\right\}$ is a Cauchy sequence in the fuzzy Banach space (Y,N) and so it converges in Y. Therefore a mapping $A:X\to Y$ defined by $A(x):=N-\lim_{n\to\infty}2^nf\left(\frac{x}{2^n}\right)$ is well defined for all $x\in X$. It means that

$$\lim_{n\to\infty}N\left(A(x)-2^nf\left(\frac{x}{2^n}\right),t\right)=1 \tag{57}$$

for all $x\in X$ and all $t>0$. In addition, it follows from (56) that

$$N\left(2^nf\left(\frac{x}{2^n}\right)-f(x),t\right)\geq N'\left(\varphi(x,2x,x),\frac{t}{\sum_{j=0}^{n-1}2^j|r|^{j+1}}\right)$$

for all $x\in X$ and all $t>0$. So

$$N(f(x)-A(x),t)$$
$$\geq\min\left\{N\left(f(x)-2^nf\left(\frac{x}{2^n}\right),(1-\varepsilon)t\right),N\left(A(x)-2^nf\left(\frac{x}{2^n}\right),\varepsilon t\right)\right\}$$

$$\geq N'\left(\varphi(x,2x,x),\ \frac{t}{\sum_{j=0}^{n-1} 2^j|r|^{j+1}}\right) \geq N'\left(\varphi(x,2x,x),\ \frac{(1-2|r|)\varepsilon t}{|r|}\right)$$

for sufficiently large n and for all $x \in X$, $t > 0$ and ε with $0 < \varepsilon < 1$. Since ε is arbitrary and N' is left continuous, we obtain

$$N(f(x) - A(x), t) \geq N'\left(\varphi(x,2x,x),\ \frac{(1-2|r|)t}{|r|}\right)$$

for all $x \in X$ and $t > 0$. It follows from (49) that

$$N\left(2^n f\left(\frac{x+y+z}{2^{n+1}}\right) + 2^n f\left(\frac{x-y+z}{2^{n+1}}\right) - 2^n f\left(\frac{x}{2^n}\right) - 2^n f\left(\frac{z}{2^n}\right), t\right)$$
$$\geq N'\left(\varphi\left(\frac{x}{2^n}, \frac{y}{2^n}, \frac{z}{2^n}\right), \frac{t}{2^n}\right) \geq N'\left(\varphi(x,y,z),\ \frac{t}{2^n|r|^n}\right)$$

for all $x, y, z \in X$, $t > 0$ and all $n \in N$. Since

$$\lim_{n\to\infty} N'\left(\varphi(x,y,z),\ \frac{t}{2^n|r|^n}\right) = 1,$$

$$N\left(2^n f\left(\frac{x+y+z}{2^{n+1}}\right) + 2^n f\left(\frac{x-y+z}{2^{n+1}}\right) - 2^n f\left(\frac{x}{2^n}\right) - 2^n f\left(\frac{z}{2^n}\right), t\right) \to 1$$

for all $x, y, z \in X$ and all $t > 0$. Therefore, we obtain, in view of (57),

$$N\left(A\left(\frac{x+y+z}{2}\right) + A\left(\frac{x-y+z}{2}\right) - A(x) - A(z), t\right)$$
$$\geq \min\left\{N\left(A\left(\frac{x+y+z}{2}\right) + A\left(\frac{x-y+z}{2}\right) - A(x) - A(z)\right.\right.$$
$$\left. -2^n f\left(\frac{x+y+z}{2^{n+1}}\right) + 2^n f\left(\frac{x-y+z}{2^{n+1}}\right) - 2^n f\left(\frac{x}{2^n}\right) - 2^n f\left(\frac{z}{2^n}\right), \frac{t}{2}\right),$$
$$\left. N\left(2^n f\left(\frac{x+y+z}{2^{n+1}}\right) + 2^n f\left(\frac{x-y+z}{2^{n+1}}\right) - 2^n f\left(\frac{x}{2^n}\right) - 2^n f\left(\frac{z}{2^n}\right), \frac{t}{2}\right)\right\}$$
$$= N\left(2^n f\left(\frac{x+y+z}{2^{n+1}}\right) + 2^n f\left(\frac{x-y+z}{2^{n+1}}\right) - 2^n f\left(\frac{x}{2^n}\right) - 2^n f\left(\frac{z}{2^n}\right), \frac{t}{2}\right)$$
$$\geq N'\left(\varphi(x,y,z),\ \frac{t}{2^{n+1}|r|^n}\right) \to 1 \quad \text{as } n \to \infty$$

which implies

$$A\left(\frac{x+y+z}{2}\right)+A\left(\frac{x-y+z}{2}\right)=A(x)+A(z)$$

for all $x, y, z \in X$. Thus $A : X \to Y$ is a mapping satisfying (1) and (51).

To prove the uniqueness, assume that there is another mapping $L : X \to Y$ which satisfies (51). Since $L(2^n x) = 2^n L(x)$ for all $x \in X$, we have

$$\begin{aligned}
&N(A(x)-L(x),t)\\
&=N\left(2^nA\left(\frac{x}{2^n}\right)-2^nL\left(\frac{x}{2^n}\right),t\right)\\
&\geq \min\left\{N\left(2^nA\left(\frac{x}{2^n}\right)-2^nf\left(\frac{x}{2^n}\right),\frac{t}{2}\right),N\left(2^nf\left(\frac{x}{2^n}\right)-2^nL\left(\frac{x}{2^n}\right),\frac{t}{2}\right)\right\}\\
&\geq N'\left(\varphi\left(\frac{x}{2^n},\frac{2x}{2^n},\frac{x}{2^n}\right),\frac{(1-2|r|)t}{|r|2^{n+1}}\right)\geq N\left(\varphi(x,2x,x),\frac{(1-2|r|)t}{|r|^{n+1}2^{n+1}}\right)\to 1\\
&\quad\text{as } n\to\infty
\end{aligned}$$

for all $t > 0$. Therefore, $A(x) = L(x)$ for all $x \in X$, which completes the proof.

Corollary 9 *Let X be a normed spaces and that (R, N') a fuzzy Banach space. Assume that there exists real number $\theta \geq 0$ and $0 < p < 2$ such that a mapping $f : X \to Y$ with $f(0) = 0$ satisfying the following inequality*

$$N\left(f\left(\frac{x+y+z}{2}\right)+f\left(\frac{x-y+z}{2}\right)-f(x)-f(z),t\right)$$
$$\geq N'\left(\theta\left(\|x\|^p+\|y\|^p+\|z\|^p\right),t\right)$$

for all $x, y, z \in X$ and $t > 0$. Then there is a unique additive mapping $A : X \to Y$ that satisfying (1) and the inequality

$$N(f(x)-A(x),t)\geq N'\left(\frac{(2^r+2)\theta\|x\|^p}{2},t\right)$$

Proof Let $\varphi(x, y, z) := \theta(\|x\|^p+\|y\|^p+\|z\|^p)$ and $|r| = \frac{1}{4}$. Applying Theorem 12, we get the desired result.

Theorem 13 *Assume that a mapping $f : X \to Y$ with $f(0) = 0$ satisfies the inequality (49) and $\varphi : X^2 \to Z$ is a mapping for which there is a constant $r \in R$ satisfying $0 < |r| < 2$ such that*

$$N'\left(\varphi(x,y,z),|r|t\right)\geq N'\left(\varphi\left(\frac{x}{2},\frac{y}{2},\frac{z}{2}\right),t\right) \tag{58}$$

for all $x, y, z \in X$ and all $t > 0$. Then we can find a unique additive mapping $A : X \to Y$ that satisfying (1) and the following inequality

$$N(f(x) - A(x), t) \geq N'\left(\frac{\varphi(x, 2x, x)}{2 - |r|}, t\right) \tag{59}$$

for all $x \in X$ and all $t > 0$.

Proof It follows from (53) that

$$N\left(\frac{f(2x)}{2} - f(x), \frac{t}{2}\right) \geq N'(\varphi(x, 2x, x), t) \tag{60}$$

for all $x \in X$ and all $t > 0$. Replacing x by $2^n x$ in (60), we obtain

$$\begin{aligned} N\left(\frac{f(2^{n+1}x)}{2^{n+1}} - \frac{f(2^n x)}{2^n}, \frac{t}{2^{n+1}}\right) &\geq N'(\varphi(2^n x, 2^{n+1}x, 2^n x), t) \\ &\geq N'\left(\varphi(x, 2x, x), \frac{t}{|r|^n}\right). \end{aligned} \tag{61}$$

So

$$N\left(\frac{f(2^{n+1}x)}{2^{n+1}} - \frac{f(2^n x)}{2^n}, \frac{|r|^n t}{2^{n+1}}\right) \geq N'(\varphi(x, 2x, x), t) \tag{62}$$

for all $x \in X$ and all $t > 0$. Proceeding as in the proof of Theorem 12, we obtain that

$$N\left(f(x) - \frac{f(2^n x)}{2^n}, \sum_{j=0}^{n-1} \frac{|r|^j t}{2^{j+1}}\right) \geq N'(\varphi(x, 2x, x), t)$$

for all $x \in X$, all $t > 0$ and any integer $n > 0$. So

$$\begin{aligned} N\left(f(x) - \frac{f(2^n x)}{2^n}, t\right) &\geq N'\left(\varphi(x, 2x, x), \frac{t}{\sum_{j=0}^{n-1} \frac{|r|^j}{2^{j+1}}}\right) \\ &\geq N'\left(\varphi(x, 2x, x), (2 - |r|)t\right). \end{aligned} \tag{63}$$

The rest of the proof is similar to the proof of Theorem 12.

Corollary 10 *Let X be a normed spaces and that (R, N') a fuzzy Banach space. Assume that there exists real number $\theta \geq 0$ and $0 < p = p_1 + p_2 + p_3 < 2$ such that a mapping $f : X \to Y$ with $f(0) = 0$ satisfying the following inequality*

$$N\left(f\left(\frac{x+y+z}{2}\right)+f\left(\frac{x-y+z}{2}\right)-f(x)-f(z),t\right)$$
$$\geq N'\left(\theta\left(\|x\|^{p_1}\cdot\|y\|^{p_2}\cdot\|z\|^{p_3}\right),t\right)$$

for all $x, y, z \in X$ and $t > 0$. Then there is a unique additive mapping $A : X \to Y$ that satisfying (1) and the inequality

$$N(f(x)-A(x),t) \geq N'\left((2^r+2)\theta\|x\|^p,t\right)$$

for all $x \in X$ and $t > 0$.

Proof Let $\varphi(x, y, z) := \theta\left(\|x\|^{p_1}\cdot\|y\|^{p_2}\cdot\|z\|^{p_3}\right)$ and $|r| = 1$. Applying Theorem 13, we get the desired result.

4.2 Fixed Point Method

Throughout this subsection, using the fixed point alternative approach we prove the Hyers-Ulam-Rassias stability of the functional equation (1) in fuzzy Banach spaces.

In this subsection, assume that X is a vector space and that (Y, N) is a fuzzy Banach space.

Theorem 14 *Let $\varphi : X^3 \to [0,\infty)$ be a function such that there exists an $L < 1$ with*

$$\varphi\left(\frac{x}{2},\frac{y}{2},\frac{z}{2}\right) \leq \frac{L\varphi(x,y,z)}{2}$$

for all $x, y, z \in X$. Let $f : X \to Y$ with $f(0) = 0$ be a mapping satisfying

$$N\left(f\left(\frac{x+y+z}{2}\right)+f\left(\frac{x-y+z}{2}\right)-f(x)-f(z),t\right) \geq \frac{t}{t+\varphi(x,y,z)} \tag{64}$$

for all $x, y, z \in X$ and all $t > 0$. Then the limit $A(x) := N - \lim_{n\to\infty} 2^n f\left(\frac{x}{2^n}\right)$ exists for each $x \in X$ and defines a unique additive mapping $A : X \to Y$ such that

$$N(f(x)-A(x),t) \geq \frac{(2-2L)t}{(2-2L)t+L\varphi(x,2x,x)} \tag{65}$$

for all $x \in X$ and $t > 0$.

Proof Putting $y = 2x$ and $z = x$ in (64) and replacing x by $\frac{x}{2}$, we have

$$N\left(2f\left(\frac{x}{2}\right) - f(x), t\right) \geq \frac{t}{t + \varphi\left(\frac{x}{2}, x, \frac{x}{2}\right)} \tag{66}$$

for all $x \in X$ and $t > 0$. Consider the set $S := \{g : X \to Y\}$ and the generalized metric d in S defined by

$$d(f, g) = \inf\left\{\mu \in R^+ : N(g(x) - h(x), \mu t) \geq \frac{t}{t + \varphi(x, 2x, x)}, \forall x \in X,\ t > 0\right\},$$

where $\inf \emptyset = +\infty$. It is easy to show that (S, d) is complete (see [33, Lemma 2.1]). Now, we consider a linear mapping $J : S \to S$ such that

$$Jg(x) := 2g\left(\frac{x}{2}\right)$$

for all $x \in X$. Let $g, h \in S$ be such that $d(g, h) = \varepsilon$. Then

$$N(g(x) - h(x), \varepsilon t) \geq \frac{t}{t + \varphi(x, 2x, x)}$$

for all $x \in X$ and $t > 0$. Hence

$$\begin{aligned} N(Jg(x) - Jh(x), L\varepsilon t) &= N\left(2g\left(\frac{x}{2}\right) - 2h\left(\frac{x}{2}\right), L\varepsilon t\right) \\ &= N\left(g\left(\frac{x}{2}\right) - h\left(\frac{x}{2}\right), \frac{L\varepsilon t}{2}\right) \geq \frac{\frac{Lt}{2}}{\frac{Lt}{2} + \varphi\left(\frac{x}{2}, x, \frac{x}{2}\right)} \\ &\geq \frac{\frac{Lt}{2}}{\frac{Lt}{2} + \frac{L\varphi(x,2x,x)}{2}} = \frac{t}{t + \varphi(x, 2x, x)} \end{aligned}$$

for all $x \in X$ and $t > 0$. Thus $d(g, h) = \varepsilon$ implies that $d(Jg, Jh) \leq L\varepsilon$. This means that $d(Jg, Jh) \leq Ld(g, h)$ for all $g, h \in S$. It follows from (66) that

$$N\left(f(x) - 2f\left(\frac{x}{2}\right), t\right) \geq \frac{t}{t + \varphi\left(\frac{x}{2}, x, \frac{x}{2}\right)} \geq \frac{t}{t + \frac{L\varphi(x,2x,x)}{2}} = \frac{\frac{2t}{L}}{\frac{2t}{L} + \varphi(x, 2x, x)}.$$

Therefore

$$N\left(f(x) - 2f\left(\frac{x}{2}\right), \frac{Lt}{2}\right) \geq \frac{t}{t + \varphi(x, 2x, x)}. \tag{67}$$

This means that $d(f, Jf) \leq \frac{L}{2}$. By Theorem 3, there exists a mapping $A : X \to Y$ satisfying the following:

(1) A is a fixed point of J, that is,

$$A\left(\frac{x}{2}\right) = \frac{A(x)}{2} \tag{68}$$

for all $x \in X$. The mapping A is a unique fixed point of J in the set $\Omega = \{h \in S : d(g, h) < \infty\}$. This implies that A is a unique mapping satisfying (68) such that there exists $\mu \in (0, \infty)$ satisfying

$$N(f(x) - A(x), \mu t) \geq \frac{t}{t + \varphi(x, 2x, x)}$$

for all $x \in X$ and $t > 0$.

(2) $d(J^n f, A) \to 0$ as $n \to \infty$. This implies the equality $N - \lim_{n\to\infty} 2^n f\left(\frac{x}{2^n}\right) = A(x)$ for all $x \in X$.

(3) $d(f, A) \leq \frac{d(f,Jf)}{1-L}$ with $f \in \Omega$, which implies the inequality $d(f, A) \leq \frac{L}{2-2L}$. This implies that the inequality (65) holds. Furthermore,

$$\begin{aligned}
&N\left(A\left(\frac{x+y+z}{2}\right) + A\left(\frac{x-y+z}{2}\right) - A(x) - A(z), t\right) \\
&\geq N - \lim_{n\to\infty}\left(2^n f\left(\frac{x+y+z}{2^{n+1}}\right) + 2^n f\left(\frac{x-y+z}{2^{n+1}}\right) - 2^n f\left(\frac{x}{2^n}\right) - 2^n f\left(\frac{z}{2^n}\right), t\right) \\
&\geq \lim_{n\to\infty} \frac{\frac{t}{2^n}}{\frac{t}{2^n} + \varphi\left(\frac{x}{2^n}, \frac{y}{2^n}, \frac{z}{2^n}\right)} \geq \lim_{n\to\infty} \frac{\frac{t}{2^n}}{\frac{t}{2^n} + \frac{L^n \varphi(x,y,z)}{2^n}} \to 1
\end{aligned}$$

for all $x, y, z \in X$, $t > 0$. So

$$N\left(A\left(\frac{x+y+z}{2}\right) + A\left(\frac{x-y+z}{2}\right) - A(x) - A(z), t\right) = 1$$

for all $x, y, z \in X$ and all $t > 0$. Thus the mapping $A : X \to Y$ is additive, as desired.

Corollary 11 *Let $\theta \geq 0$ and let p be a real number with $p > 1$. Let X be a normed vector space with norm $\|.\|$. Let $f : X \to Y$ with $f(0) = 0$ be a mapping satisfying*

$$N\left(f\left(\frac{x+y+z}{2}\right) + f\left(\frac{x-y+z}{2}\right) - f(x) - f(z), t\right) \geq \frac{t}{t + \theta\left(\|x\|^p + \|y\|^p + \|z\|^p\right)}$$

for all $x, y, z \in X$ and all $t > 0$. Then, the limit

$$A(x) := N - \lim_{n\to\infty} 2^n f\left(\frac{x}{2^n}\right) \tag{69}$$

exists for each $x \in X$ and defines a unique additive mapping $A : X \to Y$ such that

$$N(f(x) - A(x), t) \geq \frac{(2^{p+1} - 2)t}{(2^{p+1} - 2)t + (2^r + 2)\theta \|x\|^p}$$

for all $x \in X$ and $t > 0$.

Proof The proof follows from Theorem 14 by taking $\varphi(x, y, z) := \theta(\|x\|^p + \|y\|^p + \|z\|^p)$ for all $x, y, z \in X$. Then we can choose $L = 2^{-p}$ and we get the desired result.

Theorem 15 *Let $\varphi : X^3 \to [0, \infty)$ be a function such that there exists an $L < 1$ with*

$$\varphi(x, y, z) \leq 2L\varphi\left(\frac{x}{2}, \frac{y}{2}, \frac{z}{2}\right)$$

for all $x, y, z \in X$. Let $f : X \to Y$ be a mapping with $f(0) = 0$ satisfying (64). Then

$$A(x) := N - \lim_{n\to\infty} \frac{f(2^n x)}{2^n}$$

exists for each $x \in X$ and defines a unique additive mapping $A : X \to Y$ such that

$$N(f(x) - A(x), t) \geq \frac{(2 - 2L)t}{(2 - 2L)t + \varphi(x, 2x, x)} \tag{70}$$

for all $x \in X$ and all $t > 0$.

Proof Let (S, d) be the generalized metric space defined as in the proof of Theorem 14. Consider the linear mapping $J : S \to S$ such that $Jg(x) := \frac{g(2x)}{2}$ for all $x \in X$. Let $g, h \in S$ be such that $d(g, h) = \varepsilon$. Then

$$N(g(x) - h(x), \varepsilon t) \geq \frac{t}{t + \varphi(x, 2x, x)}$$

for all $x \in X$ and $t > 0$. Hence

$$\begin{aligned} N(Jg(x) - Jh(x), L\varepsilon t) &= N\left(\frac{g(2x)}{2} - \frac{h(2x)}{2}, L\varepsilon t\right) \\ &= N\Big(g(2x) - h(2x), 2L\varepsilon t\Big) \geq \frac{2Lt}{2Lt + \varphi(2x, , 4x, 2x)} \\ &\geq \frac{2Lt}{2Lt + 2L\varphi(x, 2x, x)} = \frac{t}{t + \varphi(x, 2x, x)} \end{aligned}$$

for all $x \in X$ and $t > 0$. Thus $d(g, h) = \varepsilon$ implies that $d(Jg, Jh) \leq L\varepsilon$. This means that

$$d(Jg, Jh) \leq Ld(g, h)$$

for all $g, h \in S$. It follows from (66) that

$$N\left(\frac{f(2x)}{2} - f(x), \frac{t}{2}\right) \geq \frac{t}{t + \varphi(x, 2x, x)}.$$

Therefore

$$d(f, Jf) \leq \frac{1}{2}.$$

By Theorem 3, there exists a mapping $A : X \to Y$ satisfying the following:

(1) A is a fixed point of J, that is,

$$2A(x) = A(2x) \tag{71}$$

for all $x \in X$. The mapping A is a unique fixed point of J in the set $\Omega = \{h \in S : d(g, h) < \infty\}$. This implies that A is a unique mapping satisfying (71) such that there exists $\mu \in (0, \infty)$ satisfying

$$N(f(x) - A(x), \mu t) \geq \frac{t}{t + \varphi(x, 2x, x)}$$

for all $x \in X$ and $t > 0$.

(2) $d(J^n f, A) \to 0$ as $n \to \infty$. This implies the equality $N - \lim_{n\to\infty} \frac{f(2^n x)}{2^n}$ for all $x \in X$.

(3) $d(f, A) \leq \frac{d(f, Jf)}{1-L}$ with $f \in \Omega$, which implies the inequality $d(f, A) \leq \frac{1}{2-2L}$. This implies that the inequality (70) holds.

The rest of the proof is similar to that of the proof of Theorem 14.

Corollary 12 *Let $\theta \geq 0$ and let p be a real number with $0 < p < \frac{1}{3}$. Let X be a normed vector space with norm $\|.\|$. Let $f : X \to Y$ be a mapping with $f(0) = 0$ satisfying*

$$N\left(f\left(\frac{x+y+z}{2}\right) + f\left(\frac{x-y+z}{2}\right) - f(x) - f(z), t\right) \geq \frac{t}{t + \theta\left(\|x\|^p.\|y\|^p.\|z\|^p\right)}$$

for all $x, y, z \in X$ and all $t > 0$. Then

$$A(x) := N - \lim_{n\to\infty} \frac{f(2^n x)}{2^n}$$

exists for each $x \in X$ and defines a unique additive mapping $A : X \to Y$ such that

$$N(f(x) - A(x), t) \geq \frac{(2^{1+3p} - 2)t}{(2^{1+3p} - 2)t + 2^{3p}\theta\|x\|^{3p}}.$$

for all $x \in X$.

Proof The proof follows from Theorem 15 by taking

$$\varphi(x, y, z) := \theta\left(\|x\|^p.\|y\|^p.\|z\|^p\right)$$

for all $x, y, z \in X$. Then we can choose $L = 2^{-3p}$ and we get the desired result.

References

1. T. Aoki, On the stability of the linear transformation in Banach spaces. J. Math. Soc. Jpn. **2**, 64–66 (1950)
2. L.M. Arriola, W.A. Beyer, Stability of the Cauchy functional equation over p-adic fields. Real Anal. Exch. **31**, 125–132 (2005/2006)
3. T. Bag, S.K. Samanta, Finite dimensional fuzzy normed linear spaces. J. Fuzzy Math. **11**, 687–705 (2003)
4. T. Bag, S.K. Samanta, Fuzzy bounded linear operators. Fuzzy Sets Syst. **151**, 513–547 (2005)
5. L. Cădariu, V. Radu, Fixed points and the stability of Jensen's functional equation. J. Inequal. Pure Appl. Math. **4**(1), Article ID 4 (2003)
6. L. Cădariu, V. Radu, On the stability of the Cauchy functional equation: a fixed point approach. Grazer Math. Ber. **346**, 43–52 (2004)
7. L. Cădariu, V. Radu, Fixed point methods for the generalized stability of functional equations in a single variable. Fixed Point Theory Appl. **2008**, Article ID 749392 (2008)
8. S.C. Cheng, J.N. Mordeson, Fuzzy linear operators and fuzzy normed linear spaces. Bull. Calcutta Math. Soc. **86**, 429–436 (1994)
9. P.W. Cholewa, Remarks on the stability of functional equations. Aequationes Math. **27**, 76–86 (1984)
10. J. Chung, P.K. Sahoo, On the general solution of a quartic functional equation. Bull. Korean Math. Soc. **40**, 565–576 (2003)
11. S. Czerwik, *Functional Equations and Inequalities in Several Variables* (World Scientific, River Edge, 2002)
12. D. Deses, On the representation of non-Archimedean objects. Topol. Appl. **153**, 774–785 (2005)
13. J. Diaz, B. Margolis, A fixed point theorem of the alternative for contractions on a generalized complete metric space. Bull. Am. Math. Soc. **74**, 305–309 (1968)
14. M. Eshaghi Gordji, M. Bavand Savadkouhi, Stability of mixed type cubic and quartic functional equations in random normed spaces. J. Inequal. Appl. **2009**, Article ID 527462, 9 pp. (2009)
15. M. Eshaghi Gordji, H. Khodaei, *Stability of Functional Equations* (Lap Lambert Academic Publishing, Saarbrücken, 2010)

16. M. Eshaghi Gordji, S. Abbaszadeh, C. Park, On the stability of a generalized quadratic and quartic type functional equation in quasi-Banach spaces. J. Inequal. Appl. **2009**, Article ID 153084, 26 pp. (2009)
17. M. Eshaghi Gordji, M. Bavand Savadkouhi, C. Park, Quadratic-quartic functional equations in RN-spaces. J. Inequal. Appl. **2009**, Article ID 868423, 14 pp. (2009)
18. M. Eshaghi Gordji, S. Zolfaghari, J.M. Rassias, M.B. Savadkouhi, Solution and stability of a mixed type cubic and quartic functional equation in quasi-Banach spaces. Abstr. Appl. Anal. **2009**, Article ID 417473, 14 pp. (2009)
19. W. Fechner, Stability of a functional inequality associated with the Jordan-von Neumann functional equation. Aequationes Math. **71**, 149–161 (2006)
20. P. Găvruta, A generalization of the Hyers-Ulam-Rassias stability of approximately additive mappings. J. Math. Anal. Appl. **184**, 431–436 (1994)
21. K. Hensel, Ubereine news Begrundung der Theorie der algebraischen Zahlen. Jahresber. Deutsch. Math. Verein **6**, 83–88 (1897)
22. D.H. Hyers, On the stability of the linear functional equation. Proc. Natl. Acad. Sci. U. S. A. **27**, 222–224 (1941)
23. D.H. Hyers, G. Isac, Th.M. Rassias, *Stability of Functional Equations in Several Variables* (Birkhäuser, Basel, 1998)
24. K. Jun, H. Kim, J.M. Rassias, Extended Hyers-Ulam stability for Cauchy-Jensen mappings. J. Differ. Equ. Appl. **13**, 1139–1153 (2007)
25. I. Karmosil, J. Michalek, Fuzzy metric and statistical metric spaces. Kybernetica **11**, 326–334 (1975)
26. A.K. Katsaras, Fuzzy topological vector spaces. Fuzzy Sets Syst. **12**, 143–154 (1984)
27. A.K. Katsaras, A. Beoyiannis, Tensor products of non-Archimedean weighted spaces of continuous functions. Georgian Math. J. **6**, 33–44 (1999)
28. H.A. Kenary, On the stability of a cubic functional equation in random normed spaces. J. Math. Ext. **4**, 1–11 (2009)
29. H.A. Kenary, Stability of a Pexiderial functional equation in random normed spaces. Rend. Circ. Mat. Palermo **60**, 59–68 (2011)
30. A. Khrennikov, *Non-Archimedean Analysis: Quantum Paradoxes, Dynamical Systems and Biological Models*. Mathematics and Its Applications, vol. 427 (Kluwer Academic Publishers, Dordrecht, 1997)
31. Z. Kominek, On a local stability of the Jensen functional equation. Demonstratio Math. **22**, 499–507 (1989)
32. S.V. Krishna, K.K.M. Sarma, Separation of fuzzy normed linear spaces. Fuzzy Sets Syst. **63**, 207–217 (1994)
33. D. Mihet, V. Radu, On the stability of the additive Cauchy functional equation in random normed spaces. J. Math. Anal. Appl. **343**, 567–572 (2008)
34. M. Mohammadi, Y.J. Cho, C. Park, P. Vetro, R. Saadati, Random stability of an additive-quadratic-quartic functional equation. J. Inequal. Appl. **2010**, Article ID 754210, 18 pp. (2010)
35. A. Najati, C. Park, The Pexiderized Apollonius-Jensen type additive mapping and isomorphisms between C^*-algebras. J. Differ. Equ. Appl. **14**, 459–479 (2008)
36. P.J. Nyikos, On some non-Archimedean spaces of Alexandrof and Urysohn. Topol. Appl. **91**, 1–23 (1999)
37. C. Park, Generalized Hyers-Ulam-Rassias stability of n-sesquilinear-quadratic mappings on Banach modules over C^*-algebras. J. Comput. Appl. Math. **180**, 279–291 (2005)
38. C. Park, Fixed points and Hyers-Ulam-Rassias stability of Cauchy-Jensen functional equations in Banach algebras. Fixed Point Theory Appl. **2007**, Article ID 50175 (2007)
39. C. Park, Generalized Hyers-Ulam-Rassias stability of quadratic functional equations: a fixed point approach. Fixed Point Theory Appl. **2008**, Article ID 493751 (2008)
40. C. Park, Fuzzy stability of a functional equation associated with inner product spaces. Fuzzy Sets Syst. **160**, 1632–1642 (2009)
41. J.C. Parnami, H.L. Vasudeva, On Jensen's functional equation. Aequationes Math. **43**, 211–218 (1992)

42. V. Radu, The fixed point alternative and the stability of functional equations. Fixed Point Theory **4**, 91–96 (2003)
43. Th.M. Rassias, On the stability of the linear mapping in Banach spaces. Proc. Am. Math. Soc. **72**, 297–300 (1978)
44. Th.M. Rassias, Problem 16, in *Report of the 27th International Symposium on Functional Equations*. Aequations Mathematicae, vol. 39 (1990), pp. 292–293
45. Th.M. Rassias, On the stability of the quadratic functional equation and its applications. Stud. Univ. Babes-Bolyai. **XLIII**, 89–124 (1998)
46. Th.M. Rassias, The problem of S.M. Ulam for approximately multiplicative mappings. J. Math. Anal. Appl. **246**, 352–378 (2000)
47. Th.M. Rassias, On the stability of functional equations in Banach spaces. J. Math. Anal. Appl. **251**, 264–284 (2000)
48. Th.M. Rassias, *Functional Equations, Inequalities and Applications* (Kluwer Academic Publishers Co., Dordrecht, 2003)
49. Th.M. Rassias, P. Semrl, On the behaviour of mappings which do not satisfy Hyers-Ulam stability. Proc. Am. Math. Soc. **114**, 989–993 (1992)
50. Th.M. Rassias, P. Semrl, On the Hyers-Ulam stability of linear mappings. J. Math. Anal. Appl. **173**, 325–338 (1993)
51. J. Rätz, On inequalities associated with the Jordan-von Neumann functional equation. Aequationes Math. **66**, 191–200 (2003)
52. R. Saadati, C. Park, Non-Archimedean $\mathscr{L}$-fuzzy normed spaces and stability of functional equations. Comput. Math. Appl. **60**, 2488–2496 (2010)
53. R. Saadati, M. Vaezpour, Y.J. Cho, A note to paper "On the stability of cubic mappings and quartic mappings in random normed spaces". J. Inequal. Appl. **2009**, Article ID 214530 (2009). https://doi.org/10.1155/2009/214530
54. R. Saadati, M.M. Zohdi, S.M. Vaezpour, Nonlinear L-random stability of an ACQ-functional equation. J. Inequal. Appl. **2011**, Article ID 194394, 23 pages (2011). https://doi.org/10.1155/2011/194394
55. B. Schewizer, A. Sklar, *Probabilistic Metric Spaces*. North-Holland Series in Probability and Applied Mathematics (North-Holland, New York, 1983)
56. F. Skof, Local properties and approximation of operators. Rend. Sem. Mat. Fis. Milano **53**, 113–129 (1983)
57. S.M. Ulam, *Problems in Modern Mathematics*, Science Editions (Wiley, New York, 1964)

NAN-RN Approximately Generalized Additive Functional Equations

Hassan Azadi Kenary and Themistocles M. Rassias

1 Introduction and Preliminaries

A classical question in the theory of functional equations is the following: *When is it true that a function which approximately satisfies a functional equation must be close to an exact solution of the equation?*

If the problem accepts a solution, we say that the equation is *stable*. The first stability problem concerning group homomorphisms was raised by Ulam [45] in 1940.

In the next year, Hyers [15] gave a positive answer to the above question for additive groups under the assumption that the groups are Banach spaces. In 1978, Rassias [31] proved a generalization of Hyers's theorem for additive mappings.

Theorem 1 (Th.M. Rassias) *Let f be an approximately additive mapping from a normed vector space E into a Banach space E', i.e., f satisfies the inequality*

$$|f(x+y) - f(x) - f(y)| \leq \varepsilon(\|x\|^r + \|y\|^r)$$

for all $x, y \in E$, where ε and r are constants with $\varepsilon > 0$ and $0 \leq r < 1$. Then the mapping $L : E \to E'$ defined by $L(x) := \lim_{n\to\infty} 2^{-n} f(2^n x)$ is the unique additive mapping which satisfies

H. Azadi Kenary (✉)
Department of Mathematics, College of Sciences, Yasouj University, Yasouj, Iran
e-mail: azadi@yu.ac.ir

Th. M. Rassias
Department of Mathematics, National Technical University of Athens, Athens, Greece
e-mail: trassias@math.ntua.gr

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_16

$$|f(x) - L(x)| \leq \frac{2\varepsilon|x|^r}{2 - 2^r}$$

for all $x \in E$.

The result of Rassias has influenced the development of what is now called the *Hyers-Ulam-Rassias stability problem* for functional equations. In 1994, a generalization of Rassias' theorem was obtained by Găvruta [13] by replacing the bound $\varepsilon(\|x\|^p + \|y\|^p)$ by a general control function $\phi(x, y)$.

The functional equation

$$f(x + y) + f(x - y) = 2f(x) + 2f(y)$$

is called a *quadratic functional equation*. In particular, every solution of the quadratic functional equation is said to be a *quadratic mapping*. In 1983, a generalized Hyers-Ulam stability problem for the quadratic functional equation was proved by Skof [44] for mappings $f : X \to Y$, where X is a normed space and Y is a Banach space. In 1984, Cholewa [6] noticed that the theorem of Skof is still true if the relevant domain X is replaced by an Abelian group and, in 2002, Czerwik [7] proved the generalized Hyers-Ulam stability of the quadratic functional equation.

The stability problems of several functional equations have been extensively investigated by a number of authors and there are many interesting results concerning this problem [1–5, 9–12, 16–19, 23–25, 27–42].

In 1897, Hensel [14] has introduced a normed space which does not have the Archimedean property. It turned out that non-Archimedean spaces have many nice applications (see [8, 20, 21, 26]).

A *valuation* is a function $|\cdot|$ from a field $\mathbb{K}$ into $[0, \infty)$ such that 0 is the unique element having the 0 valuation, $|rs| = |r||s|$ and the triangle inequality holds, i.e.,

$$|r + s| \leq \max\{|r|, |s|\}.$$

A field $\mathbb{K}$ is called a *valued field* if $\mathbb{K}$ carries a valuation. The usual absolute values of $\mathbb{R}$ and $\mathbb{C}$ are examples of valuations.

Let us consider a valuation which satisfies a stronger condition than the triangle inequality. If the triangle inequality is replaced by

$$|r + s| \leq \max\{|r|, |s|\}$$

for all $r, s \in \mathbb{K}$, then the function $|\cdot|$ is called a *non-Archimedean valuation* and the field is called a *non-Archimedean field*. Clearly, $|1| = |-1| = 1$ and $|n| \leq 1$ for all $n \geq 1$. A trivial example of a non-Archimedean valuation is the function $|\cdot|$ taking everything except for 0 into 1 and $|0| = 0$.

Definition 1 Let X be a vector space over a field $\mathbb{K}$ with a non-Archimedean valuation $|\cdot|$. A function $\|\cdot\| : X \to [0, \infty)$ is called a *non-Archimedean norm* if the following conditions hold:

(a) $\|x\| = 0$ if and only if $x = 0$ for all $x \in X$;
(b) $\|rx\| = |r|\|x\|$ for all $r \in \mathbb{K}$ and $x \in X$;
(c) the strong triangle inequality holds:

$$\|x + y\| \leq \max\{\|x\|, \|y\|\}$$

for all $x, y \in X$.

Then $(X, \|\cdot\|)$ is called a *non-Archimedean normed space*.

Definition 2 Let $\{x_n\}$ be a sequence in a non-Archimedean normed space X.

(a) A sequence $\{x_n\}_{n=1}^{\infty}$ in a non-Archimedean space is a *Cauchy sequence* iff, the sequence $\{x_{n+1} - x_n\}_{n=1}^{\infty}$ converges to zero.
(b) The sequence $\{x_n\}$ is said to be *convergent* if, for any $\varepsilon > 0$, there are a positive integer N and $x \in X$ such that

$$\|x_n - x\| \leq \varepsilon$$

for all $n \geq N$. Then the point $x \in X$ is called the *limit* of the sequence $\{x_n\}$, which is denote by $\lim_{n\to\infty} x_n = x$.
(c) If every Cauchy sequence in X converges, then the non-Archimedean normed space X is called a *non-Archimedean Banach space*.

In the sequel, we adopt the usual terminology, notions and conventions of the theory of random normed spaces as in [43].

Throughout this paper (in random stability section), let Γ^+ denote the set of all probability distribution functions $F : \mathbb{R} \cup [-\infty, +\infty] \to [0, 1]$ such that F is left-continuous and nondecreasing on $\mathbb{R}$ and $F(0) = 0$, $F(+\infty) = 1$. It is clear that the set

$$D^+ = \{F \in \Gamma^+ : l^- F(-\infty) = 1\},$$

where $l^- f(x) = \lim_{t\to x^-} f(t)$, is a subset of Γ^+. The set Γ^+ is partially ordered by the usual point-wise ordering of functions, that is, $F \leq G$ if and only if $F(t) \leq G(t)$ for all $t \in \mathbb{R}$. For any $a \geq 0$, the element $H_a(t)$ of D^+ is defined by

$$H_a(t) = \begin{cases} 0, \text{ if } t \leq a, \\ 1, \text{ if } t > a. \end{cases}$$

We can easily show that the maximal element in Γ^+ is the distribution function $H_0(t)$.

Definition 3 A function $T : [0, 1]^2 \to [0, 1]$ is a *continuous triangular norm* (briefly, a t-norm) if T satisfies the following conditions:

(a) T is commutative and associative;
(b) T is continuous;

(c) $T(x, 1) = x$ for all $x \in [0, 1]$;
(d) $T(x, y) \le T(z, w)$ whenever $x \le z$ and $y \le w$ for all $x, y, z, w \in [0, 1]$.

Three typical examples of continuous t-norms are as follows:

$$T(x, y) = xy, \quad T(x, y) = \max\{a + b - 1, 0\}, \quad T(x, y) = \min(a, b).$$

Recall that, if T is a t-norm and $\{x_n\}$ is a sequence in $[0, 1]$, then $T_{i=1}^{n} x_i$ is defined recursively by $T_{i=1}^{1} x_1 = x_1$ and $T_{i=1}^{n} x_i = T(T_{i=1}^{n-1} x_i, x_n)$ for all $n \ge 2$. $T_{i=n}^{\infty} x_i$ is defined by $T_{i=1}^{\infty} x_{n+i}$.

Definition 4 A random normed space (briefly, RN-space) is a triple (X, μ, T), where X is a vector space, T is a continuous t-norm and $\mu : X \to D^+$ is a mapping such that the following conditions hold:

(a) $\mu_x(t) = H_0(t)$ for all $x \in X$ and $t > 0$ if and only if $x = 0$;
(b) $\mu_{\alpha x}(t) = \mu_x(\frac{t}{|\alpha|})$ for all $\alpha \in \mathbb{R}$ with $\alpha \neq 0$, $x \in X$ and $t \ge 0$;
(c) $\mu_{x+y}(t + s) \ge T(\mu_x(t), \mu_y(s))$ for all $x, y \in X$ and $t, s \ge 0$.

Every normed space $(X, \|\cdot\|)$ defines a random normed space (X, μ, T_M), where $\mu_u(t) = \frac{t}{t+\|u\|}$ for all $t > 0$ and T_M is the minimum t-norm. This space X is called the *induced random normed space*.

If the t-norm T is such that $\sup_{0<a<1} T(a, a) = 1$, then every RN-space (X, μ, T) is a metrizable linear topological space with the topology τ (called the *μ-topology* or the *(ε, δ)-topology*, where $\varepsilon > 0$ and $\lambda \in (0, 1)$) induced by the base $\{U(\varepsilon, \lambda)\}$ of neighborhoods of θ, where

$$U(\varepsilon, \lambda) = \{x \in X : \mu_x(\varepsilon) > 1 - \lambda\}.$$

Definition 5 Let (X, μ, T) be an RN-space.

(1) A sequence $\{x_n\}$ in X is said to be *convergent* to a point $x \in X$ (write $x_n \to x$ as $n \to \infty$) if $\lim_{n\to\infty} \mu_{x_n - x}(t) = 1$ for all $t > 0$.
(2) A sequence $\{x_n\}$ in X is called a *Cauchy sequence* in X if $\lim_{n\to\infty} \mu_{x_n - x_m}(t) = 1$ for all $t > 0$.
(3) The RN-space (X, μ, T) is said to be *complete* if every Cauchy sequence in X is convergent.

Theorem 2 ([43]) *If (X, μ, T) is an RN-space and $\{x_n\}$ is a sequence such that $x_n \to x$, then* $\lim_{n\to\infty} \mu_{x_n}(t) = \mu_x(t)$.

Definition 6 Let X be a set. A function $d : X \times X \to [0, \infty]$ is called a generalized metric on X if d satisfies the following conditions:

(a) $d(x, y) = 0$ if and only if $x = y$ for all $x, y \in X$;
(b) $d(x, y) = d(y, x)$ for all $x, y \in X$;
(c) $d(x, z) \le d(x, y) + d(y, z)$ for all $x, y, z \in X$.

Theorem 3 *Let (X,d) be a complete generalized metric space and $J : X \to X$ be a strictly contractive mapping with Lipschitz constant $L < 1$. Then, for all $x \in X$, either*

$$d(J^n x, J^{n+1} x) = \infty \tag{1}$$

for all nonnegative integers n or there exists a positive integer n_0 such that

(a) $d(J^n x, J^{n+1} x) < \infty$ *for all* $n_0 \geq n_0$*;*
(b) the sequence $\{J^n x\}$ *converges to a fixed point* y^* *of* J*;*
(c) y^* *is the unique fixed point of* J *in the set* $Y = \{y \in X : d(J^{n_0} x, y) < \infty\}$*;*
(d) $d(y, y^*) \leq \frac{1}{1-L} d(y, Jy)$ *for all* $y \in Y$*.*

In this paper, we prove the generalized Hyers-Ulam-Rassias stability of the following functional equation:

$$f\Big(\sum_{i=1}^{m} \alpha_i x_i\Big) = \Big[\prod_{i=1}^{m} f(x_i)\Big] . \Big[\sum_{i=1}^{m} \Big[\alpha_i \Big(\prod_{j=1,\ j\neq i}^{m} f(x_j)\Big)\Big]\Big]^{-1} \tag{2}$$

for arbitrary but fixed real numbers $(\alpha_1, \alpha_2, \cdots, \alpha_m) \neq (0, 0, \cdots, 0)$, so that $\alpha = \sum_{i=1}^{m} \alpha_i > 1$ and $|\alpha| \neq 1$, in various spaces.

2 Non-Archimedean Stability of Eq. (2)

In this section, using the fixed point alternative approach, we prove the generalized Hyers-Ulam stability of functional equation (2) in non-Archimedean Spaces. Throughout this paper, assume that X is a non-Archimedean normed vector space and that Y is a non-Archimedean Banach space.

Theorem 4 *Let $\zeta : X^m \to [0, \infty)$ be a function such that there exists $L < 1$ with*

$$\zeta\Big(\frac{x_1}{\alpha}, \frac{x_2}{\alpha}, \cdots, \frac{x_m}{\alpha}\Big) \leq |\alpha| L \zeta(x_1, x_2, \cdots, x_m) \tag{3}$$

for all $x_1, x_2, \cdots, x_m \in X$. If $f : X \to Y$ is a mapping satisfying

$$\Big\| f\Big(\sum_{i=1}^{m} \alpha_i x_i\Big) - \Big[\prod_{i=1}^{m} f(x_i)\Big] . \Big[\sum_{i=1}^{m} \Big[\alpha_i \Big(\prod_{j=1,\ j\neq i}^{m} f(x_j)\Big)\Big]\Big]^{-1} \Big\| \leq \zeta(x_1, x_2, \cdots, x_m). \tag{4}$$

for all $x_1, x_2, \cdots, x_m \in X$, then there is a unique mapping $C : X \to Y$ such that

$$\| f(x) - C(x) \| \leq \frac{|\alpha| L \zeta(x, x, \cdots, x)}{1 - L}. \tag{5}$$

Proof Putting $x_1 = x_2 = \cdots = x_m = x$ in (4), we get

$$\left\| \frac{1}{\alpha} f\left(\frac{x}{\alpha}\right) - f(x) \right\| \leq \zeta\left(\frac{x}{\alpha}, \frac{x}{\alpha}, \cdots, \frac{x}{\alpha}\right) \tag{6}$$

for all $x \in G$. Consider the set

$$S := \{g : X \to Y\}$$

and the generalized metric d in S defined by

$$d(f, g) = \inf_{\mu \in (0,+\infty)} \Big\{ \|g(x) - h(x)\| \leq \mu\zeta(x, x, \cdots, x), \ \forall x \in X \Big\}, \tag{7}$$

where $\inf \emptyset = +\infty$. It is easy to show that (S, d) is complete (see [22], Lemma 2.1). Now, we consider a linear mapping $J : S \to S$ such that

$$Jh(x) := \frac{1}{\alpha} h\left(\frac{x}{\alpha}\right) \tag{8}$$

for all $x \in X$. Let $g, h \in S$ be such that $d(g, h) = \varepsilon$. Then we have

$$\|g(x) - h(x)\| \leq \varepsilon\zeta(x, x, \cdots, x) \tag{9}$$

for all $x \in X$ and so

$$\begin{aligned} \|Jg(x) - Jh(x)\| &= \left\| \frac{1}{\alpha} g\left(\frac{x}{\alpha}\right) - \frac{1}{\alpha} h\left(\frac{x}{\alpha}\right) \right\| \\ &\leq \frac{1}{|\alpha|} \varepsilon\zeta\left(\frac{x}{\alpha}, \frac{x}{\alpha}, \cdots, \frac{x}{\alpha}\right) \\ &\leq L\varepsilon \end{aligned}$$

for all $x \in X$. Thus $d(g, h) = \varepsilon$ implies that $d(Jg, Jh) \leq L\varepsilon$. This means that $d(Jg, Jh) \leq Ld(g, h)$ for all $g, h \in S$. It follows from (6) that

$$d(f, Jf) \leq |\alpha| L. \tag{10}$$

By Theorem 3, there exists a mapping $C : X \to Y$ satisfying the following:

(1) C is a fixed point of J, that is,

$$C\left(\frac{x}{\alpha}\right) = \alpha C(x) \tag{11}$$

for all $x \in X$. The mapping C is a unique fixed point of J in the set

$$\Omega = \{h \in S : d(g,h) < \infty\}.$$

This implies that C is a unique mapping satisfying (11) such that there exists $\mu \in (0,\infty)$ satisfying

$$\|f(x) - C(x)\| \leq \mu\zeta(x, x, \cdots, x) \tag{12}$$

for all $x \in X$.

(2) $d(J^n f, C) \to 0$ as $n \to \infty$. This implies the equality

$$\lim_{n\to\infty} \frac{1}{\alpha^n} f\Big(\frac{x}{\alpha^n}\Big) = C(x) \tag{13}$$

for all $x \in X$.

(3) $d(f,C) \leq \frac{d(f,Jf)}{1-L}$ with $f \in \Omega$, which implies the inequality

$$d(f,C) \leq \frac{|\alpha|L}{1-L}.$$

This implies that the inequality (5) holds.

It follows from (3) and (4) that

$$\begin{aligned}
&\left\| C\Big(\sum_{i=1}^{m} \alpha_i x_i\Big) - \Big[\prod_{i=1}^{m} C(x_i)\Big]. \Big[\sum_{i=1}^{m} \Big[\alpha_i\Big(\prod_{j=1,\ j\neq i}^{m} C(x_j)\Big)\Big]\Big]^{-1} \right\| \\
&= \lim_{n\to\infty} \frac{1}{|\alpha|^n} \left\| f\left(\sum_{i=1}^{m} \frac{\alpha_i x_i}{\alpha^n}\right) - \Big[\prod_{i=1}^{m} f\left(\frac{x_i}{\alpha^n}\right)\Big]. \Big[\sum_{i=1}^{m} \Big[\alpha_i\Big(\prod_{j=1,\ j\neq i}^{m} f\left(\frac{x_j}{\alpha^n}\right)\Big)\Big]\Big]^{-1} \right\| \\
&\leq \lim_{n\to\infty} \frac{1}{|\alpha|^n} \zeta\left(\frac{x_1}{\alpha^n}, \frac{x_2}{\alpha^n}, \cdots, \frac{x_m}{\alpha^n}\right) \\
&\leq \lim_{n\to\infty} \frac{1}{|\alpha|^n} \cdot |\alpha|^n L^n \zeta(x_1, x_2, \cdots, x_m) \\
&= 0
\end{aligned}$$

for all $x_1, x_2, \cdots, x_m \in X$. So

$$C\Big(\sum_{i=1}^{m} \alpha_i x_i\Big) - \Big[\prod_{i=1}^{m} C(x_i)\Big]. \Big[\sum_{i=1}^{m} \Big[\alpha_i\Big(\prod_{j=1,\ j\neq i}^{m} C(x_j)\Big)\Big]\Big]^{-1} = 0$$

for all $x_1, x_2, \cdots, x_m \in X$. Hence $C : X \to Y$ satisfying (2). This completes the proof.

Corollary 1 *Let $\theta \geq 0$ and r be a real number with $0 < r < 1$. Let $f : X \to Y$ be a mapping satisfying*

$$\left\| f\Big(\sum_{i=1}^{m} \alpha_i x_i\Big) - \Big[\prod_{i=1}^{m} f(x_i)\Big] . \Big[\sum_{i=1}^{m} \Big[\alpha_i\Big(\prod_{j=1,\ j\neq i}^{m} f(x_j)\Big)\Big]\Big]^{-1} \right\| \leq \theta\Big(\sum_{i=1}^{m} \|x_i\|^r\Big) \tag{14}$$

for all $x_1, x_2, \cdots, x_m \in X$. Then the limit

$$C(x) = \lim_{n\to\infty} \frac{1}{\alpha^n} f\Big(\frac{x}{\alpha^n}\Big)$$

exists for all $x \in X$ and defines a unique mapping $C : X \to Y$ such that

$$\|f(x) - C(x)\| \leq \begin{cases} \frac{m|\alpha|^2\theta\|x\|^r}{|\alpha|^r - |\alpha|} & \textit{if } |\alpha| < 1, \\ \\ \frac{m|\alpha|^{r+1}\theta\|x\|^r}{|\alpha| - |\alpha|^r} & \textit{if } |\alpha| > 1. \end{cases} \tag{15}$$

for all $x \in X$.

Proof The proof follows from Theorem 4 if we take

$$\zeta(x_1, x_2, \cdots, x_m) = \theta\Big(\sum_{i=1}^{m} \|x_i\|^r\Big) \tag{16}$$

for all $x_1, x_2, \cdots, x_m \in X$. In fact, if we choose $L = \begin{cases} |\alpha|^{1-r} \text{ if } |\alpha| < 1 \\ |\alpha|^{r-1} \text{ if } |\alpha| > 1 \end{cases}$, then we get the desired result.

Corollary 2 *Let $\theta \geq 0$ and r_i be positive real numbers with $0 < r = \sum_{i=1}^{m} r_i < 1$. Let $f : X \to Y$ be a mapping satisfying*

$$\left\| f\Big(\sum_{i=1}^{m} \alpha_i x_i\Big) - \Big[\prod_{i=1}^{m} f(x_i)\Big] . \Big[\sum_{i=1}^{m} \Big[\alpha_i\Big(\prod_{j=1,\ j\neq i}^{m} f(x_j)\Big)\Big]\Big]^{-1} \right\| \leq \theta\Big(\prod_{i=1}^{m} \|x_i\|^{r_i}\Big) \tag{17}$$

for all $x_1, x_2, \cdots, x_m \in X$. Then the limit

$$C(x) = \lim_{n\to\infty} \frac{1}{\alpha^n} f\Big(\frac{x}{\alpha^n}\Big)$$

exists for all $x \in X$ and defines a unique mapping $C : X \to Y$ such that

$$\|f(x) - C(x)\| \leq \begin{cases} \frac{|\alpha|\theta\|x\|^r}{|\alpha|^{r-1} - 1} & \textit{if } |\alpha| < 1 \\ \\ \frac{|\alpha|^r\theta\|x\|^r}{1 - |\alpha|^{r-1}} & \textit{if } |\alpha| > 1 \end{cases} \tag{18}$$

for all $x \in X$.

Proof The proof follows from Theorem 4 if we take

$$\zeta(x_1, x_2, \cdots, x_m) = \theta\Big(\prod_{i=1}^{m} \|x_i\|^{r_i}\Big) \tag{19}$$

for all $x_1, x_2, \cdots, x_m \in X$. In fact, if we choose $L = \begin{cases} |\alpha|^{1-r} \text{ if } |\alpha| < 1 \\ |\alpha|^{r-1} \text{ if } |\alpha| > 1 \end{cases}$, then we get the desired result.

Theorem 5 *Let $\zeta : X^m \to [0, \infty)$ be a function such that there exists an $L < 1$ with*

$$\zeta(\alpha x_1, \alpha x_2, \cdots, \alpha x_m) \leq \frac{L}{|\alpha|}\zeta(x_1, x_2, \cdots, x_m) \tag{20}$$

for all $x_1, x_2, \cdots, x_m \in X$. Let $f : X \to Y$ be a mapping satisfying

$$\Big\| f\Big(\sum_{i=1}^{m} \alpha_i x_i\Big) - \Big[\prod_{i=1}^{m} f(x_i)\Big] . \Big[\sum_{i=1}^{m} \Big[\alpha_i\Big(\prod_{j=1,\ j\neq i}^{m} f(x_j)\Big)\Big]\Big]^{-1} \Big\| \leq \zeta(x_1, x_2, \cdots, x_m). \tag{21}$$

Then the limit

$$C(x) := \lim_{n\to+\infty} \alpha^n f(\alpha^n x)$$

exists for all $x \in X$ and defines a unique mapping $C : X \to Y$ such that

$$\|f(x) - C(x)\| \leq \frac{|\alpha|\zeta(x, x, \cdots, x)}{1 - L} \tag{22}$$

for all $x \in X$.

Proof It follows from (6) that

$$\|f(x) - \alpha f(\alpha x)\| \leq |\alpha|\zeta(x, x, \cdots, x) \tag{23}$$

for all $x \in X$. Let (S, d) be the generalized metric space defined in the proof of the Theorem (4). Consider a linear mapping $J : S \to S$ such that

$$Jh(x) := \alpha h(\alpha x) \tag{24}$$

for all $x \in X$. It follows from (23) that

$$d(f, Jf) \leq |\alpha| < +\infty. \tag{25}$$

By Theorem (3), there exists a mapping $C : X \to Y$ satisfying the following:

(1) C is a fixed point of J, that is,

$$C(\alpha x) = \frac{1}{\alpha} C(x) \tag{26}$$

for all $x \in X$. The mapping C is a unique fixed point of J in the set

$$\Omega = \{h \in S : d(g, h) < \infty\}.$$

This implies that C is a unique mapping satisfying (26) such that there exists $\mu \in (0, \infty)$ satisfying

$$\|f(x) - C(x)\| \leq \mu \zeta(x, x, \cdots, x) \tag{27}$$

for all $x \in X$.

(2) $d(J^n f, C) \to 0$ as $n \to \infty$. This implies the equality

$$\lim_{n \to \infty} \alpha^n f(\alpha^n x) = C(x) \tag{28}$$

for all $x \in X$.

(3) $d(f, C) \leq \frac{d(f, Jf)}{1-L}$ with $f \in \Omega$, which implies the inequality

$$d(f, C) \leq \frac{|\alpha|}{1 - L}.$$

This implies that the inequality (22) holds. This completes the proof.

Corollary 3 *Let $\theta \geq 0$ and r be a real number with $r > 1$. Let $f : X \to Y$ be a mapping satisfying*

$$\left\| f\Big(\sum_{i=1}^{m} \alpha_i x_i\Big) - \Big[\prod_{i=1}^{m} f(x_i)\Big] . \Big[\sum_{i=1}^{m} \Big[\alpha_i \Big(\prod_{j=1,\ j \neq i}^{m} f(x_j)\Big)\Big]\Big]^{-1} \right\| \leq \theta \Big(\sum_{i=1}^{m} \|x_i\|^r\Big) \tag{29}$$

for all $x_1, x_2, \cdots, x_m \in X$. Then it follows that, for all $x \in X$,

$$C(x) = \lim_{n \to \infty} \alpha^n f(\alpha^n x) \tag{30}$$

exists and defines a unique mapping $C : X \to Y$ such that

$$\|f(x) - C(x)\| \leq \begin{cases} \frac{m|\alpha|^{r+1}\theta\|x\|^r}{|\alpha|^r - |\alpha|} & \textit{if } |\alpha| > 1 \\ \\ \frac{m|\alpha|^2\theta\|x\|^r}{|\alpha| - |\alpha|^r} & \textit{if } |\alpha| < 1 \end{cases} \tag{31}$$

for all $x \in X$.

Proof The proof follows from Theorem 5 if we take

$$\zeta(x_1, x_2, \cdots, x_m) = \theta\Big(\sum_{i=1}^{m} \|x_i\|^r\Big) \tag{32}$$

for all $x_1, x_2, \cdots, x_m \in X$. In fact, if we choose $L = \begin{cases} |\alpha|^{1-r} \text{ if } |\alpha| > 1 \\ |\alpha|^{r-1} \text{ if } |\alpha| < 1 \end{cases}$, then we get the desired result.

Corollary 4 *Let $\theta \geq 0$ and r be positive real number with $r \in (\frac{1}{m}, \infty)$. Let $f : X \to Y$ be a mapping satisfying*

$$\Big\| f\Big(\sum_{i=1}^{m} \alpha_i x_i\Big) - \Big[\prod_{i=1}^{m} f(x_i)\Big] . \Big[\sum_{i=1}^{m}\Big[\alpha_i\Big(\prod_{j=1,\ j\neq i}^{m} f(x_j)\Big)\Big]\Big]^{-1}\Big\| \leq \theta\Big(\prod_{i=1}^{m} \|x_i\|^r\Big) \tag{33}$$

for all $x_1, x_2, \cdots, x_m \in X$. Then the limit

$$C(x) = \lim_{n\to\infty} \alpha^n f(\alpha^n x) \tag{34}$$

exists for all $x \in X$ and defines a unique mapping $C : X \to Y$ such that

$$\|f(x) - C(x)\| \leq \begin{cases} \frac{|\alpha|^{mr+1}\theta\|x\|^r}{|\alpha|^{mr}-|\alpha|} & \textit{if } |\alpha| > 1 \\ \\ \frac{|\alpha|^2\theta\|x\|^{mr}}{|\alpha|-|\alpha|^{mr}} & \textit{if } |\alpha| < 1 \end{cases} \tag{35}$$

for all $x \in X$.

Proof The proof follows from Theorem 4 if we take

$$\zeta(x_1, x_2, \cdots, x_m) = \theta\Big(\prod_{i=1}^{m} \|x_i\|^r\Big) \tag{36}$$

for all $x_1, x_2, \cdots, x_m \in X$. In fact, if we choose $L = \begin{cases} |\alpha|^{mr-1} \text{ if } |\alpha| < 1 \\ |\alpha|^{1-mr} \text{ if } |\alpha| > 1 \end{cases}$, then we get the desired result.

3 Random Stability of Eq. (2)

In this section, we prove the generalized Hyers-Ulam stability of the functional equation (2) in random normed spaces by using direct and fixed point alternative methods.

3.1 Direct Method

Theorem 6 *Let X be a real linear space, $(Z, \mu', \min)$ be an RN-space and $\phi : X^m \to Z$ be a function such that there exists $0 < \beta < \alpha$ such that*

$$\mu'_{\phi\left(\frac{x_1}{\alpha},\frac{x_2}{\alpha},\cdots,\frac{x_m}{\alpha}\right)}(t) \geq \mu'_{\phi(x_1,x_2,\cdots,x_m)}\left(\frac{t}{\beta}\right) \tag{37}$$

for all $x_1, x_2, \cdots, x_m \in X$ and $t > 0$ and

$$\lim_{n\to\infty} \mu'_{\phi\left(\frac{x_1}{\alpha^n},\frac{x_2}{\alpha^n},\cdots,\frac{x_m}{\alpha^n}\right)}(\alpha^n t) = 1$$

for all $x_1, x_2, \cdots, x_m \in X$ and $t > 0$. Let $(Y, \mu, \min)$ be a complete RN-space. If $f : X \to Y$ be a mapping such that

$$\mu_{f(\sum_{i=1}^m \alpha_i x_i)-\left[\prod_{i=1}^m f(x_i)\right].\left[\sum_{i=1}^m \left[\alpha_i\left(\prod_{j=1,\ j\neq i}^m f(x_j)\right)\right]\right]^{-1}}(t) \geq \mu'_{\phi(x_1,x_2,\cdots,x_m)}(t) \tag{38}$$

for all $x_1, x_2, \cdots, x_m \in X$ and $t > 0$. Then the limit

$$C(x) = \lim_{n\to\infty} \frac{1}{\alpha^n} f\left(\frac{x}{\alpha^n}\right)$$

exists for all $x \in X$ and defines a unique mapping $C : X \to Y$ such that

$$\mu_{f(x)-C(x)}(t) \geq \mu'_{\phi(x,x,\cdots,x)}\left(\frac{(\alpha-\beta)t}{\alpha^2}\right). \tag{39}$$

for all $x \in X$ and $t > 0$.

Proof Putting $x_1 = x_2 = \cdots = x_m = x$ in (38), we see that

$$\mu_{f(\alpha x)-\frac{1}{\alpha}f(x)}(t) \geq \mu'_{\phi(x,x,\cdots,x)}(t) \tag{40}$$

for all $x \in X$. Replacing x by $\frac{x}{\alpha^n}$ in (40) and using (37), we obtain

$$\begin{aligned}\mu_{\frac{1}{\alpha^{n-1}}f(\frac{x}{\alpha^{n-1}})-\frac{1}{\alpha^n}f(\frac{x}{\alpha^n})}(t) &\geq \mu'_{\phi(\frac{x}{\alpha^n},\frac{x}{\alpha^n},\cdots,\frac{x}{\alpha^n})}(\alpha^{n-1}t) \\ &\geq \mu'_{\phi(x,x,\cdots,x)}\left(\frac{\alpha^{n-1}t}{\beta^n}\right).\end{aligned} \tag{41}$$

Since

$$\frac{1}{\alpha^n} f\left(\frac{x}{\alpha^n}\right) - f(x) = \sum_{k=0}^{n-1} \frac{1}{\alpha^{k+1}} f\left(\frac{x}{\alpha^{k+1}}\right) - \frac{1}{\alpha^k} f\left(\frac{x}{\alpha^k}\right)$$

therefore

$$\begin{aligned} \mu_{\frac{1}{\alpha^n} f(\frac{x}{\alpha^n}) - f(x)} \left(\sum_{k=0}^{n-1} \frac{t\beta^k}{\alpha^{k-1}}\right) &= \mu_{\sum_{k=0}^{n-1} \frac{1}{\alpha^{k+1}} f\left(\frac{x}{\alpha^{k+1}}\right) - \frac{1}{\alpha^k} f\left(\frac{x}{\alpha^k}\right)} \left(\sum_{k=0}^{n-1} \frac{t\beta^k}{\alpha^{k-1}}\right) \\ &\geq T_{k=0}^{n-1} \left(\mu_{\frac{1}{\alpha^{k+1}} f\left(\frac{x}{\alpha^{k+1}}\right) - \frac{1}{\alpha^k} f\left(\frac{x}{\alpha^k}\right)} \left(\frac{t\beta^k}{\alpha^{k-1}}\right)\right) \quad (42) \\ &\geq T_{k=0}^{n-1} \left(\mu'_{\phi(x,x,\cdots,x)}(t)\right) \\ &= \mu'_{\phi(x,x,\cdots,x)}(t). \end{aligned}$$

This implies that

$$\mu_{\frac{1}{\alpha^n} f(\frac{x}{\alpha^n}) - f(x)}(t) \geq \mu'_{\phi(x,x,\cdots,x)} \left(\frac{t}{\sum_{k=0}^{n-1} \frac{\beta^k}{\alpha^{k-1}}}\right). \quad (43)$$

Replacing x by $\frac{x}{\alpha^p}$ in (43), we obtain

$$\mu_{\frac{1}{\alpha^{n+p}} f\left(\frac{x}{\alpha^{n+p}}\right) - \frac{1}{\alpha^p} f\left(\frac{x}{\alpha^p}\right)}(t) \geq \mu'_{\phi(x,x,\cdots,x)} \left(\frac{t}{\sum_{k=p}^{n+p-1} \frac{\beta^k}{\alpha^{k-1}}}\right). \quad (44)$$

Since

$$\lim_{p,n\to\infty} \mu'_{\phi(x,x,\cdots,x)}\left(\frac{t}{\sum_{k=p}^{n+p-1} \frac{\beta^k}{\alpha^{k-1}}}\right) = 1,$$

it follows that $\left\{\frac{1}{\alpha^n} f(\frac{x}{\alpha^n})\right\}_{n=1}^{+\infty}$ is a Cauchy sequence in complete RN-space $(Y, \mu, \min)$ and so there exists a point $C(x) \in Y$ such that

$$C(x) = \lim_{n\to\infty} \frac{1}{\alpha^n} f\left(\frac{x}{\alpha^n}\right).$$

Fix $x \in X$ and put $p = 0$ in (44). Then we obtain

$$\mu_{\frac{1}{\alpha^n} f(\frac{x}{\alpha^n}) - f(x)}(t) \geq \mu'_{\phi(x,x,\cdots,x)} \left(\frac{t}{\sum_{k=0}^{n-1} \frac{\beta^k}{\alpha^{k-1}}}\right) \quad (45)$$

and so, for any $\varepsilon > 0$,

$$\begin{aligned}\mu_{C(x)-f(x)}(t+\varepsilon) &\geq T\left(\mu_{C(x)-\frac{1}{\alpha^n}f\left(\frac{x}{\alpha^n}\right)}(\varepsilon), \mu_{\frac{1}{\alpha^n}f\left(\frac{x}{\alpha^n}\right)-f(x)}(t)\right) \\ &\geq T\left(\mu_{C(x)-\frac{1}{\alpha^n}f\left(\frac{x}{\alpha^n}\right)}(\varepsilon), \mu'_{\phi(x,x,\cdots,x)}\left(\frac{t}{\sum_{k=0}^{n-1}\frac{\beta^k}{\alpha^{k-1}}}\right)\right).\end{aligned} \tag{46}$$

Taking $n \to \infty$ in (46), we get

$$\mu_{C(x)-f(x)}(t+\varepsilon) \geq \mu'_{\phi(x,x,\cdots,x)}\left(\frac{(\alpha-\beta)t}{\alpha^2}\right). \tag{47}$$

Since ε is arbitrary, by taking $\varepsilon \to 0$ in (47), we get

$$\mu_{C(x)-f(x)}(t) \geq \mu'_{\phi(x,x,\cdots,x)}\left(\frac{(\alpha-\beta)t}{\alpha^2}\right). \tag{48}$$

Replacing $x_1, x_2, \cdots, x_m$ by $\frac{x_1}{\alpha^n}, \frac{x_2}{\alpha^n}, \cdots, \frac{x_m}{\alpha^n}$, respectively, in (38), we get

$$\begin{aligned}&\mu_{\frac{1}{\alpha^n}\left[f\left(\sum_{i=1}^m \alpha_i\left(\frac{x_i}{\alpha^n}\right)\right)-\left[\prod_{i=1}^m f\left(\frac{x_i}{\alpha^n}\right)\right)\right].\left[\sum_{i=1}^m\left[\alpha_i\left(\prod_{j=1,\ j\neq i}^m f\left(\frac{x_j}{\alpha^n}\right)\right)\right]\right]^{-1}\right]}(t) \\ &\geq \mu'_{\phi\left(\frac{x_1}{\alpha^n},\frac{x_2}{\alpha^n},\cdots,\frac{x_m}{\alpha^n}\right)}(\alpha^n t)\end{aligned} \tag{49}$$

for all $x_1, x_2, \cdots, x_m \in X$ and $t > 0$. Since

$$\lim_{n\to\infty}\mu'_{\phi\left(\frac{x_1}{\alpha^n},\frac{x_2}{\alpha^n},\cdots,\frac{x_m}{\alpha^n}\right)}(\alpha^n t) = 1$$

we conclude that C satisfies (2).

To prove the uniqueness of the mapping C, assume that there exist another mapping $D : X \to Y$ which satisfies (39). Then we have

$$\begin{aligned}\mu_{C(x)-D(x)}(t) &= \lim_{n\to\infty}\mu_{\frac{1}{\alpha^n}C\left(\frac{x}{\alpha^n}\right)-\frac{1}{\alpha^n}D\left(\frac{x}{\alpha^n}\right)}(t) \\ &\geq \lim_{n\to\infty}\min\left\{\mu_{\frac{1}{\alpha^n}C\left(\frac{x}{\alpha^n}\right)-\frac{1}{\alpha^n}f\left(\frac{x}{\alpha^n}\right)}\left(\frac{t}{2}\right), \mu_{\frac{1}{\alpha^n}f\left(\frac{x}{\alpha^n}\right)-\frac{1}{\alpha^n}D\left(\frac{x}{\alpha^n}\right)}\left(\frac{t}{2}\right)\right\} \\ &\geq \lim_{n\to\infty}\mu'_{\phi\left(\frac{x}{\alpha^n},\frac{x}{\alpha^n},\cdots,\frac{x}{\alpha^n}\right)}\left(\frac{\alpha^n(\alpha-\beta)}{2\alpha^2}\right) \\ &\geq \lim_{n\to\infty}\mu'_{\phi(x,x,\cdots,x)}\left(\frac{\alpha^n(\alpha-\beta)t}{2\alpha^2\beta^n}\right).\end{aligned} \tag{50}$$

Since $\lim_{n\to\infty}\frac{\alpha^n(\alpha-\beta)}{2\alpha^2\beta^n} = \infty$ we get

$$\lim_{n\to\infty} \mu'_{\phi(x,x,\cdots,x)}\left(\frac{\alpha^n(\alpha-\beta)t}{2\alpha^2\beta^n}\right) = 1.$$

Therefore, it follows that $\mu_{C(x)-D(x)}(t) = 1$ for all $t > 0$ and so $C(x) = D(x)$. This completes the proof.

Corollary 5 *Let X be a real linear space, $(Z, \mu', \min)$ be an RN-space and $(Y, \mu, \min)$ be a complete RN-space. Let p be a real number with $0 < p < 1$ and $z_0 \in Z$. If $f : X \to Y$ be a mapping such that*

$$\mu_{f(\sum_{i=1}^m \alpha_i x_i)-\left[\prod_{i=1}^m f(x_i)\right].\left[\sum_{i=1}^m\left[\alpha_i\left(\prod_{j=1,\ j\neq i}^m f(x_j)\right)\right]\right]^{-1}}(t) \geq \mu'_{(\sum_{i=1}^m \|x_i\|^p)z_0}(t) \tag{51}$$

for all $x_1, x_2, \cdots, x_m \in X$ and $t > 0$. Then the limit

$$C(x) = \lim_{n\to\infty} \frac{1}{\alpha^n} f\left(\frac{x}{\alpha^n}\right) \tag{52}$$

exists for all $x \in X$ and defines a unique mapping $C : X \to Y$ such

$$\mu_{f(x)-C(x)}(t) \geq \mu'_{\|x\|^p z_0}\left(\frac{(\alpha-\alpha^p)t}{m\alpha^2}\right) \tag{53}$$

for all $x \in X$ and $t > 0$.

Proof Let $\beta = \alpha^p$ and $\phi : X^m \to Z$ be a mapping defined by $\phi(x_1, x_2, \cdots, x_m) = (\sum_{i=1}^m \|x_i\|^p)z_0$. Applying Theorem 6, we get desired result.

Corollary 6 *Let X be a real linear space, $(Z, \mu', \min)$ be an RN-space and $(Y, \mu, \min)$ be a complete RN-space. Let $z_0 \in Z$ and $f : X \to Y$ be a mapping such that*

$$\mu_{f(\sum_{i=1}^m \alpha_i x_i)-\left[\prod_{i=1}^m f(x_i)\right].\left[\sum_{i=1}^m\left[\alpha_i\left(\prod_{j=1,\ j\neq i}^m f(x_j)\right)\right]\right]^{-1}}(t) \geq \mu'_{\delta z_0}(t) \tag{54}$$

for all $x_1, x_2, \cdots, x_m \in X$ and $t > 0$. Then the limit

$$C(x) = \lim_{n\to\infty} \frac{1}{\alpha^n} f\left(\frac{x}{\alpha^n}\right) \tag{55}$$

exists for all $x \in X$ and defines a unique mapping $C : X \to Y$ such

$$\mu_{f(x)-C(x)}(t) \geq \mu_{\delta z_0}\left(\frac{t}{2\alpha}\right) \tag{56}$$

for all $x \in X$ and $t > 0$.

Proof Let $\beta = \frac{\alpha}{2}$ and $\phi : X^m \to Z$ be a mapping defined by $\phi(x_1, x_2, \cdots, x_m) = \delta z_0$. Applying Theorem 6, we get desired result.

Theorem 7 *Let X be a real linear space, $(Z, \mu', \min)$ be an RN-space and $\phi : X^m \to Z$ be a function such that there exists $0 < \beta < \frac{1}{\alpha}$ such that*

$$\mu'_{\phi(\alpha x_1, \alpha x_2, \cdots, \alpha x_m)}(t) \geq \mu'_{\beta\phi(x_1, x_2, \cdots, x_m)}(t) \tag{57}$$

for all $x_1, x_2, \cdots, x_m \in X$ and $t > 0$ and

$$\lim_{n\to\infty} \mu'_{\phi(\alpha^n x_1, \alpha^n x_2, \cdots, \alpha^n x_m)}\left(\frac{t}{\alpha^n}\right) = 1$$

for all $x_1, x_2, \cdots, x_m \in X$ and $t > 0$. Let $(Y, \mu, \min)$ be a complete RN-space. If $f : X \to Y$ be a mapping such that

$$\mu_{f(\sum_{i=1}^m \alpha_i x_i) - \left[\prod_{i=1}^m f(x_i)\right].\left[\sum_{i=1}^m \left[\alpha_i \left(\prod_{j=1,\ j\neq i}^m f(x_j)\right)\right]\right]^{-1}}(t) \geq \mu'_{\phi(x_1, x_2, \cdots, x_m)}(t) \tag{58}$$

for all $x_1, x_2, \cdots, x_m \in X$ and $t > 0$. Then the limit

$$C(x) = \lim_{n\to\infty} \alpha^n f(\alpha^n x)$$

exists for all $x \in X$ and defines a unique mapping $C : X \to Y$ such that

$$\mu_{f(x)-C(x)}(t) \geq \mu'_{\phi(x,x,\cdots,x)}\left(\frac{(1-\alpha\beta)t}{\alpha}\right). \tag{59}$$

for all $x \in X$ and $t > 0$.

Proof By (40), we find that

$$\mu_{\alpha f(\alpha x) - f(x)}(t) \geq \mu'_{\phi(x,x,\cdots,x)}\left(\frac{t}{\alpha}\right) \tag{60}$$

Replacing x by $\alpha^n x$ in (60) and using (57), we obtain

$$\mu_{\alpha^{n+1} f(\alpha^{n+1} x) - \alpha^n f(\alpha^n x)}(t) \geq \mu'_{\phi(x,x,\cdots,x)}\left(\frac{t}{\alpha(\alpha\beta)^n}\right). \tag{61}$$

The rest of the proof is similar to the proof of the Theorem 6.

Corollary 7 *Let X be a real linear space, $(Z, \mu', \min)$ be an RN-space and $(Y, \mu, \min)$ be a complete RN-space. Let $p_i \in \mathbb{R}^+$ with $p = \sum_{i=1}^m p_i > 1$ and $z_0 \in Z$. If $f : X \to Y$ be a mapping such that*

$$\mu_{f(\sum_{i=1}^m \alpha_i x_i)-\left[\prod_{i=1}^m f(x_i)\right].\left[\sum_{i=1}^m\left[\alpha_i\left(\prod_{j=1,\ j\neq i}^m f(x_j)\right)\right]\right]^{-1}}(t) \geq \mu'_{(\prod_{i=1}^m \|x_i\|^{p_i})z_0}(t) \tag{62}$$

for all $x_1, x_2, \cdots, x_m \in X$ and $t > 0$. Then the limit

$$C(x) = \lim_{n\to\infty} \alpha^n f(\alpha^n x) \tag{63}$$

exists for all $x \in X$ and defines a unique mapping $C : X \to Y$ such

$$\mu_{f(x)-C(x)}(t) \geq \mu'_{\|x\|^p z_0}\left(\frac{(\alpha^p - \alpha)t}{\alpha^{p+1}}\right) \tag{64}$$

for all $x \in X$ and $t > 0$.

Proof Let $\beta = \alpha^{-p}$ and $\phi : X^m \to Z$ be a mapping defined by $\phi(x_1, x_2, \cdots, x_m) = (\prod_{i=1}^m \|x_i\|^{p_i})z_0$. Applying Theorem 7, we get desired result.

3.2 Fixed Point Method

Theorem 8 *Let X be a linear space, (Y, μ, T_M) be a complete RN-space and Φ be a mapping from X^m to D^+($\Phi(x_1, x_2, \cdots, x_m)$ is denoted by $\Phi_{x_1,x_2,\cdots,x_m}$) such that there exists $0 < \beta < \frac{1}{\alpha}$ such that*

$$\Phi_{\frac{x_1}{\alpha},\frac{x_2}{\alpha},\cdots,\frac{x_m}{\alpha}}(t) \leq \Phi_{x_1,x_2,\cdots,x_m}(\beta t) \tag{65}$$

for all $x_1, x_2, \cdots, x_m \in X$ and $t > 0$. Let $f : X \to Y$ be a mapping satisfying

$$\mu_{f(\sum_{i=1}^m \alpha_i x_i)-\left[\prod_{i=1}^m f(x_i)\right].\left[\sum_{i=1}^m\left[\alpha_i\left(\prod_{j=1,\ j\neq i}^m f(x_j)\right)\right]\right]^{-1}}(t) \geq \Phi_{x_1,x_2,\cdots,x_m}(t) \tag{66}$$

for all $x_1, x_2, \cdots, x_m \in X$ and $t > 0$. Then it follows that, for all $x \in X$,

$$A(x) := \lim_{n\to\infty} \alpha^n f(\alpha^n x)$$

exists and $A : X \to Y$ is a unique mapping such that

$$\mu_{f(x)-A(x)}(t) \geq \Phi_{x,x,\cdots,x}\left(\frac{(1-\alpha\beta)t}{\alpha}\right) \tag{67}$$

for all $x \in X$ and $t > 0$.

Proof Putting $x_1 = x_2 = \cdots = x_m = x$ in (66), we obtain

$$\mu_{\alpha f(\alpha x) - f(x)}(t) \geq \Phi_{x,x,\cdots,x}\left(\frac{t}{\alpha}\right) \tag{68}$$

for all $x \in X$ and $t > 0$. Consider the set

$$S := \{g : X \to Y\} \tag{69}$$

and the generalized metric d in S defined by

$$d(f, g) = \inf\{u \in \mathbb{R}^+ : \mu_{g(x)-h(x)}(ut) \geq \Phi_{x,x,\cdots,x}(t),\ \forall x \in X,\ t > 0\}, \tag{70}$$

where $\inf \emptyset = +\infty$. It is easy to show that (S, d) is complete (see [22], Lemma 2.1).

Now, we consider a linear mapping $J : S \to S$ such that

$$Jh(x) := \alpha h(\alpha x) \tag{71}$$

for all $x \in X$.

First, we prove that J is a strictly contractive mapping with the Lipschitz constant $\alpha\beta$. In fact, let $g, h \in S$ be such that $d(g, h) < \varepsilon$. Then we have

$$\mu_{g(x)-h(x)}(\varepsilon t) \geq \Phi_{x,x,\cdots,x}(t) \tag{72}$$

for all $x \in X$ and $t > 0$ and so

$$\begin{aligned} \mu_{Jg(x)-Jh(x)}(\alpha\beta\varepsilon t) &= \mu_{\alpha g(\alpha x)-\alpha h(\alpha x)}(\alpha\beta\varepsilon t) \\ &= \mu_{g(\alpha x)-h(\alpha x)}(\beta\varepsilon t) \\ &\geq \Phi_{\alpha x,\alpha x,\cdots,\alpha x}(\beta t) \\ &\geq \Phi_{x,x,\cdots,x}(t) \end{aligned} \tag{73}$$

for all $x \in X$ and $t > 0$. Thus $d(g, h) < \varepsilon$ implies that $d(Jg, Jh) < \alpha\beta\varepsilon$. This means that

$$d(Jg, Jh) \leq \alpha\beta d(g, h) \tag{74}$$

for all $g, h \in S$. It follows from (68) that

$$d(f, Jf) \leq \alpha. \tag{75}$$

By Theorem (3), there exists a mapping $A : X \to Y$ satisfying the following:

(1) A is a fixed point of J, that is,

$$A(\alpha x) = \frac{1}{\alpha} A(x) \tag{76}$$

for all $x \in X$. The mapping A is a unique fixed point of J in the set

$$\Omega = \{h \in S : d(g, h) < \infty\}. \tag{77}$$

This implies that A is a unique mapping satisfying (76) such that there exists $u \in (0, \infty)$ satisfying

$$\mu_{f(x)-A(x)}(ut) \geq \Phi_{x,x,\cdots,x}(t) \tag{78}$$

for all $x \in X$ and $t > 0$.

(2) $d(J^n f, A) \to 0$ as $n \to \infty$. This implies the equality

$$\lim_{n\to\infty} \alpha^n f(\alpha^n x) = A(x) \tag{79}$$

for all $x \in X$.

(3) $d(f, A) \leq \frac{d(f,Jf)}{1-\alpha\beta}$ with $f \in \Omega$, which implies the inequality

$$d(f, A) \leq \frac{\alpha}{1-\alpha\beta} \tag{80}$$

and so

$$\mu_{f(x)-A(x)}\left(\frac{\alpha t}{1-\alpha\beta}\right) \geq \Phi_{x,x,\cdots,x}(t) \tag{81}$$

for all $x \in X$ and $t > 0$. This implies that the inequality (67) holds. Replacing $x_1, x_2, \cdots, x_m$ by $\alpha^n x_1, \alpha^n x_2, \cdots, \alpha^n x_m$, respectively in (66), we obtain

$$\begin{aligned}&\mu_{\alpha^n\left[f(\sum_{i=1}^m \alpha_i(\alpha^n x_i)) - \left[\prod_{i=1}^m f(\alpha^n x_i)\right] . \left[\sum_{i=1}^m \left[\alpha_i\left(\prod_{j=1,\ j\neq i}^m f(\alpha^n x_j)\right)\right]\right]^{-1}\right]}(t)\\ &\geq \Phi_{\alpha^n x_1, \alpha^n x_2, \cdots, \alpha^n x_m}\left(\frac{t}{\alpha^n}\right)\end{aligned} \tag{82}$$

for all $x_1, x_2, \cdots, x_m \in X$, $t > 0$ and $n \geq 1$ and so, from (65), it follows that

$$\Phi_{\alpha^n x_1, \alpha^n x_2, \cdots, \alpha^n x_m}\left(\frac{t}{\alpha^n}\right) \geq \Phi_{x_1, x_2, \cdots, x_m}\left(\frac{t}{(\alpha\beta)^n}\right) \tag{83}$$

Since

$$\lim_{n\to\infty} \Phi_{x_1, x_2, \cdots, x_m}\left(\frac{t}{(\alpha\beta)^n}\right) = 1$$

for all $x_1, x_2, \cdots, x_m \in X$ and $t > 0$, then we have

$$\mu_{A(\sum_{i=1}^m \alpha_i x_i) - \left[\prod_{i=1}^m A(x_i)\right] \cdot \left[\sum_{i=1}^m \left[\alpha_i \left(\prod_{j=1,\ j\neq i}^m A(x_j)\right)\right]\right]^{-1}}(t) = 1$$

for all $x_1, x_2, \cdots, x_m \in X$ and $t > 0$. Thus the mapping $A : X \to Y$ satisfy (2). This completes the proof.

Corollary 8 *Let $\theta \geq 0$ and p be a real number with $p > 1$. Let $f : X \to Y$ be a mapping satisfying*

$$\mu_{f(\sum_{i=1}^m \alpha_i x_i) - \left[\prod_{i=1}^m f(x_i)\right] \cdot \left[\sum_{i=1}^m \left[\alpha_i \left(\prod_{j=1,\ j\neq i}^m f(x_j)\right)\right]\right]^{-1}}(t) \geq \frac{t}{t + \theta\left(\sum_{i=1}^m \|x_i\|^p\right)} \tag{84}$$

for all $x_1, x_2, \cdots, x_n \in X$ and $t > 0$. Then the limit

$$A(x) = \lim_{n\to\infty} \alpha^n f(\alpha^n x) \tag{85}$$

exists for all $x \in X$ and $A : X \to Y$ is a unique mapping such that

$$\mu_{f(x)-A(x)}(t) \geq \frac{(\alpha^p - \alpha)t}{(\alpha^p - \alpha)t + m\alpha^{p+1}\theta\|x\|^p} \tag{86}$$

for all $x \in X$ and $t > 0$.

Proof The proof follows from Theorem 8 if we take

$$\Phi_{x_1,x_2,\cdots,x_m}(t) = \frac{t}{t + \theta\left(\sum_{i=1}^m \|x_i\|^p\right)} \tag{87}$$

for all $x_1, x_2, \cdots, x_m \in X$ and $t > 0$. In fact, if we choose $\beta = \frac{1}{\alpha^p}$, then we get the desired result.

Theorem 9 *Let X be a linear space, (Y, μ, T_M) be a complete RN-space and Φ be a mapping from X^m to D^+ ($\Phi(x_1, x_2, \cdots, x_m)$ is denoted by $\Phi_{x_1,x_2,\cdots,x_m}$) such that for some $0 < \beta < \alpha$*

$$\Phi_{\alpha x_1,\alpha x_2,\cdots,\alpha x_m}(t) \leq \Phi_{x_1,x_2,\cdots,x_m}(\beta t) \tag{88}$$

for all $x_1, x_2, \cdots, x_m \in X$ and $t > 0$. Let $f : X \to Y$ be a mapping satisfying

$$\mu_{f(\sum_{i=1}^m \alpha_i x_i) - \left[\prod_{i=1}^m f(x_i)\right] \cdot \left[\sum_{i=1}^m \left[\alpha_i \left(\prod_{j=1,\ j\neq i}^m f(x_j)\right)\right]\right]^{-1}}(t) \geq \Phi_{x_1,x_2,\cdots,x_m}(t) \tag{89}$$

for all $x, y \in X$ and $t > 0$. Then the limit

$$A(x) := \lim_{n\to\infty} \frac{1}{\alpha^n} f\Big(\frac{x}{\alpha^n}\Big) \tag{90}$$

exists for all $x \in X$ and $A : X \to Y$ is a unique mapping such that

$$\mu_{f(x)-A(x)}(t) \geq \varPhi_{x,x,\cdots,x}\Big(\frac{(\alpha-\beta)t}{\alpha\beta}\Big). \tag{91}$$

for all $x \in X$ and $t > 0$.

Proof Substituting $x_1 = x_2 = \cdots, x_m = x$ in (89), we obtain

$$\mu_{f(x)-\frac{1}{\alpha}f(\frac{x}{\alpha})}(t) \geq \varPhi_{\frac{x}{\alpha},\frac{x}{\alpha},\cdots,\frac{x}{\alpha}}(t) \geq \varPhi_{x,x,\cdots,x}\left(\frac{t}{\beta}\right) \tag{92}$$

for all $x \in X$. Let (S, d) be the generalized metric space defined in the proof of the Theorem 8. Consider a linear mapping $J : S \to S$ such that

$$Jh(x) := \frac{1}{\alpha} h\Big(\frac{x}{\alpha}\Big) \tag{93}$$

for all $x \in X$. It follows from (92) that

$$d(f, Jf) \leq \beta. \tag{94}$$

By Theorem 3, there exists a mapping $A : X \to Y$ satisfying the following:

(1) A is a fixed point of J, that is,

$$\alpha A(x) = A\Big(\frac{x}{\alpha}\Big) \tag{95}$$

for all $x \in X$. The mapping A is a unique fixed point of J in the set

$$\varOmega = \{h \in S : d(g, h) < \infty\}.$$

This implies that A is a unique mapping satisfying (95) such that there exists $u \in (0, \infty)$ satisfying

$$\mu_{f(x)-A(x)}(ut) \geq \varPhi_{x,x,\cdots,x}(t) \tag{96}$$

for all $x \in X$.

(2) $d(J^n f, A) \to 0$ as $n \to \infty$. This implies the equality

$$\lim_{n\to\infty} \frac{1}{\alpha^n} f\Big(\frac{x}{\alpha^n}\Big) = A(x) \tag{97}$$

for all $x \in X$.

(3) $d(f, A) \leq \frac{d(f, Jf)}{1-L}$ with $f \in \Omega$, which implies the inequality

$$d(f, A) \leq \frac{\alpha\beta}{\alpha - \beta}.$$

So and so

$$\mu_{f(x)-A(x)}\left(\frac{\alpha\beta t}{\alpha - \beta}\right) \geq \Phi_{x,x,\cdots,x}(t) \tag{98}$$

This implies that the inequality (91) holds. The rest of the proof is similar to the proof of Theorem 8.

Corollary 9 *Let $\theta \geq 0$ and p be a real number with $0 < p < 1$. Let $f : X \to Y$ be a mapping satisfying*

$$\mu_{f(\sum_{i=1}^{m} \alpha_i x_i) - \left[\prod_{i=1}^{m} f(x_i)\right] . \left[\sum_{i=1}^{m}\left[\alpha_i\left(\prod_{j=1,\ j\neq i}^{m} f(x_j)\right)\right]\right]^{-1}}(t) \geq \frac{t}{t + \theta\left(\sum_{i=1}^{m} \|x_i\|^p\right)} \tag{99}$$

for all $x_1, x_2, \cdots, x_m \in X$ and $t > 0$. Then the limit

$$A(x) = \lim_{n\to\infty} \frac{1}{\alpha^n} f\left(\frac{x}{\alpha^n}\right) \tag{100}$$

exists for all $x \in X$ and $A : X \to Y$ is a unique mapping such that

$$\mu_{f(x)-A(x)}(t) \geq \frac{(\alpha - \alpha^p)t}{(\alpha - \alpha^p)t + m.\alpha^{p+1}\theta\|x\|^p} \tag{101}$$

for all $x \in X$ and $t > 0$.

Proof The proof follows from Theorem 9 if we take

$$\Phi_{x_1,x_2,\cdots,x_m}(t) = \frac{t}{t + \theta\left(\sum_{i=1}^{m} \|x_i\|^p\right)} \tag{102}$$

for all $x_1, x_2, \cdots, x_m \in X$ and $t > 0$. In fact, if we choose $\beta = \alpha^p$, then we get the desired result.

References

1. L.M. Arriola, W.A. Beyer, Stability of the Cauchy functional equation over p-adic fields. Real Anal. Exchange **31**, 125–132 (2005/06)
2. H. Azadi Kenary, Stability of a Pexiderial functional equation in random normed spaces. Rend. Circ. Mat. Palermo **60**, 59–68 (2011). https://doi.org/10.1007/s12215-011-0027-5

3. H. Azadi Kenary, C. Park, Direct and fixed point methods approach to the generalized Hyers–Ulam stability for a functional equation having monomials as solutions. Iran. J. Sci. Technol. Trans. A **A4**, 301–307 (2011)
4. H. Azadi Kenary, J.R. Lee, C. Park, Non-Archimedean stability of an AQQ functional equation. J. Comput. Anal. Appl. **14**(2), 211–227 (2012)
5. H. Azadi Kenary, H. Rezaei, S. Talebzadeh, S.J. Lee, Stabilities of cubic mappings in various normed spaces: direct and fixed point methods. J. Appl. Math. **2012**, Article ID 546819, 28 pp. (2012)
6. P.W. Cholewa, Remarks on the stability of functional equations. Aequationes Math. **27**, 76–86 (1984)
7. S. Czerwik, *Functional Equations and Inequalities in Several Variables* (World Scientific, River Edge, 2002)
8. D. Deses, *On the representation of non-Archimedean objects*. Topol. Appl. **153**, 774–785 (2005)
9. M. Eshaghi Gordji, M. Bavand Savadkouhi, Stability of mixed type cubic and quartic functional equations in random normed spaces. J. Inequal. Appl. **2009**, Article ID 527462, 9 pp. (2009)
10. M. Eshaghi Gordji, M.B. Savadkouhi, Stability of cubic and quartic functional equations in non-Archimedean spaces. Acta Applicandae Math. **110**, 1321–1329 (2010)
11. M. Eshaghi Gordji, M. Bavand Savadkouhi, C. Park, Quadratic-quartic functional equations in RN-spaces. J. Inequal. Appl. **2009**, Article ID 868423, 14 pp. (2009)
12. M. Eshaghi Gordji, S. Zolfaghari, J.M. Rassias, M.B. Savadkouhi, Solution and stability of a mixed type cubic and quartic functional equation in quasi-Banach spaces. Abstr. Appl. Anal. **2009**, Article ID 417473, 14 pp. (2009)
13. P. Găvruta, A generalization of the Hyers-Ulam-Rassias stability of approximately additive mappings. J. Math. Anal. Appl. **184**, 431–436 (1994)
14. K. Hensel, Ubereine news Begrundung der Theorie der algebraischen Zahlen. Jahresber. Deutsch. Math. Verein **6**, 83–88 (1897)
15. D.H. Hyers, On the stability of the linear functional equation. Proc. Natl. Acad. Sci. U. S. A. **27**, 222–224 (1941)
16. S.M. Jung, Hyers-Ulam-Rassias stability of Jensen's equation and its application. Proc. Am. Math. Soc. **126**, 3137–3143 (1998)
17. S.-M. Jung, M.Th. Rassias, A linear functional equation of third order associated to the Fibonacci numbers. Abstr. Appl. Anal. **2014**, Article ID 137468 (2014)
18. S.-M. Jung, D. Popa, M.Th. Rassias, On the stability of the linear functional equation in a single variable on complete metric groups. J. Glob. Optim. **59**, 165–171 (2014)
19. S.-M. Jung, M.Th. Rassias, C. Mortici, On a functional equation of trigonometric type. Appl. Math. Comput. **252**, 294–303 (2015)
20. A.K. Katsaras, A. Beoyiannis, Tensor products of non-Archimedean weighted spaces of continuous functions. Georgian Math. J. **6**, 33–44 (1999)
21. A. Khrennikov, *Non-Archimedean Analysis: Quantum Paradoxes, Dynamical Systems and Biological Models*, vol. 427. Mathematics and Its Applications (Kluwer Academic Publishers, Dordrecht, 1997)
22. D. Mihet, V. Radu, On the stability of the additive Cauchy functional equation in random normed spaces. J. Math. Anal. Appl. **343**, 567–572 (2008)
23. A.K. Mirmostafaee, Approximately additive mappings in non-Archimedean normed spaces. Bull. Korean Math. Soc. **46**, 387–400 (2009)
24. C. Mortici, M.Th. Rassias, S.-M. Jung, On the stability of a functional equation associated with the Fibonacci numbers. Abstr. Appl. Anal. **2014**, Article ID 546046, 6 pp. (2014)
25. M.S. Moslehian, Th.M. Rassias, Stability of functional equations in non-Archimedean spaces. Appl. Anal. Discrete Math. **1**, 325–334 (2007)
26. P.J. Nyikos, On some non-Archimedean spaces of Alexandrof and Urysohn. Topol. Appl. **91**, 1–23 (1999)

27. C. Park, Generalized Hyers-Ulam-Rassias stability of n-sesquilinear-quadratic mappings on Banach modules over C^*-algebras. J. Comput. Appl. Math. **180**, 279–291 (2005)
28. C. Park, Fixed points and Hyers-Ulam-Rassias stability of Cauchy-Jensen functional equations in Banach algebras. Fixed Point Theory Appl. **2007**, Article ID 50175 (2007)
29. C. Park, Generalized Hyers-Ulam-Rassias stability of quadratic functional equations: a fixed point approach. Fixed Point Theory Appl. **2008**, Article ID 493751 (2008)
30. C. Park, Fuzzy stability of a functional equation associated with inner product spaces. Fuzzy Sets Syst. **160**, 1632–1642 (2009)
31. Th.M. Rassias, On the stability of the linear mapping in Banach spaces. Proc. Am. Math. Soc. **72**, 297–300 (1978)
32. Th.M. Rassias, Problem 16; 2. Report of the 27th International Symposium on Functional Equations. Aequations Math. **39**, 292–293 (1990)
33. Th.M. Rassias, On the stability of the quadratic functional equation and its applications. Studia Univ. Babes-Bolyai. **XLIII** 89–124 (1998)
34. Th.M. Rassias, The problem of S.M. Ulam for approximately multiplicative mappings. J. Math. Anal. Appl. **246**, 352–378 (2000)
35. Th.M. Rassias, On the stability of functional equations in Banach spaces. J. Math. Anal. Appl. **251**, 264–284 (2000)
36. Th.M. Rassias, *Functional Equations, Inequalities and Applications* (Kluwer Academic Publishers Co., Dordrecht, 2003)
37. Th.M. Rassias, P. Semrl, On the behaviour of mappings which do not satisfy Hyers-Ulam stability. Proc. Am. Math. Soc. **114** 989–993 (1992)
38. Th.M. Rassias, P. Semrl, On the Hyers-Ulam stability of linear mappings. J. Math. Anal. Appl. **173**, 325–338 (1993)
39. K. Ravi, B.V.S. Kumar, Ulam stability of generalized reciprocal functional equation in several variables. Int. J. Appl. Math. Stat. **19**(D10), 1–19 (2010)
40. R. Saadati, C. Park, Non-Archimedean $\mathscr{L}$-fuzzy normed spaces and stability of functional equations. Comput. Math. Appl. **60**(8), 2488–2496 (2010)
41. R. Saadati, M. Vaezpour, Y.J. Cho, A note to paper "On the stability of cubic mappings and quartic mappings in random normed spaces". J. Inequal. Appl. **2009**, Article ID 214530. https://doi.org/10.1155/2009/214530
42. R. Saadati, M.M. Zohdi, S.M. Vaezpour, Nonlinear L-random stability of an ACQ functional equation. J. Inequal. Appl. **2011**, Article ID 194394, 23 pp. https://doi.org/10.1155/2011/194394
43. B. Schewizer, A. Sklar, *Probabilistic Metric Spaces*. North-Holland Series in Probability and Applied Mathematics (North-Holland, New York, 1983)
44. F. Skof, Local properties and approximation of operators. Rend. Sem. Mat. Fis. Milano **53**, 113–129 (1983)
45. S.M. Ulam, *Problems in Modern Mathematics*. Science Editions (Wiley, New York, 1964)

On the HUR-Stability of Quadratic Functional Equations in Fuzzy Banach Spaces

Hassan Azadi Kenary and Themistocles M. Rassias

1 Introduction and Preliminaries

Let X be a real vector space. A function $N : X \times \mathbb{R} \to [0, 1]$ is called a fuzzy norm on X if for all $x, y \in X$ and all $s, t \in \mathbb{R}$,

(*N1*) $N(x, t) = 0$ for $t \leq 0$;
(*N2*) $x = 0$ if and only if $N(x, t) = 1$ for all $t > 0$;
(*N3*) $N(cx, t) = N\left(x, \frac{t}{|c|}\right)$ if $c \neq 0$;
(*N4*) $N(x + y, c + t) \geq min\{N(x, s), N(y, t)\}$;
(*N5*) $N(x, .)$ is a non-decreasing function of $\mathbb{R}$ and $\lim_{t\to\infty} N(x, t) = 1$;
(*N6*) for $x \neq 0$, $N(x, .)$ is continuous on $\mathbb{R}$.

Example 1 Let $(X, \|.\|)$ be a normed linear space and $\alpha, \beta > 0$. Then

$$N(x, t) = \begin{cases} \frac{\alpha t}{\alpha t+\beta\|x\|} & t > 0, x \in X \\ 0 & t \leq 0, x \in X \end{cases}$$

is a fuzzy norm on X.

Definition 1 Let (X, N) be a fuzzy normed vector space. A sequence $\{x_n\}$ in X is said to be convergent or converge if there exists an $x \in X$ such that $\lim_{t\to\infty} N(x_n -$

H. Azadi Kenary (✉)
Department of Mathematics, College of Sciences, Yasouj University, Yasouj, Iran
e-mail: azadi@yu.ac.ir

Th. M. Rassias
Department of Mathematics, National Technical University of Athens, Athens, Greece
e-mail: trassias@math.ntua.gr

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_17

$x, t) = 1$ for all $t > 0$. In this case, x is called the limit of the sequence $\{x_n\}$ in X and we denote it by $N - \lim_{t\to\infty} x_n = x$.

Definition 2 Let (X, N) be a fuzzy normed vector space. A sequence $\{x_n\}$ in X is called Cauchy if for each $\varepsilon > 0$ and each $t > 0$ there exists an $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ and all $p > 0$, we have $N(x_{n+p} - x_n, t) > 1 - \varepsilon$.

It is well known that every convergent sequence in a fuzzy normed vector space is Cauchy. If each Cauchy sequence is convergent, then the fuzzy norm is said to be complete and the fuzzy normed vector space is called a fuzzy Banach space.

We say that a mapping $f : X \to Y$ between fuzzy normed vector spaces X and Y is continuous at a point $x \in X$ if for each sequence $\{x_n\}$ converging to $x_0 \in X$, then the sequence $\{f(x_n)\}$ converges to $f(x_0)$. If $f : X \to Y$ is continuous at each $x \in X$, then $f : X \to Y$ is said to be continuous on X (see [2]).

Definition 3 Let X be a set. A function $d : X \times X \to [0, \infty]$ is called a generalized metric on X if d satisfies the following conditions:

(1) $d(x, y) = 0$ if and only if $x = y$ for all $x, y \in X$;
(2) $d(x, y) = d(y, x)$ for all $x, y \in X$;
(3) $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in X$.

Theorem 1 *Let (X,d) be a complete generalized metric space and $J : X \to X$ be a strictly contractive mapping with Lipschitz constant $L < 1$. Then, for all $x \in X$, either*

$$d(J^n x, J^{n+1} x) = \infty$$

for all nonnegative integers n or there exists a positive integer n_0 such that

(1) $d(J^n x, J^{n+1} x) < \infty$ *for all $n_0 \geq n_0$;*
(2) *the sequence $\{J^n x\}$ converges to a fixed point y^* of J;*
(3) *y^* is the unique fixed point of J in the set $Y = \{y \in X : d(J^{n_0} x, y) < \infty\}$;*
(4) $d(y, y^*) \leq \frac{1}{1-L} d(y, Jy)$ *for all $y \in Y$.*

The stability problem of functional equations was originated from a question of Ulam [37] concerning the stability of group homomorphisms. Hyers [13] gave a first affirmative partial answer to the question of Ulam for Banach spaces. Hyers' Theorem was generalized by Th. M. Rassias [29] for linear mappings by considering an unbounded Cauchy difference.

Theorem 2 (Th.M. Rassias) *Let $f : E \to E'$ be a mapping from a normed vector space E into a Banach space E' subject to the inequality $\|f(x + y) - f(x) - f(y)\| \leq \varepsilon(\|x\|^p + \|y\|^p)$ for all $x, y \in E$, where ε and p are constants with $\varepsilon > 0$ and $0 \leq p < 1$. Then the limit $L(x) = \lim_{n\to\infty} \frac{f(2^n x)}{2^n}$ exists for all $x \in E$ and $L : E \to E'$ is the unique linear mapping which satisfies $\|f(x) - L(x)\| \leq \frac{2\varepsilon}{2-2^p} \|x\|^p$ for all $x \in E$. Also, if for each $x \in E$ the function $f(tx)$ is continuous in $t \in \mathbb{R}$, then L is linear.*

The functional equation $f(x+y)+f(x-y)=2f(x)+2f(y)$ is called a *quadratic functional equation*. In particular, every solution of the quadratic functional equation is said to be a *quadratic mapping*. The Hyers-Ulam stability of the quadratic functional equation was proved by Skof [36] for mappings $f : X \to Y$, where X is a normed space and Y is a Banach space. Cholewa [5] noticed that the theorem of Skof is still true if the relevant domain X is replaced by an Abelian group. Czerwik [6] proved the Hyers-Ulam stability of the quadratic functional equation.

In this paper, we consider the following quadratic functional equations

$$f\left(\sum_{i=1}^{n} a_i x_i\right) + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} f(a_i x_i \pm a_j x_j) = (3n-2) \sum_{i=1}^{n} a_i^2 f(x_i), \tag{1}$$

where $a_1, \cdots, a_n \in \mathbb{Z} - \{0\}$ and some $l \in \{1, 2, \cdots, n-1\}$, $a_l \neq 1$ and $a_n = 1$, where n is a positive integer greater or at lease equal to two, in fuzzy Banach spaces.

The stability problems of several functional equations have been extensively investigated by a number of authors, and there are many interesting results concerning this problem (see [4, 7–10, 12, 14–16, 21–28, 30–35]).

Katsaras [18] defined a fuzzy norm on a vector space to construct a fuzzy vector topological structure on the space. Some mathematicians have defined fuzzy norms on a vector space from various points of view (see [11, 19, 28]).

In particular, Bag and Samanta [1], following Cheng and Mordeson [3], gave an idea of fuzzy norm in such a manner that the corresponding fuzzy metric is of Karmosil and Michalek type [17]. They established a decomposition theorem of a fuzzy norm into a family of crisp norms and investigated some properties of fuzzy normed spaces [2].

2 Fuzzy Stability of Quadratic Functional Equation (1): A Fixed Point Method

In this section, using the fixed point alternative approach we prove the Hyers-Ulam-Rassias stability of functional equation (1) in fuzzy Banach spaces. Throughout this paper, assume that X is a vector space and that (Y, N) is a fuzzy Banach space.

Theorem 3 *Let $\varphi : X^n \to [0, \infty)$ be a function such that there exists an $L < 1$ with*

$$\varphi\left(\frac{x_1}{a_1}, \frac{x_2}{a_2}, \cdots, \frac{x_n}{a_n}\right) \leq \frac{L\varphi(a_1, a_2, \cdots, a_n)}{a_1^2}$$

for all $x_1, x_2, \cdots, x_n \in X$ *and all* $a_1 \neq 1$. *Let* $f : X \to Y$ *with* $f(0) = 0$ *is a mapping satisfying*

$$N\left(f\left(\sum_{i=1}^{n} a_i x_i\right) + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} f(a_i x_i \pm a_j x_j) - (3n-2)\sum_{i=1}^{n} a_i^2 f(x_i), t\right) \geq \frac{t}{t + \varphi(x_1, \cdots, x_n)} \tag{2}$$

for all $x_1, \cdots, x_n \in X$ *and all* $t > 0$. *Then, the limit*

$$Q(x) := N\text{-}\lim_{m\to\infty} a_1^{2m} f\left(\frac{x}{a_1^m}\right)$$

exists for each $x \in X$ *and defines a unique quadratic mapping* $Q : X \to Y$ *such that*

$$N(f(x) - Q(x), t) \geq \frac{(a_1^2(3n-2) - a_1^2(3n-2)L)t}{(a_1^2(3n-2) - a_1^2(3n-2)L)t + L\varphi(x, 0, \cdots, 0)}. \tag{3}$$

for all $x \in X$ *and all* $t > 0$.

Proof Putting $x_1 = x$ and $x_2 = \cdots = x_n = 0$ in (2) and using $f(0) = 0$, we have

$$N\left(f(a_1 x) - a_1^2 f(x), \frac{t}{3n-2}\right) \geq \frac{t}{t + \varphi(x, 0, \cdots, 0)}. \tag{4}$$

Replacing x by $\frac{x}{a_1}$ in (4), we obtain

$$N\left(a_1^2 f\left(\frac{x}{a_1}\right) - f(x), \frac{t}{3n-2}\right) \geq \frac{t}{t + \varphi\left(\frac{x}{a_1}, 0, \cdots, 0\right)} \tag{5}$$

for all $x \in X$ and $t > 0$. Consider the set $S := \{g : X \to Y ;\ g(0) = 0\}$ and the generalized metric d in S defined by

$$d(f, g) = \inf\left\{\mu \in \mathbb{R}^+ : N(g(x) - h(x), \mu t) \geq \frac{t}{t + \varphi(x, 0, \cdots, 0)}, \forall x \in X,\ t > 0\right\},$$

where $\inf \emptyset = +\infty$. It is easy to show that (S, d) is complete (see [20]). Now, we consider a linear mapping $J : S \to S$ such that $Jg(x) := a_1^2 g\left(\frac{x}{a_1}\right)$ for all $x \in X$. Let $g, h \in S$ be such that $d(g, h) = \varepsilon$. Then $N(g(x) - h(x), \varepsilon t) \geq \frac{t}{t + \varphi(x, 0, \cdots, 0)}$ for all $x \in X$ and $t > 0$. Hence

$$N(Jg(x) - Jh(x), L\varepsilon t) = N\left(a_1^2 g\left(\frac{x}{a_1}\right) - a_1^2 h\left(\frac{x}{a_1}\right), L\varepsilon t\right)$$
$$= N\left(g\left(\frac{x}{a_1}\right) - g\left(\frac{x}{a_1}\right), \frac{L\varepsilon t}{a_1^2}\right) \geq \frac{\frac{Lt}{a_1^2}}{\frac{Lt}{a_1^2} + \varphi\left(\frac{x}{a_1}, 0, \cdots, 0\right)}$$
$$\geq \frac{\frac{Lt}{a_1^2}}{\frac{Lt}{a_1^2} + \frac{L\varphi(x,0,\cdots,0)}{a_1^2}} = \frac{t}{t + \varphi(x, 0, \cdots, 0)}$$

for all $x \in X$ and $t > 0$. Thus $d(g, h) = \varepsilon$ implies that $d(Jg, Jh) \leq L\varepsilon$. This means that $d(Jg, Jh) \leq Ld(g, h)$ for all $g, h \in S$. It follows from (5) that

$$N\left(a_1^2 f\left(\frac{x}{a_1}\right) - f(x), \frac{Lt}{a_1^2(3n-2)}\right) \geq \frac{t}{t + \varphi\,(x, 0, \cdots, 0)} \tag{6}$$

for all $x \in X$ and $t > 0$. This implies that $d(f, Jf) \leq \frac{L}{a_1^2(3n-2)}$. By Theorem 2.1, there exists a mapping $Q : X \to Y$ satisfying the following:

(1) Q is a fixed point of J, that is,

$$Q\left(\frac{x}{a_1}\right) = \frac{Q(x)}{a_1^2} \tag{7}$$

for all $x \in X$. The mapping Q is a unique fixed point of J in the set $\Omega = \{h \in S : d(g, h) < \infty\}$. This implies that Q is a unique mapping satisfying (7) such that there exists $\mu \in (0, \infty)$ satisfying $N(f(x) - Q(x), \mu t) \geq \frac{t}{t+\varphi(x,0,\cdots,0)}$ for all $x \in X$ and $t > 0$.

(2) $d(J^m f, Q) \to 0$ as $m \to \infty$. This implies the equality

$$N\text{-}\lim_{m\to\infty} a_1^{2m} f\left(\frac{x}{a_1^m}\right) = Q(x)$$

for all $x \in X$.

(3) $d(f, Q) \leq \frac{d(f,Jf)}{1-L}$ with $f \in \Omega$, which implies the inequality $d(f, Q) \leq \frac{L}{a_1^2(3n-2)-a_1^2(3n-2)L}$. This implies that the inequality (3) holds. Furthermore, since

$$N\left(Q\left(\sum_{i=1}^{n} a_i x_i\right) + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Q(a_i x_i \pm a_j x_j) - (3n-2)\sum_{i=1}^{n} a_i^2 Q(x_i), t\right)$$

$$= N - \lim_{m\to\infty}\left(a_1^{2m} f\left(\sum_{i=1}^{n}\frac{a_i x_i}{a_1^m}\right) + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_1^{2m} f\left(\frac{a_i x_i \pm a_j x_j}{a_1^m}\right)\right.$$

$$\left. -(3n-2)\sum_{i=1}^{n} a_i^2 a_1^{2m} f\left(\frac{x_i}{a_1^m}\right), t\right)$$

$$\geq \lim_{m\to\infty}\frac{\frac{t}{a_1^{2m}}}{\frac{t}{a_1^{2m}} + \varphi\left(\frac{x_1}{a_1^m}, \frac{x_2}{a_1^m}, \cdots, \frac{x_n}{a_1^m}\right)} \geq \lim_{m\to\infty}\frac{\frac{t}{a_1^{2m}}}{\frac{t}{a_1^{2m}} + \frac{L^m \varphi(x_1, x_2, \cdots, x_n)}{a_1^{2m}}} \to 1$$

for all $x_1, x_2, \cdots, x_n \in X$, $t > 0$ and all $m \in \mathbb{N}$. Hence

$$N\left(Q\left(\sum_{i=1}^{n} a_i x_i\right) + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Q(a_i x_i \pm a_j x_j) - (3n-2)\sum_{i=1}^{n} a_i^2 Q(x_i), t\right) = 1$$

for all $x_1, x_2, \cdots, x_n \in X$ and all $t > 0$. Thus the mapping $Q : X \to Y$ is quadratic, as desired. This completes the proof.

Corollary 1 *Let $\theta \geq 0$ and let p be a real number with $p > 2$. Let X be a normed vector space with norm $\|.\|$. Let $f : X \to Y$ with $f(0) = 0$ be a mapping satisfying the following inequality*

$$N\left(f\left(\sum_{i=1}^{n} a_i x_i\right) + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} f(a_i x_i \pm a_j x_j) - (3n-2)\sum_{i=1}^{n} a_i^2 f(x_i), t\right)$$

$$\geq \frac{t}{t + \theta\left(\sum_{i=1}^{n}\|x_i\|^p\right)}$$

for all $x_1, x_2, \cdots, x_n \in X$ and all $t > 0$. Then, the limit $Q(x) := N\text{-}\lim_{m\to\infty} a_1^{2m} f\left(\frac{x}{a_1^m}\right)$ exists for each $x \in X$ and defines a unique quadratic mapping $Q : X \to Y$ such that

$$N(f(x) - Q(x), t) \geq \frac{a_1^2(3n-2)(|a_1|^p - a_1^2)t}{a_1^2(3n-2)(|a_1|^p - a_1^2)t + a_1^2\theta\|x\|^p}$$

for all $x \in X$ and $t > 0$.

Proof The proof follows from Theorem 3.1 by taking $\varphi(x_1, x_2, \cdots, x_n) := \theta\left(\sum_{i=1}^{n}\|x_i\|^p\right)$ for all $x_1, x_2, \cdots, x_n \in X$. Then we can choose $L = |a_1|^{2-p}$ and we get the desired result.

Theorem 4 *Let $\varphi : X^n \to [0, \infty)$ be a function such that there exists an $L < 1$ with $\varphi(x_1, x_2, \cdots, x_n) \leq a_1^2 L\varphi\left(\frac{x_1}{a_1}, \frac{x_2}{a_1}, \cdots, \frac{x_n}{a_1}\right)$ for all $x, y \in X$. Let $f : X \to Y$*

be a mapping with $f(0)=0$ satisfying (2). Then $Q(x) := N\text{-}\lim_{m\to\infty} \frac{f(a_1^m x)}{a_1^{2m}}$ exists for each $x \in X$ and defines a unique quadratic mapping $Q : X \to Y$ such that

$$N(f(x)-Q(x),t) \geq \frac{(a_1^2(3n-2)-a_1^2(3n-2)L)t}{(a_1^2(3n-2)-a_1^2(3n-2)L)t+\varphi(x,0,\cdots,0)} \tag{8}$$

for all $x \in X$ and all $t > 0$.

Proof Let (S,d) be the generalized metric space defined as in the proof of Theorem 2.1. Consider the linear mapping $J : S \to S$ such that $Jg(x) := \frac{g(a_1x)}{a_1^2}$ for all $x \in X$. Let $g,h \in S$ be such that $d(g,h)=\varepsilon$. Then $N(g(x)-h(x),\varepsilon t) \geq \frac{t}{t+\varphi(x,0,\cdots,0)}$ for all $x \in X$ and $t > 0$. Hence

$$\begin{aligned}
N(Jg(x)-Jh(x),L\varepsilon t) &= N\left(\frac{g(a_1x)}{a_1^2}-\frac{h(a_1x)}{a_1^2},L\varepsilon t\right)\\
&= N\Big(g(a_1x)-h(a_1x),a_1^2L\varepsilon t\Big) \geq \frac{a_1^2Lt}{a_1^2Lt+\varphi(a_1x,,0,\cdots,0)}\\
&\geq \frac{a_1^2Lt}{a_1^2Lt+a_1^2L\varphi(x,0,\cdots,0)} = \frac{t}{t+\varphi(x,0,\cdots,0)}
\end{aligned}$$

for all $x \in X$ and $t > 0$. Thus $d(g,h)=\varepsilon$ implies that $d(Jg,Jh) \leq L\varepsilon$. This means that $d(Jg,Jh) \leq Ld(g,h)$ for all $g,h \in S$. It follows from (4) that

$$N\left(f(x)-\frac{f(a_1x)}{a_1^2},\frac{t}{a_1^2(3n-2)}\right) \geq \frac{t}{t+\varphi(x,0,\cdots,0)} \tag{9}$$

for all $x \in X$ and $t > 0$. Therefore

$$d(f,Jf) \leq \frac{1}{a_1^2(3n-2)}.$$

By Theorem 2.1, there exists a mapping $Q : X \to Y$ satisfying the following:

(1) Q is a fixed point of J, that is,

$$2^2Q(x) = Q(2x) \tag{10}$$

for all $x \in X$. The mapping Q is a unique fixed point of J in the set $\Omega = \{h \in S : d(g,h) < \infty\}$. This implies that Q is a unique mapping satisfying (10) such that there exists $\mu \in (0,\infty)$ satisfying $N(f(x)-Q(x),\mu t) \geq \frac{t}{t+\varphi(x,0,\cdots,0)}$ for all $x \in X$ and $t > 0$.

(2) $d(J^m f, Q) \to 0$ as $m \to \infty$. This implies the equality $Q(x) = N\text{-}\lim_{m\to\infty} \frac{f(a_1^m x)}{a_1^{2m}}$ for all $x \in X$.

(3) $d(f, Q) \leq \frac{d(f,Jf)}{1-L}$ with $f \in \Omega$, which implies the inequality $d(f, Q) \leq \frac{1}{a_1^2(3n-2)-a_1^2(3n-2)L}$. This implies that the inequality (8) holds. The rest of the proof is similar to that of the proof of Theorem 2.1.

Corollary 2 *Let* $\theta \geq 0$ *and let* p *be a real number with* $0 < p < 2$. *Let* X *be a normed vector space with norm* $\|.\|$. *Let* $f : X \to Y$ *be a mapping with* $f(0) = 0$ *satisfying*

$$N\left(f\left(\sum_{i=1}^{n} a_i x_i\right) + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} f(a_i x_i \pm a_j x_j) - (3n-2)\sum_{i=1}^{n} a_i^2 f(x_i), t\right)$$
$$\geq \frac{t}{t + \theta\left(\sum_{i=1}^{n} \|x_i\|^p\right)}$$

for all $x_1, x_2, \cdots, x_n \in X$ *and all* $t > 0$. *Then, the limit* $Q(x) := N\text{-}\lim_{m\to\infty} \frac{f(a_1^m x)}{a_1^{2m}}$ *exists for each* $x \in X$ *and defines a unique quadratic mapping* $Q : X \to Y$ *such that*

$$N(f(x) - Q(x), t) \geq \frac{a_1^2(3n-2)(a_1^2 - |a_1|^p)t}{a_1^2(3n-2)(a_1^2 - |a_1|^p)t + a_1^2\theta\|x\|^p}$$

for all $x \in X$.

Proof The proof follows from Theorem 2.2 by taking $\varphi(x_1, \cdots, x_n) := \theta\left(\sum_{i=1}^{n} \|x_i\|^p\right)$ for all $x_1, \cdots, x_n \in X$. Then we can choose $L = |a_1|^{p-2}$ and we get the desired result.

3 Fuzzy Stability of Functional Equation (1): A Direct Method

In this section, using direct method, we prove the Hyers-Ulam-Rassias stability of functional equation (1) in fuzzy Banach spaces. Throughout this section, we assume that X is a linear space, (Y, N) is a fuzzy Banach space and (Z, N') is a fuzzy normed spaces. Moreover, we assume that $N(x, .)$ is a left continuous function on $\mathbb{R}$.

Theorem 5 *Assume that a mapping* $f : X \to Y$ *with* $f(0) = 0$ *satisfies the inequality*

$$N\left(f\left(\sum_{i=1}^{n} a_i x_i\right) + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} f(a_i x_i \pm a_j x_j) - (3n-2)\sum_{i=1}^{n} a_i^2 f(x_i), t\right) \geq N'(\varphi(x_1, \cdots, x_n), t) \quad (11)$$

for all $x_1, \cdots, x_n \in X$, $t > 0$ *and* $\varphi : X^n \to Z$ *is a mapping for which there is a constant* $r \in \mathbb{R}$ *satisfying* $0 < a_1^2|r| < 1$ *such that*

$$N'\left(\varphi\left(\frac{x_1}{a_1}, \frac{x_2}{a_1}, \cdots, \frac{x_n}{a_1}\right), t\right) \geq N'\left(\varphi(x_1, \cdots, x_n), \frac{t}{|r|}\right) \quad (12)$$

for all $x_1, \cdots, x_n \in X$ *and all* $t > 0$. *Then there exists a unique quadratic mapping* $Q : X \to Y$ *satisfying (1) and the inequality*

$$N(f(x) - Q(x), t) \geq N'\left(\varphi(x, 0, \cdots, 0), \frac{(3n-2)(1 - a_1^2|r|)t}{|r|}\right) \quad (13)$$

for all $x \in X$ *and all* $t > 0$.

Proof It follows from (12) that

$$N'\left(\varphi\left(\frac{x_1}{a_1^j}, \frac{x_2}{a_1^j}, \cdots, \frac{x_n}{a_1^j}\right), t\right) \geq N'\left(\varphi(x_1, \cdots, x_n), \frac{t}{|r|^j}\right)$$

for all $x_1, \cdots, x_n \in X$ and all $t > 0$. Putting $x_1 = x$ and $x_2 = \cdots = x_n = 0$ in (11), using $f(0) = 0$ and then replacing x by $\frac{x}{2}$, we have

$$N\left(a_1^2 f\left(\frac{x}{a_1}\right) - f(x), \frac{t}{3n-2}\right) \geq N'\left(\varphi\left(\frac{x}{a_1}, 0, \cdots, 0\right), t\right) \quad (14)$$

for all $x \in X$ and all $t > 0$. Replacing x by $\frac{x}{a_1^j}$ in (14), we have

$$N\left(a_1^{2j+2} f\left(\frac{x}{a_1^{j+1}}\right) - a_1^{2j} f\left(\frac{x}{a_1^j}\right), \frac{a_1^{2j} t}{3n-2}\right) \geq N'\left(\varphi\left(\frac{x}{a_1^{j+1}}, 0, \cdots, 0\right), t\right) \geq N'\left(\varphi(x, 0, \cdots, 0), \frac{t}{|r|^{j+1}}\right) \quad (15)$$

for all $x \in X$, all $t > 0$ and any integer $j \geq 0$. So

$$\begin{aligned}
&N\left(f(x)-a_1^{2m}f\left(\frac{x}{a_1^m}\right),\sum_{j=0}^{m-1}\frac{a_1^{2j}|r|^{j+1}t}{3n-2}\right)\\
&=N\left(\sum_{j=0}^{m-1}a_1^{2j+2}f\left(\frac{x}{a_1^{j+1}}\right)-a_1^{2j}f\left(\frac{x}{a_1^j}\right),\sum_{j=0}^{m-1}\frac{a_1^{2j}|r|^{j+1}t}{3n-2}\right)\\
&\geq \min_{0\leq j\leq n-1}\left\{N\left(a_1^{2j+2}f\left(\frac{x}{a_1^{j+1}}\right)-a_1^{2j}f\left(\frac{x}{a_1^j}\right),\frac{a_1^{2j}|r|^{j+1}t}{3n-2}\right)\right\}\\
&\geq N'(\varphi(x,0,\cdots,0),t)
\end{aligned}$$

which yields

$$\begin{aligned}
&N\left(a_1^{2m+2p}f\left(\frac{x}{a_1^{m+p}}\right)-a_1^{2p}f\left(\frac{x}{a_1^p}\right),\sum_{j=0}^{m-1}\frac{a_1^{2j+2p}|r|^{j+1}t}{3n-2}\right)\\
&\geq N'\left(\varphi\left(\frac{x}{a_1^p},0,\cdots,0\right),t\right)\\
&\geq N'\left(\varphi(x,0,\cdots,0),\frac{t}{|r|^p}\right)
\end{aligned}$$

for all $x\in X, t>0$ and any integers $n>0,\ p\geq 0$. So

$$\begin{aligned}
&N\left(a_1^{2m+2p}f\left(\frac{x}{a_1^{m+p}}\right)-a_1^{2p}f\left(\frac{x}{a_1^p}\right),\sum_{j=0}^{m-1}\frac{a_1^{2j+2p}|r|^{j+p+1}t}{3n-2}\right)\\
&\geq N'\left(\varphi(x,0,\cdots,0),t\right)
\end{aligned}$$

for all $x\in X, t>0$ and any integers $n>0,\ p\geq 0$. Hence one obtains

$$N\left(a_1^{2m+2p}f\left(\frac{x}{a_1^{m+p}}\right)-a_1^{2p}f\left(\frac{x}{a_1^p}\right),t\right)\geq N'\left(\varphi(x,0,..,0),\frac{t}{\sum_{j=0}^{m-1}\frac{a_1^{2j+2p}|r|^{j+p+1}}{3n-2}}\right) \quad (16)$$

for all $x\in X, t>0$ and any integers $n>0,\ p\geq 0$. Since, the series $\sum_{j=0}^{+\infty}a_1^{2j}|r|^j$ is convergent series, we see by taking the limit $p\to\infty$ in the last inequality that the sequence $\left\{a_1^{2m}f\left(\frac{x}{a_1^m}\right)\right\}$ is a Cauchy sequence in the fuzzy Banach space (Y,N) and so it converges in Y. Therefore a mapping $Q:X\to Y$ defined by

$$Q(x):=N-\lim_{m\to\infty}a_1^{2m}f\left(\frac{x}{a_1^m}\right)$$

is well defined for all $x \in X$. It means that

$$\lim_{m\to\infty} N\left(Q(x) - a_1^{2m} f\left(\frac{x}{a_1^m}\right), t\right) = 1 \tag{17}$$

for all $x \in X$ and all $t > 0$. In addition, it follows from (16) that

$$N\left(f(x) - a_1^{2m} f\left(\frac{x}{a_1^m}\right), t\right) \geq N'\left(\varphi(x, 0, .., 0), \frac{t}{\sum_{j=0}^{m-1} \frac{a_1^{2j}|r|^{j+1}}{3n-2}}\right)$$

for all $x \in X$ and all $t > 0$. So

$$\begin{aligned}
&N(f(x) - Q(x), t)\\
&\geq \min\left\{N\left(f(x) - a_1^{2m} f\left(\frac{x}{a_1^m}\right), (1-\varepsilon)t\right), N\left(A(x)a_1^{2m} f\left(\frac{x}{a_1^m}\right), \varepsilon t\right)\right\}\\
&\geq N'\left(\varphi(x, 0, .., 0), \frac{t}{\sum_{j=0}^{m-1} \frac{a_1^{2j}|r|^{j+1}}{3n-2}}\right) \geq N'\left(\varphi(x, 0, \cdots, 0), \frac{(3n-2)(1-a_1^2|r|)\varepsilon t}{|r|}\right)
\end{aligned}$$

for sufficiently large m and for all $x \in X$, $t > 0$ and ε with $0 < \varepsilon < 1$. Since ε is arbitrary and N' is left continuous, we obtain

$$N(f(x) - Q(x), t) \geq N'\left(\varphi(x, 0, \cdots, 0), \frac{(3n-2)(1-a_1^2|r|)t}{|r|}\right)$$

for all $x \in X$ and $t > 0$. It follows from (11) that

$$\begin{aligned}
&N\left(a_1^{2m} f\left(\sum_{i=1}^{n} \frac{a_i x_i}{a_1^m}\right) + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_1^{2m} f\left(\frac{a_i x_i \pm a_j x_j}{a_1^m}\right)\right.\\
&\left.-(3n-2)\sum_{i=1}^{n} a_i^2 a_1^{2m} f\left(\frac{x_i}{a_1^m}\right), t\right)\\
&\geq N'\left(\varphi\left(\frac{x_1}{a_1^m}, \frac{x_2}{a_1^m}, \cdots, \frac{x_n}{a_1^m}\right), \frac{t}{a_1^{2m}}\right) \geq N'\left(\varphi(x_1, x_2, \cdots, x_n), \frac{t}{a_1^{2m}|r|^m}\right)
\end{aligned}$$

for all $x_1, x_2, \cdots, x_n \in X$, $t > 0$ and all $n \in \mathbb{N}$. Since

$$\lim_{m\to\infty} N'\left(\varphi(x_1, x_2, \cdots, x_n), \frac{t}{a_1^{2m}|r|^m}\right) = 1$$

and so

$$N\left(a_1^{2m} f\left(\sum_{i=1}^{n} \frac{a_i x_i}{a_1^m}\right) + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_1^{2m} f\left(\frac{a_i x_i \pm a_j x_j}{a_1^m}\right)\right.$$
$$\left. -(3n-2)\sum_{i=1}^{n} a_i^2 a_1^{2m} f\left(\frac{x_i}{a_1^m}\right), t\right) \to 1$$

for all $x_1, x_2, \cdots, x_n \in X$ and all $t > 0$. Therefore, we obtain in view of (17)

$$N\left(Q\left(\sum_{i=1}^{n} a_i x_i\right) + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Q(a_i x_i \pm a_j x_j) - (3n-2)\sum_{i=1}^{n} a_i^2 Q(x_i), t\right)$$
$$\geq \min\left\{N\left(Q\left(\sum_{i=1}^{n} a_i x_i\right) + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Q(a_i x_i \pm a_j x_j) - (3n-2)\sum_{i=1}^{n} a_i^2 Q(x_i)\right.\right.$$
$$-a_1^{2m} f\left(\sum_{i=1}^{n} \frac{a_i x_i}{a_1^m}\right) - \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_1^{2m} f\left(\frac{a_i x_i \pm a_j x_j}{a_1^m}\right)$$
$$\left.+(3n-2)\sum_{i=1}^{n} a_i^2 a_1^{2m} f\left(\frac{x_i}{a_1^m}\right), \frac{t}{2}\right),$$
$$N\left(a_1^{2m} f\left(\sum_{i=1}^{n} \frac{a_i x_i}{a_1^m}\right) + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_1^{2m} f\left(\frac{a_i x_i \pm a_j x_j}{a_1^m}\right)\right.$$
$$\left.\left.-(3n-2)\sum_{i=1}^{n} a_i^2 a_1^{2m} f\left(\frac{x_i}{a_1^m}\right), \frac{t}{2}\right)\right\}$$
$$\geq N\left(a_1^{2m} f\left(\sum_{i=1}^{n} \frac{a_i x_i}{a_1^m}\right) + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_1^{2m} f\left(\frac{a_i x_i \pm a_j x_j}{a_1^m}\right)\right.$$
$$\left.-(3n-2)\sum_{i=1}^{n} a_i^2 a_1^{2m} f\left(\frac{x_i}{a_1^m}\right), \frac{t}{2}\right)$$
$$\geq N'\left(\varphi(x_1, x_2, \cdots, x_n), \frac{t}{2a_1^{2m}|r|^m}\right) \to 1 \quad \text{as } m \to \infty$$

which implies

$$Q\left(\sum_{i=1}^{n} a_i x_i\right) + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q(a_i x_i \pm a_j x_j) = (3n-2) \sum_{i=1}^{n} a_i^2 Q(x_i)$$

for all $x_1, x_2, \cdots, x_n \in X$. Thus $Q : X \to Y$ is a mapping satisfying the Eq. (1) and the inequality (13). Thus the mapping $Q : X \to Y$ is quadratic, as desired. To prove the uniqueness, let there is another mapping $R : X \to Y$ which satisfies the inequality (13). Since $R(x) = a_1^{2m} R\left(\frac{x}{a_1^m}\right)$ for all $x \in X$, we have

$$\begin{aligned}
&N(Q(x) - R(x), t)\\
&= N\left(a_1^{2m} Q\left(\frac{x}{a_1^m}\right) - a_1^{2m} R\left(\frac{x}{a_1^m}\right), t\right)\\
&\geq \min\left\{N\left(a_1^{2m} Q\left(\frac{x}{a_1^m}\right) - a_1^{2m} f\left(\frac{x}{a_1^m}\right), \frac{t}{2}\right),\right.\\
&\left. N\left(a_1^{2m} f\left(\frac{x}{a_1^m}\right) - a_1^{2m} R\left(\frac{x}{a_1^m}\right), \frac{t}{2}\right)\right\}\\
&\geq N'\left(\varphi\left(\frac{x}{a_1^m}, 0, \cdots, 0\right), \frac{(3n-2)(1-a_1^2|r|)t}{a_1^{2m}|r|}\right)\\
&\geq N\left(\varphi(x, 0, \cdots, 0), \frac{(3n-2)(1-a_1^2|r|)t}{a_1^{2m}|r|^{m+1}}\right) \to 1
\end{aligned}$$

as $m \to \infty$ for all $t > 0$. Therefore $Q(x) = R(x)$ for all $x \in X$. This completes the proof.

Corollary 3 *Let X be a normed spaces and that $(\mathbb{R}, N')$ a fuzzy Banach space. Assume that there exists real number $\theta \geq 0$ and $p > 1$ such that a mapping $f : X \to Y$ with $f(0) = 0$ satisfies the following inequality*

$$N\left(f\left(\sum_{i=1}^{n} a_i x_i\right) + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} f(a_i x_i \pm a_j x_j) - (3n-2) \sum_{i=1}^{n} a_i^2 f(x_i), t\right) \geq N'\left(\theta\left(\sum_{i=1}^{n} \|x_i\|^p\right), t\right) \tag{18}$$

for all $x_1, x_2, \cdots, x_n \in X$ and $t > 0$. Then there is a unique quadratic mapping $Q : X \to Y$ that satisfying (1) and the inequality

$$N(f(x) - Q(x), t) \geq N'\left(\frac{\theta\|x\|^p}{(3n-2)(a_1^{2p} - a_1^2)}, t\right)$$

Proof Let $\varphi(x_1, x_2, \cdots, x_n) := \theta\left(\sum_{i=1}^n \|x_i\|^p\right)$ and $|r| = |a_1|^{-2p}$. Apply Theorem 5, we get desired results.

Theorem 6 *Assume that a mapping $f : X \to Y$ with $f(0) = 0$ satisfies the inequality (11) and $\varphi : X^n \to Z$ is a mapping for which there is a constant $r \in \mathbb{R}$ satisfying $0 < |r| < a_1^2$ such that*

$$N'\left(\frac{\varphi(x_1, x_2, \cdots, x_n)}{|r|}, t\right) \geq N'\left(\varphi\left(\frac{x_1}{a_1}, \frac{x_2}{a_1}, \cdots, \frac{x_n}{a_1}\right), t\right) \tag{19}$$

for all $x, y \in X$ and all $t > 0$. Then there exists a unique quadratic mapping $Q : X \to Y$ that satisfying (1) and the following inequality

$$N(f(x) - Q(x), t) \geq N'\left(\frac{\varphi(x, 0, \cdots, 0)}{(3n-2)(a_1^2 - |r|)}, t\right) \tag{20}$$

for all $x \in X$ and all $t > 0$.

Proof Putting $x_1 = x$ and $x_2 = \cdots = x_n = 0$ in (11), using $f(0) = 0$, we have

$$N\left(\frac{f(a_1 x)}{a_1^2} - f(x), \frac{t}{a_1^2(3n-2)}\right) \geq N'(\varphi(x, 0, \cdots, 0), t) \tag{21}$$

for all $x \in X$ and all $t > 0$. Replacing x by $a_1^m x$ in (21), we obtain

$$\begin{aligned} N\left(\frac{f(a_1^{m+1}x)}{a_1^{2m+2}} - \frac{f(a_1^m x)}{a_1^{2m}}, \frac{t}{a_1^{2m+2}(3n-2)}\right) &\geq N'(\varphi(a_1^m x, 0, \cdots, 0), t) \\ &\geq N'\left(\varphi(x, 0, \cdots, 0), \frac{t}{|r|^m}\right). \end{aligned} \tag{22}$$

So

$$N\left(\frac{f(a_1^{m+1}x)}{a_1^{2m+2}} - \frac{f(a_1^m x)}{a_1^{2m}}, \frac{|r|^m t}{a_1^{2m+2}(3n-2)}\right) \geq N'(\varphi(x, 0, \cdots, 0), t) \tag{23}$$

for all $x \in X$ and all $t > 0$. Proceeding as in the proof of Theorem 5, we obtain that

$$N\left(f(x) - \frac{f(a_1^m x)}{a_1^{2m}}, \sum_{j=0}^{m-1} \frac{|r|^j t}{a_1^{2j+2}(3n-2)}\right) \geq N'(\varphi(x, 0, \cdots, 0), t)$$

for all $x \in X$, all $t > 0$ and any integer $n > 0$. So

$$N\left(f(x) - \frac{f(a_1^m x)}{a_1^{2m}}, t\right) \geq N'\left(\varphi(x, 0, \cdots, 0), \frac{t}{\sum_{j=0}^{m-1} \frac{|r|^j}{a_1^{2j+2}(3n-2)}}\right)$$
$$\geq N'\left(\frac{\varphi(x, 0, \cdots, 0)}{(3n-2)(a_1^2 - |r|)}, t\right).$$

The rest of the proof is similar to the proof of Theorem 5.

Corollary 4 *Let X be a normed spaces and that $(\mathbb{R}, N')$ a fuzzy Banach space. Assume that there exists real number $\theta \geq 0$ and $0 < p < 1$ such that a mapping $f : X \to Y$ with $f(0) = 0$ satisfies (18). Then there is a unique quadratic mapping $Q : X \to Y$ that satisfying (1) and the inequality*

$$N(f(x) - Q(x), t) \geq N'\left(\frac{\theta \|x\|^p}{(3n-2)(a_1^2 - a_1^{2p})}, t\right)$$

Proof Let $\varphi(x_1, x_2, \cdots, x_n) := \theta\left(\sum_{i=1}^{n} \|x_i\|^p\right)$ and $|r| = |a_1|^{2p}$. Apply Theorem 6, we get desired results.

References

1. T. Bag, S.K. Samanta, Finite dimensional fuzzy normed linear spaces. J. Fuzzy Math. **11**, 687–705 (2003)
2. T. Bag, S.K. Samanta, Fuzzy bounded linear operators. Fuzzy Sets Syst. **151**, 513–547 (2005)
3. S.C. Cheng, J.N. Mordeson, Fuzzy linear operators and fuzzy normed linear spaces. Bull. Calcutta Math. Soc. **86**, 429–436 (1994)
4. Y.J. Cho, R. Saadati, Lattice non-Archimedean random stability of ACQ functional equation. Adv. Differ. Equ. **2011**, 31 (2011)
5. P.W. Cholewa, Remarks on the stability of functional equations. Aequationes Math. **27**, 76–86 (1984)
6. S. Czerwik, On the stability of the quadratic mapping in normed spaces. Abh. Math. Semin. Univ. Hambg. **62**, 239–248 (1992)
7. M. Eshaghi Gordji, M. Bavand Savadkouhi, Stability of mixed type cubic and quartic functional equations in random normed spaces. J. Inequal. Appl. **2009**, Article ID 527462, 9 pp. (2009)
8. M. Eshaghi Gordji, H. Khodaei, *Stability of Functional Equations* (Lap Lambert Academic Publishing, Saarbrücken, 2010)
9. M. Eshaghi Gordji, M. Bavand Savadkouhi, C. Park, Quadratic-quartic functional equations in RN-spaces. J. Inequal. Appl. **2009**, Article ID 868423, 14 pp. (2009)
10. M. Eshaghi Gordji, S. Zolfaghari, J.M. Rassias, M.B. Savadkouhi, Solution and stability of a mixed type cubic and quartic functional equation in quasi-Banach spaces. Abstr. Appl. Anal. **2009**, Article ID 417473, 14 pp. (2009)
11. C. Felbin, Finite-dimensional fuzzy normed linear space. Fuzzy Sets Syst. **48**, 239–248 (1992)

12. P. Găvruta, A generalization of the Hyers-Ulam-Rassias stability of approximately additive mappings. J. Math. Anal. Appl. **184**, 431–436 (1994)
13. D.H. Hyers, On the stability of the linear functional equation. Proc. Natl. Acad. Sci. U. S. A. **27**, 222–224 (1941)
14. S.-M. Jung, M.Th. Rassias, A linear functional equation of third order associated to the Fibonacci numbers. Abstr. Appl. Anal. **2014**, Article ID 137468 (2014)
15. S.-M. Jung, M.Th. Rassias, C. Mortici, On a functional equation of trigonometric type. Appl. Math. Comput. **252**, 294–303 (2015)
16. S.-M. Jung, D. Popa, M.Th. Rassias, On the stability of the linear functional equation in a single variable on complete metric groups. J. Glob. Optim. **59**, 165–171 (2014)
17. I. Karmosil, J. Michalek, Fuzzy metric and statistical metric spaces. Kybernetica **11**, 326–334 (1975)
18. A.K. Katsaras, Fuzzy topological vector spaces. Fuzzy Sets Syst. **12**, 143–154 (1984)
19. S.V. Krishna, K.K.M. Sarma, Separation of fuzzy normed linear spaces. Fuzzy Sets Syst. **63**, 207–217 (1994)
20. D. Mihet, V. Radu, On the stability of the additive Cauchy functional equation in random normed spaces. J. Math. Anal. Appl. **343**, 567–572 (2008)
21. M. Mohammadi, Y.J. Cho, C. Park, P. Vetro, R. Saadati, Random stability of an additive-quadratic-quartic functional equation. J. Inequal. Appl. **2010**, Article ID 754210, 18 pp. (2010)
22. A. Najati, Y.J. Cho, Generalized Hyers–Ulam stability of the Pexiderized Cauchy functional equation in non-Archimedean spaces. Fixed Point Theory Appl. **2011**, Article ID 309026, 11 pp. (2011)
23. A. Najati, J.I. Kang, Y.J. Cho, Local stability of the Pexiderized Cauchy and Jensen's equations in fuzzy spaces. J. Inequal. Appl. **2011**, 78 (2011)
24. C. Park, On the stability of the linear mapping in Banach modules. J. Math. Anal. Appl. **275**, 711–720 (2002)
25. C. Park, Generalized Hyers-Ulam-Rassias stability of n-sesquilinear-quadratic mappings on Banach modules over C^*-algebras. J. Comput. Appl. Math. **180**, 279–291 (2005)
26. C. Park, Fixed points and Hyers-Ulam-Rassias stability of Cauchy-Jensen functional equations in Banach algebras. Fixed Point Theory Appl. **2007**, Art. ID 50175 (2007)
27. C. Park, Generalized Hyers-Ulam-Rassias stability of quadratic functional equations: a fixed point approach. Fixed Point Theory Appl. **2008**, Art. ID 493751 (2008)
28. C. Park, Fuzzy stability of a functional equation associated with inner product spaces. Fuzzy Sets Syst. **160**, 1632–1642 (2009)
29. Th.M. Rassias, On the stability of the linear mapping in Banach spaces. Proc. Am. Math. Soc. **72**, 297–300 (1978)
30. Th.M. Rassias, On the stability of the quadratic functional equation and it's application. Stud. Univ. Babes-Bolyai **XLIII**, 89–124 (1998)
31. Th.M. Rassias, On the stability of functional equations in Banach spaces. J. Math. Anal. Appl. **251**, 264–284 (2000)
32. Th.M. Rassias, P. Šemrl, On the Hyers-Ulam stability of linear mappings. J. Math. Anal. Appl. **173**, 325–338 (1993)
33. R. Saadati, C. Park, Non-Archimedean $\mathscr{L}$-fuzzy normed spaces and stability of functional equations. Comput. Math. Appl. **60**(8), 2488–2496 (2010)
34. R. Saadati, M. Vaezpour, Y.J. Cho, A note to paper "On the stability of cubic mappings and quartic mappings in random normed spaces". J. Inequal. Appl. **2009**, Article ID 214530 (2009). https://doi.org/10.1155/2009/214530
35. R. Saadati, M.M. Zohdi, S.M. Vaezpour, Nonlinear L-random stability of an ACQ functional equation. J. Inequal. Appl. **2011**, Article ID 194394, 23 pp. (2011). https://doi.org/10.1155/2011/194394
36. F. Skof, Local properties and approximation of operators. Rend. Semin. Mat. Fis. Milano **53**, 113–129 (1983)
37. S.M. Ulam, *Problems in Modern Mathematics* (Wiley, New York, 1964)

Asymptotic Orbits in Hill's Problem When the Larger Primary is a Source of Radiation

Vassilis S. Kalantonis, Angela E. Perdiou, and Christos N. Douskos

1 Introduction

The dynamics around the collinear libration points of the restricted three-body problem or its Hill limiting case has attracted the interest of many researchers in the last decades both from theoretical and practical point of view (see, e.g., [9, 22, 26] and references therein). A special case of motion around these points is when the third body of negligible mass tracks orbits which asymptotically depart from and arrive at the collinear equilibrium points themselves or the Lyapunov periodic orbits existing in their vicinity (see, for example, [3, 5, 6, 14]).

Orbits which start asymptotically from an equilibrium point (or a Lyapunov orbit) and terminate asymptotically at the same point (or orbit) are called homoclinic while orbits which asymptotically start from an equilibrium point (or a Lyapunov orbit) and terminate, in the same way, at another equilibrium point (or another periodic orbit) are called heteroclinic. Asymptotic orbits at a collinear equilibrium point can be considered as the limiting case of asymptotic orbits to a Lyapunov periodic orbit since they are orbits which emanate from the collinear point and terminate at it or another equilibrium point, instead of from finite periodic orbits around them. This means that, an asymptotic orbit at a collinear equilibrium point can be used as a reference orbit since its existence indicates the existence, in its immediate

V. S. Kalantonis (✉)
Department of Electrical and Computer Engineering, University of Patras, Patras, Greece
e-mail: kalantonis@upatras.gr

A. E. Perdiou
Department of Civil Engineering, University of Patras, Patras, Greece
e-mail: aperdiou@upatras.gr

C. N. Douskos
Department of Computer Engineering and Informatics, University of Patras, Patras, Greece
e-mail: douskosc@upatras.gr

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_18

neighbourhood, of an infinity of orbits asymptotic to the Lyapunov periodic orbits [10]. These orbits have been studied by Deprit and Henrard [4] and Perdios and Markellos [19], in the framework of the restricted three-body problem. On the other hand, orbits asymptotic to the Lyapunov periodic orbits emanating from the collinear equilibrium points are important from both theoretical and practical point of view since they cause the destruction of invariant tori while they can also be used for the design of trajectories for space missions [1, 7, 11, 25].

In the present work, we study homoclinic orbits at both the collinear equilibrium points and the Lyapunov periodic orbits in a modification of Hill's problem where the larger primary, i.e. the Sun, is a source of radiation. A similar study but only for the case of asymptotic orbits to periodic orbits has been done by Papadakis [17] for the corresponding photogravitational restricted three-body problem. To determine orbits which asymptotically terminate at these points we use fourth order expansions with respect to a small orbital parameter while for the determination of asymptotic orbits to the Lyapunov periodic orbits we compute the corresponding unstable manifolds by applying a similar analysis based on the iso-energetic stability indices. These analytical solutions have been used for obtaining appropriate initial conditions for the numerical integration of the equations of motion and the accurate computation of the asymptotic orbits. Finally, in the case of homoclinic orbits to the Lyapunov periodic orbits, certain Poincaré surface of section portraits have been constructed in order to detect transversality of the stable and unstable manifolds and several asymptotic orbits have been determined for the specific value of the radiation factor $Q_1 = 0.5$. Our paper is organized as follows: In Sect. 2, we recall the equations of motion of the model-problem as well as we discuss its equilibrium points. In Sect. 3, we present our results for homoclinic orbits at the collinear equilibria while the corresponding homoclinic orbits to the Lyapunov periodic orbits are shown and described in Sect. 4. Finally, in Sect. 5, we conclude.

2 Equations of Motion and Equilibrium Points

The Hill problem where the primary is a source of radiation is derived from the corresponding restricted three-body problem in a similar way as the classical Hill problem is derived from the classical restricted problem [15]. The restricted three-body problem with radiation [23] describes a different point of view for the motion of small particles than that of the classical restricted problem and many researchers have studied it (see [2, 12, 16, 18, 24], among others). This extended model takes into account only the radiation pressure component of the radiation drag, this being the most powerful component besides the gravitational forces and has been proposed, for example, for the consideration of the mass loss process from very luminous stars (see [13, 20, 21]).

In rotating coordinates, the Hill problem in which the larger primary is a source of radiation is described by the following equations of motion:

$$\ddot{x} - 2\dot{y} = \frac{\partial W}{\partial x} = 3x - \frac{x}{r^3} - Q_1, \qquad \ddot{y} + 2\dot{x} = \frac{\partial W}{\partial y} = -\frac{y}{r^3}, \tag{1}$$

where the potential function W is:

$$W = \frac{3x^2}{2} - Q_1 x + \frac{1}{r}, \qquad r = \sqrt{x^2 + y^2}, \tag{2}$$

and the equation of the Jacobi integral is $2W - (\dot{x}^2 + \dot{y}^2) = \Gamma$, where Γ is the Jacobi constant. These equations result from the restricted three-body problem with radiation by placing the origin at the smaller primary, appropriate rescaling of lengths and applying the transformation $q_1 = 1 - Q_1 \mu^{1/3}$ for the radiation factor, where μ is the usual mass parameter.

The problem admits two collinear equilibrium points; L_1 on the negative axis and L_2 on the positive axis. The x-axis is an axis of symmetry but, contrary to the classical Hill problem, the y-axis is not. The positions of these equilibria are given by the following exact formula [15]:

$$x_{L_i} = \frac{\sigma}{9}\left[f(Q_1) + \sigma Q_1 + \frac{Q_1^2}{f(Q_1)}\right], \tag{3}$$

where

$$f(Q_1) = \frac{3^{5/3}}{2^{1/3}}\left[1 + g(Q_1) + \sqrt{1 + 2g(Q_1)}\right]^{1/3}, \qquad g(Q_1) = \frac{2\sigma}{9}(\frac{Q_1}{3})^3, \tag{4}$$

with $\sigma = (-1)^i$, $i = 1, 2$. To study the stability of the equilibrium points, we transfer the origin to the equilibrium point L_i by setting $x = x_{L_i} + \xi$, $y = \eta$, $i = 1, 2$, and linearize the equations of motion obtaining the system:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}, \qquad \mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ A_1 & 0 & 0 & 2 \\ 0 & B_1 & -2 & 0 \end{bmatrix}, \tag{5}$$

where $\mathbf{x} = (\xi, \eta, \dot{\xi}, \dot{\eta})^{\mathrm{T}}$ and:

$$A_1 = 3 + \frac{2}{r_0^3}, \qquad B_1 = -\frac{1}{r_0^3}. \tag{6}$$

The characteristic equation of System (5) has two real and two pure imaginary roots, so, due to the real roots the equilibrium points are unstable. Of interest here, for our next section, are the real roots λ_0 generating the stable and unstable manifolds of the equilibrium points:

$$\lambda_0 = \pm\sqrt{w_1 + \sqrt{w_1^2 - w_2}}, \qquad w_1 = \frac{A_1 + B_1 - 4}{2}, \qquad w_2 = A_1 B_1. \tag{7}$$

3 Asymptotic Orbits at Collinear Equilibria

The equations of motion to fourth order terms are:

$$\begin{aligned} \ddot{\xi} - 2\dot{\eta} &= A_1\xi + A_2\xi^2 + A_3\eta^2 + A_4\xi\eta^2 + A_5\xi^3 + A_6\xi^4 + A_7\eta^4 + A_8\xi^2\eta^2, \\ \ddot{\eta} + 2\dot{\xi} &= B_1\eta + B_2\xi\eta + B_3\xi^2\eta + B_4\eta^3 + B_5\xi\eta^3 + B_6\xi^3\eta, \end{aligned} \tag{8}$$

where the higher order coefficients of the right-hand sides are expressions depending on the radiation factor, through the distance $r_0 = |x_{L_i}|$, $i = 1, 2$ of the equilibrium point, and are given by:

$$\begin{aligned} &A_2 = -\sigma\frac{3}{r_0^4}, && A_3 = \sigma\frac{3}{2r_0^4}, && A_4 = -\frac{6}{r_0^5}, && A_5 = \frac{4}{r_0^5}, \\ &A_6 = -\sigma\frac{5}{r_0^6}, && A_7 = -\sigma\frac{15}{8r_0^6}, && A_8 = \sigma\frac{15}{r_0^6}, && B_2 = \sigma\frac{3}{r_0^4}, \\ &B_3 = -\frac{6}{r_0^5}, && B_4 = \frac{3}{2r_0^5}, && B_5 = -\sigma\frac{15}{2r_0^6}, && B_6 = \sigma\frac{10}{r_0^6}. \end{aligned} \tag{9}$$

We now express the solution of the above system in the form of series expansions in terms of a small orbital parameter ϵ:

$$\xi(t) = \sum_{j=1}^{4} \epsilon^j \phi_j(t), \qquad \eta(t) = \sum_{j=1}^{4} \epsilon^j \omega_j(t). \tag{10}$$

By substituting this formal solution into System (8) the following systems are derived:

$$\begin{aligned} \ddot{\phi}_1 - 2\dot{\omega}_1 &= A_1\phi_1, \\ \ddot{\omega}_1 + 2\dot{\phi}_1 &= B_1\omega_1, \end{aligned} \tag{11}$$

$$\begin{aligned} \ddot{\phi}_2 - 2\dot{\omega}_2 &= A_1\phi_2 + A_2\phi_1^2 + A_3\omega_1^2, \\ \ddot{\omega}_2 + 2\dot{\phi}_2 &= B_1\omega_2 + B_2\phi_1\omega_1, \end{aligned} \tag{12}$$

$$\begin{aligned}
\ddot{\phi}_3 - 2\dot{\omega}_3 &= A_1\phi_3 + 2A_2\phi_1\phi_2 + 2A_3\omega_1\omega_2 + A_4\phi_1\omega_1^2 + A_5\phi_1^3,\\
\ddot{\omega}_3 + 2\dot{\phi}_3 &= B_1\omega_3 + B_2\phi_1\omega_2 + B_2\phi_2\omega_1 + B_3\phi_1^2\omega_1 + B_4\omega_1^3,
\end{aligned} \tag{13}$$

$$\begin{aligned}
\ddot{\phi}_4 - 2\dot{\omega}_4 = {} & A_1\phi_4 + A_2\phi_2^2 + 2A_2\phi_1\phi_3 + A_3\omega_2^2 + 2A_3\omega_1\omega_3 + A_4\phi_2\omega_1^2 + \\
& +2A_4\phi_1\omega_1\omega_2 + 3A_5\phi_1^2\phi_2 + A_6\phi_1^4 + A_7\omega_1^4 + A_8\phi_1^2\omega_1^2,\\
\ddot{\omega}_4 + 2\dot{\phi}_4 = {} & B_1\omega_4 + B_2\phi_2\omega_2 + B_2\phi_1\omega_3 + B_2\phi_3\omega_1 + 2B_3\phi_1\phi_2\omega_1 + \\
& +B_3\phi_1^2\omega_2 + 3B_4\omega_1^2\omega_2 + B_5\phi_1\omega_1^3 + B_6\phi_1^3\omega_1.
\end{aligned} \tag{14}$$

In order to determine an asymptotic orbit to the collinear equilibrium point L_i, $i = 1, 2$, we consider the solutions which correspond to the real eigenvalues. The solution of the first-order System (11), corresponding to $\lambda_0 > 0$, is directly obtained in the form $\xi(t) = \epsilon h_1 e^{\lambda_0 t}$, $\eta(t) = \epsilon g_1 e^{\lambda_0 t}$, where:

$$h_1 = 1, \qquad g_1 = \frac{\lambda_0^2 - A_1}{2\lambda_0} = \frac{2\lambda_0}{B_1 - \lambda_0^2}. \tag{15}$$

To first order, the corresponding eigenvectors are [4]:

$$\begin{array}{llll}
\text{I} & : \ \xi = \epsilon e^{\lambda_0 t}, & \eta = \epsilon g_1 e^{\lambda_0 t}, & \text{outgoing eigenvector,}\\
\text{II} & : \ \xi = -\epsilon e^{\lambda_0 t}, & \eta = -\epsilon g_1 e^{\lambda_0 t}, & \text{outgoing eigenvector,}\\
\text{III} & : \ \xi = \epsilon e^{-\lambda_0 t}, & \eta = -\epsilon g_1 e^{-\lambda_0 t}, & \text{incoming eigenvector,}\\
\text{IV} & : \ \xi = -\epsilon e^{-\lambda_0 t}, & \eta = \epsilon g_1 e^{-\lambda_0 t}, & \text{incoming eigenvector,}
\end{array}$$

and homoclinic solutions at the collinear equilibrium points can be formed by combining the outgoing eigenvector I with the incoming eigenvector III, or by combining the outgoing eigenvector II with the incoming eigenvector IV. The orbits arising from the above eigenvectors are called asymptotic orbits of kind I, II, III, IV, respectively. The fourth order solution corresponding to the positive eigenvalue is found in the form:

$$\xi(t) = \sum_{j=1}^{4} \epsilon^j h_j e^{j\lambda_0 t}, \qquad \eta(t) = \sum_{j=1}^{4} \epsilon^j g_j e^{j\lambda_0 t}, \tag{16}$$

where the coefficients h_j, g_j, $j = 2, 3, 4$, are obtained by solving successively Systems (12)–(14). The initial conditions for an outgoing asymptotic orbit of kind I up to fourth order terms are given by:

$$\begin{aligned}
x_{0,\text{I}} &= x_{L_2} + \sum_{j=1}^{4} \epsilon^j h_j, & y_{0,\text{I}} &= \sum_{j=1}^{4} \epsilon^j g_j,\\
\dot{x}_{0,\text{I}} &= \sum_{j=1}^{4} j\epsilon^j h_j \lambda_0, & \dot{y}_{0,\text{I}} &= \sum_{j=1}^{4} j\epsilon^j g_j \lambda_0,
\end{aligned} \tag{17}$$

while the initial conditions for an outgoing asymptotic orbit of kind II are given by the same expressions by changing the value of ϵ to $-\epsilon$. Due to the symmetry

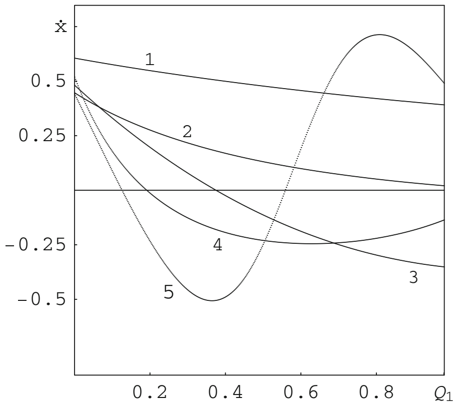


Fig. 1 The function $\dot{x}(Q_1)$ for the first five cuts of the orbit with the x-axis in the case of the negative equilibrium point L_1

of the problem w.r.t. the x-axis we are able to determine the initial conditions of the incoming asymptotic orbits III, IV from those of the outgoing orbits I, II, respectively, by changing the signs of y_0 and $\dot{x}_0$.

Due to the same symmetry, transversality of the unstable with the stable manifold of the equilibrium point is detected when the orbit reaches this axis perpendicularly, i.e. with $\dot{x}(Q_1) = 0$. Therefore, we scan the Q_1-axis and for each value of Q_1 integrate the equations of motion using the initial conditions (17) up to the n-th crossing of the orbit with the x-axis. The roots of the function $\dot{x}(Q_1)$ will indicate the existence of homoclinic orbits at the collinear equilibrium point. In Fig. 1 we show the behaviour of this function for the first five crossings of the orbit with the x-axis in the case of the "negative" equilibrium point L_1. When a homoclinic orbit has been detected, we can determine it accurately by applying well-known differential corrections procedures or by a more refined scanning. Numerical data of many homoclinic orbits are available for the readers. In Fig. 2 we show homoclinic orbits (continuous lines) at the positive collinear equilibrium point for the values of the radiation factor $Q_1 = 0.01471899$, $Q_1 = 0.03551808$ and $Q_1 = 0.72629021$ (from left to right) together with the corresponding zero-velocity curves (dashed lines).

4 Asymptotic Orbits to Lyapunov Periodic Orbits

In order to detect homoclinic orbits to periodic orbits of the Lyapunov family we shall use again the symmetry property of the problem w.r.t. the x-axis. For a particular Lyapunov orbit, the initial four-dimensional phase space is reduced to a three-dimensional subspace of iso-energetic orbits due to the equation of the Jacobi integral, for the specific value of the Jacobi constant of the Lyapunov orbit. The final reduction of the phase space to a two-dimensional sub-space is carried out by considering the cuts of the unstable manifold of the Lyapunov orbit with the

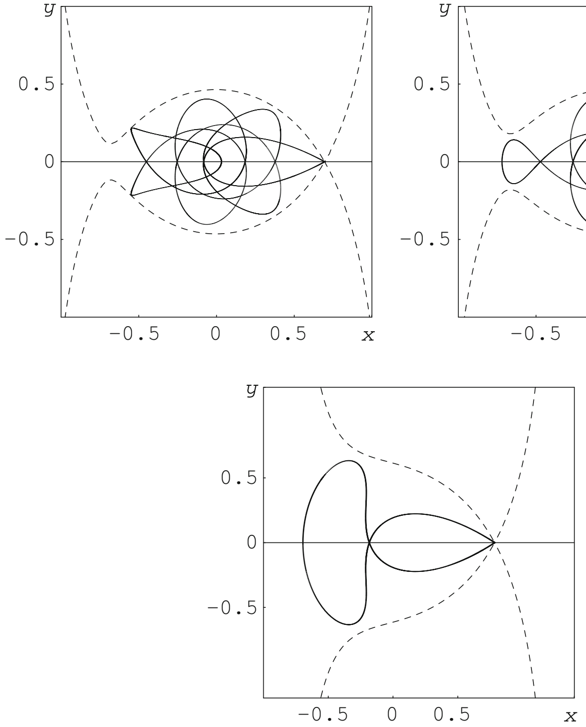


Fig. 2 Homoclinic orbits at collinear equilibrium point L_2

$y = 0$ plane, i.e. the x-axis. If the unstable manifold of the Lyapunov orbit has a perpendicular intersection with the x-axis, i.e. the horizontal component of the velocity is zero ($\dot{x} = 0$), then transversality is achieved and a homoclinic orbit to the Lyapunov orbit exists. The numerical construction of the corresponding stable–unstable manifolds is based on a linear analysis [22] and can be briefly described as follows.

Let that, for a specific value Γ_0 of the Jacobi constant, $(x_0, \dot{x}_0)$ is a fixed point on the Poincaré surface of section $y = 0$ for positive direction of the flow, i.e. $\dot{y}_0 > 0$. This fixed point corresponds to a periodic orbit with initial conditions $(x, y, \dot{x}, \dot{y}) = (x_0, 0, \dot{x}_0, [2W(x_0, 0) - \dot{x}_0^2 - \Gamma_0]^{1/2})$. The differential of the Poincaré map is given by:

$$\dot{\mathbf{X}} = \mathbf{B}\mathbf{X}_0, \quad \mathbf{B} = \begin{bmatrix} a_h & b_h \\ c_h & d_h \end{bmatrix}, \quad \mathbf{X} = (x, \dot{x})^{\mathrm{T}}, \quad \mathbf{X}_0 = (x_0, \dot{x}_0)^{\mathrm{T}}, \tag{18}$$

where $a_h = \partial x/\partial x_0$, $b_h = \partial x/\partial \dot{x}_0$, $c_h = \partial \dot{x}/\partial x_0$ and $d_h = \partial \dot{x}/\partial \dot{x}_0$ are the iso-energetic horizontal stability parameters as these were defined by Hénon [8] with $a_h d_h - b_h c_h = 1$ for the area preservation of the map. The eigenvalues of the linear map (18) are determined by the characteristic equation $|\mathbf{B} - \lambda\mathbf{I}| = 0$ which, due to symmetry of the problem with respect to the Ox-axis as well as the area preservation property, takes the following form:

$$\lambda^2 - 2a_h\lambda + 1 = 0, \tag{19}$$

therefore, obviously the motion is in general bounded when $|a_h| < 1$ while is unbounded when $|a_h| > 1$. In order to determine suitable initial conditions for the computation of the unstable manifold we take its linear approximation $\dot{\mathbf{X}} = \mathbf{X}_0 + \varepsilon\Theta_{\mathbf{u}}$ in the direction of the eigenvector which corresponds to the real eigenvalue $\lambda_u = a_h + (a_h^2 - 1)^{1/2}$ with modulus larger than one. The corresponding linear approximation may be written in the form:

$$\begin{pmatrix} x \\ \dot{x} \end{pmatrix} = \begin{pmatrix} x_0 \\ \dot{x}_0 \end{pmatrix} + \varepsilon \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \tag{20}$$

where $\varepsilon = \varepsilon_0\lambda_u^\delta = \varepsilon_0 e^{-\delta \ln \lambda_s}$, with $\lambda_s = a_h - (a_h^2 - 1)^{1/2}$ being the smallest real eigenvalue, ε_0 is a sufficiently small constant in order to ensure the validity of the above linear approximation and $\delta \in [0, 1)$. Also, the components of the corresponding eigenvector associated to the eigenvalue λ_u can be easily found to be:

$$\theta_1 = 1 \quad \text{and} \quad \theta_2 = \frac{\lambda_u - a_h}{b_h} = \frac{c_h}{\lambda_u - d_h}. \tag{21}$$

Thus, the suitable initial conditions for the unstable manifold will be:

$$\begin{aligned} x &= x_0 + \varepsilon_0\lambda_u^\delta, \\ y &= 0, \\ \dot{x} &= \dot{x}_0 + \varepsilon_0\lambda_u^\delta\theta_2, \\ \dot{y} &= [2W(x_0, 0) - \dot{x}_0^2 - \Gamma_0]^{1/2}, \end{aligned} \tag{22}$$

while the density of the unstable manifold in the phase space is succeeded by obtaining several values of δ in the interval $[0, 1)$.

In Fig. 3a we show the unstable manifold of a Lyapunov periodic orbit around L_1, up to its third cut with the x-axis, for the Jacobi constant value $\Gamma = 4.967021$. In Fig. 3b the phase portraits, in the $(x, \dot{x})$ plane, of the unstable manifolds (first cuts with the x-axis) for various Lyapunov orbits are presented. Each curve corresponds to a specific periodic orbit and the points of the curve located on the x-axis, i.e. with $\dot{x} = 0$, denote the existence of homoclinic orbits. The tangential curve (dashed line) marks the onset of homoclinic orbits. Note that the “center” of the innermost elliptic curve represents the limiting case of the unstable manifolds of the Lyapunov orbits, i.e. the one-dimensional unstable manifold of the equilibrium point L_1. In

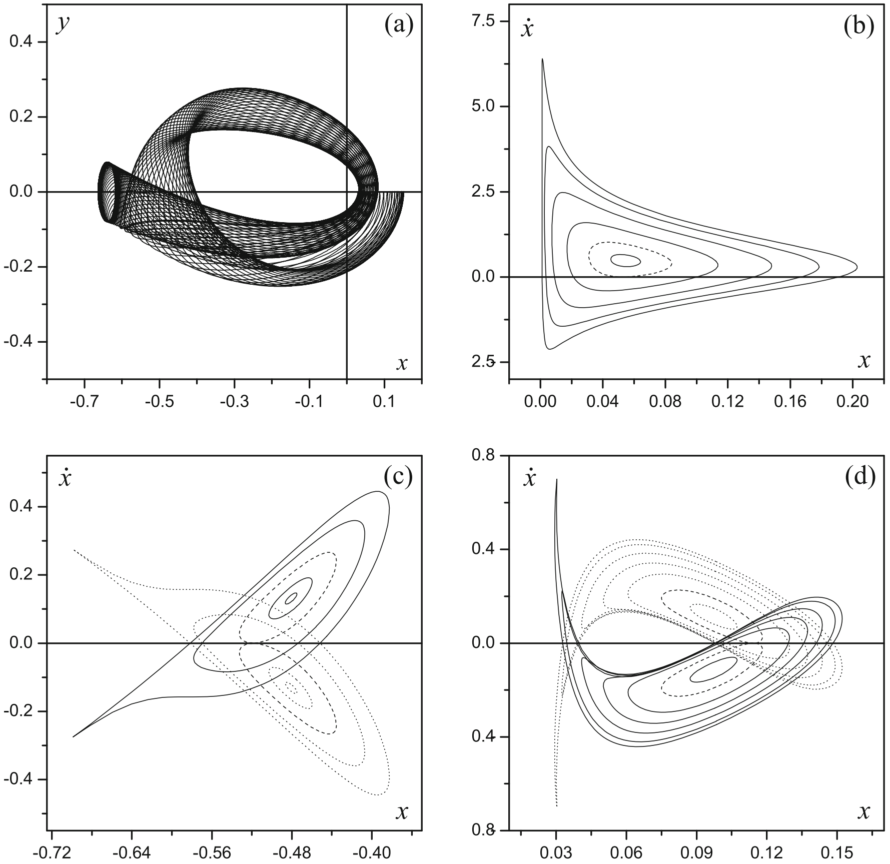


Fig. 3 (**a**) The unstable manifold of a Lyapunov orbit around L_1. The phase portraits of the unstable manifolds of various Lyapunov orbits around L_1 for the (**b**) $n = 1$ cuts, (**c**) $n = 2$ cuts and (**d**) $n = 3$ cuts

Fig. 3c and d we show the phase portraits of the second and third cuts of the unstable manifolds for various Lyapunov orbits together with the portraits of the corresponding stable manifolds (dotted lines). In these two cases we observe that transversality of the stable with the unstable manifolds occurs also outside the x-axis indicating the existence of non-symmetric homoclinic orbits. Similarly, in Fig. 4a the unstable manifold of a Lyapunov periodic orbit around the positive equilibrium point L_2, up to its third cut with the x-axis, for $\Gamma = 3.591264$, is shown. In Fig. 4b, c and d we show the phase portraits of the first, second and third cuts of the unstable manifolds for various Lyapunov orbits around L_2. As we see, in the case of the Lyapunov family emanating from L_2, homoclinic orbits exist only for the $n = 1$ cuts. In Fig. 5 several homoclinic orbits to Lyapunov periodic orbits are shown.

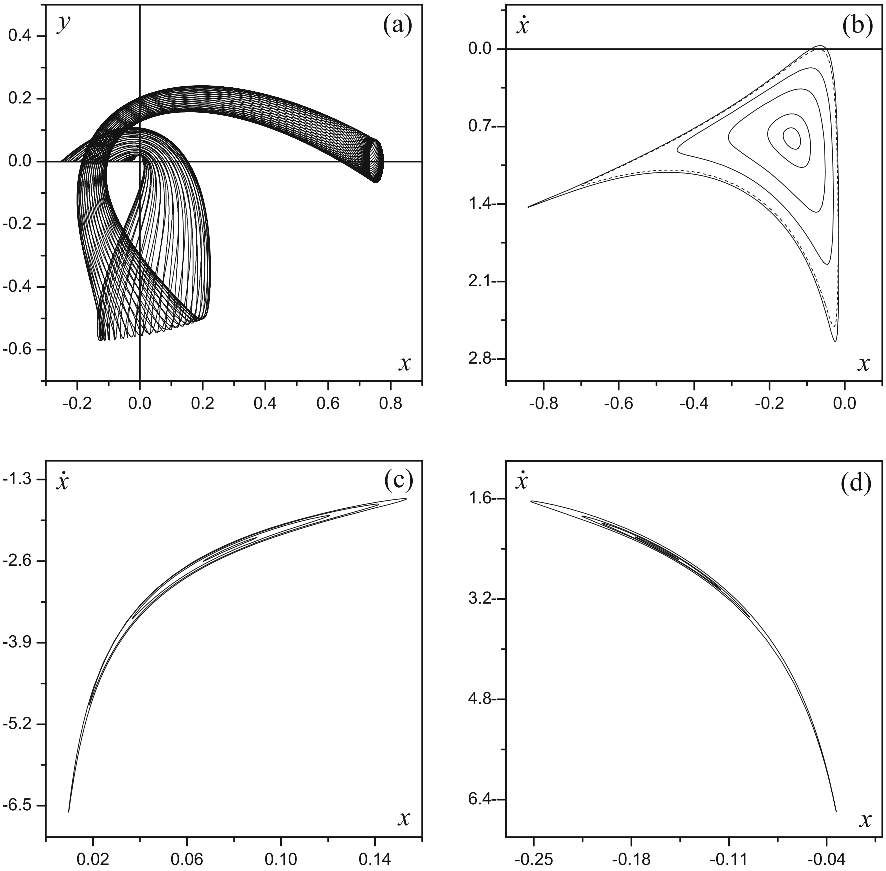


Fig. 4 (**a**) The unstable manifold of a Lyapunov orbit around L_2. The phase portraits of the unstable manifolds of various Lyapunov orbits around L_2 for the (**b**) $n = 1$ cuts, (**c**) $n = 2$ cuts and (**d**) $n = 3$ cuts

5 Conclusions

Asymptotic orbits at collinear equilibrium points were studied in a special version of the well-known Hill problem where the larger primary, representing the Sun, is a source of radiation. In particular, orbits which start asymptotically from a collinear equilibrium point and terminate at the same point asymptotically were determined. This kind of motion is a special case of non-escape motion in Celestial Mechanics since the moving particle (e.g. a natural or an artificial satellite) is trapped in the vicinity of an unstable equilibrium point for infinite time. In addition, asymptotic orbits at the highly unstable periodic orbits emanating from the collinear equilibrium points (the so-called Lyapunov orbits) were also computed. The corresponding

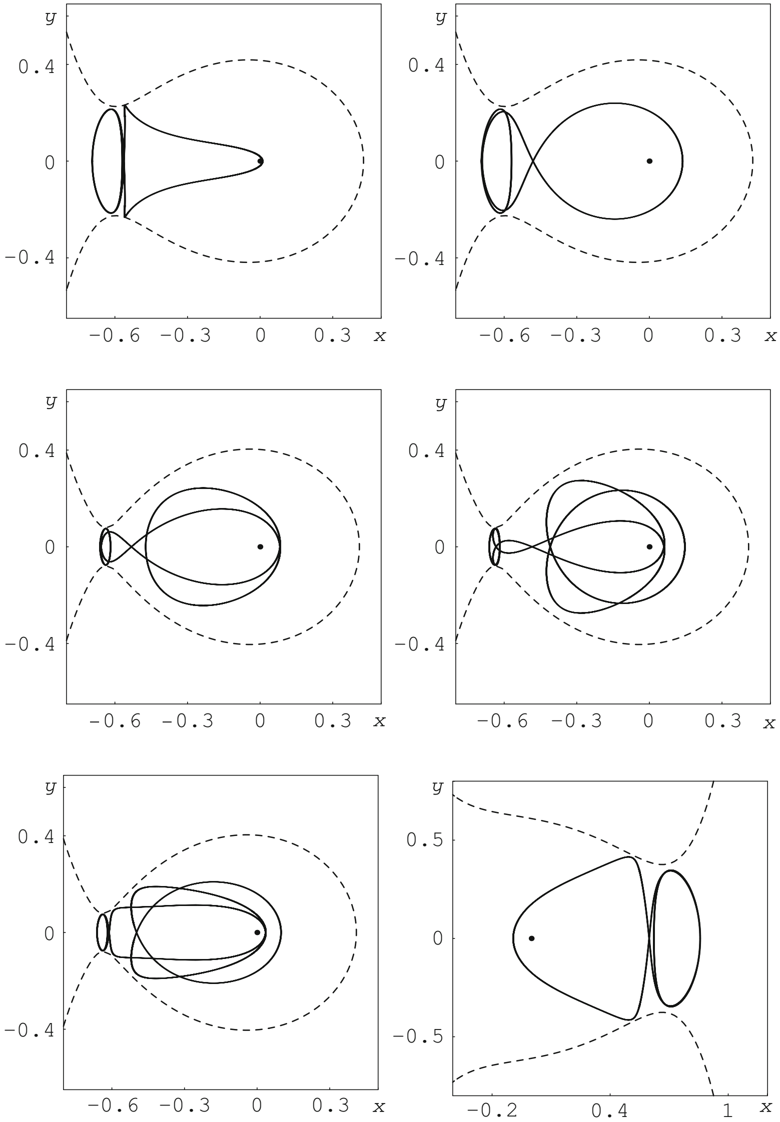


Fig. 5 Homoclinic orbits to Lyapunov orbits around L_1 and L_2 (last orbit)

asymptotic orbits are also of special importance from a practical point of view (e.g. for the design of low energy transfer) since they are used to connect the initial and final orbit of a space mission design which belong to their unstable and stable manifold, respectively. In both cases of the aforementioned homoclinic orbits, semi-analytical solutions were obtained which were used for the numerical computation of the corresponding unstable manifolds. Especially, for the latter case of asymptotic orbits, we additionally used appropriate Poincaré surface of sections in order to detect transversality of the stable and unstable manifolds and locate the corresponding homoclinic orbits.

Natural extensions of the present work would be to study the way where the homoclinic orbits to the Lyapunov periodic orbits vary with respect to the unique parameter of the problem Q_1 as well as to consider also possible heteroclinic connections at both Lyapunov periodic orbits and collinear equilibrium points.

Acknowledgement The authors would like to thank Prof. V.V. Markellos for valuable discussions during this work.

References

1. C.C. Conley, Low energy transit orbits in the restricted three-body problem. SIAM J. Appl. Math. **16**, 732–746 (1968)
2. M.K. Das, P. Narang, S. Mahajan, M. Yuasa, Effect of radiation on the stability of equilibrium points in the binary stellar systems: RW–Monocerotis, Kr*ü*ger 60. Astrophys. Space Sci. **314**, 261–274 (2008)
3. K.E. Davis, R.L. Anderson, D.J. Scheeres, G.H. Born, The use of invariant manifolds for transfers between unstable periodic orbits of different energies. Celest. Mech. Dyn. Astron. **107**, 471–485 (2010)
4. A. Deprit, J. Henrard, Symmetric doubly asymptotic orbits in the restricted three-body problem. Astron. J. **70**, 271–274 (1965)
5. A. Deprit, J. Henrard, Construction of orbits asymptotic to a periodic orbit. Astron. J. **74**, 308–316 (1969)
6. G. Gómez, J.M. Mondelo, The dynamics around the collinear equilibrium points of the RTBP. Phys. D **157**(4), 283–321 (2001)
7. G. Gómez, M. Marcote, J.M. Mondelo, The invariant manifold structure of the spatial Hill's problem. Dyn. Syst. Int. J. **20**, 115–147 (2005)
8. M. Hénon, Exploration Numérique du Problème Restreint II – Masses Égales, Stabilité des Orbites Périodiques. Ann. Astrophys. **28**, 992–1007 (1965)
9. K.C. Howell, D.C. Davis, A.F. Haapala, Application of periapse maps for the design of trajectories near the smaller primary in multi–body regimes. Math. Probl. Eng. **2012**, Article ID 351759, 22 pp. (2012)
10. V.S. Kalantonis, C.N. Douskos, E.A.Perdios, Numerical determination of homoclinic and heteroclinic orbits at collinear equilibria in the restricted three-body problem with oblateness. Celest. Mech. Dyn. Astron. **94**, 135–153 (2006)
11. W.S. Koon, M.W. Lo, J.E. Marsden, S.D. Ross, Heteroclinic connections between periodic orbits and resonance transitions in celestial mechanics. Chaos **10**(2), 427–469 (2000)
12. A.L. Kunitsyn, A.T. Tureshbaev, On the collinear libration points in the photo-gravitational three-body problem. Celest. Mech. **35**, 105–112 (1985)
13. A.L. Kunitsyn, E.N. Polyakhova, The restricted photogravitational three-body problem: a modern state. Astron. Astrophys. Trans. **6**, 283–293 (1995)
14. J. Llibre, R. Martínez, C. Simó, Transversality of the invariant manifolds associated to the Lyapunov family of the periodic orbits near L_2 in the restricted three-body problem. J. Differ. Equ. **58**, 104–156 (1985)
15. V.V. Markellos, A.E. Roy, M.J. Velgakis, S.S. Kanavos, A photogravitational Hill problem and radiation effects on Hill stability of orbits. Astrophys. Space Sci. **271**, 293–301 (2000)

16. Z. Niedzielska, Nonlinear stability of the libration points in the photogravitational restricted three body problem. Celest. Mech. Dyn. Astron. **58**, 203–213 (1994)
17. K.E. Papadakis, Homoclinic and heteroclinic orbits in the photogravitational restricted three–body problem. Astrophys. Space Sci. **302**, 67–82 (2006)
18. N. Pathak, V.O. Thomas, Evolution of the f family orbits in the photo gravitational Sun-Saturn system with oblateness. Int. J. Astron. Astrophys. **6**, 254–271 (2016)
19. E.A. Perdios, V.V. Markellos, Symmetric doubly-asymptotic periodic orbits at collinear equilibria. Astrophys. Space Sci. **166**, 129–149 (1990)
20. D.W. Schuerman, Roche potentials including radiation effects. Astrophys. Space Sci. **19**, 351–358 (1972)
21. D.W. Schuerman, The restricted three–body problem including radiation pressure. Astrophys. J. **238**, 337–342 (1980)
22. C. Simó, T.J. Stuchi, Central stable/unstable manifolds and the destruction of KAM tori in the planar Hill problem. Phys. D **140**, 1–32 (2000)
23. J.F.L. Simmons, A.J.C. McDonald, J.C. Brown, The restricted 3-body problem with radiation pressure. Celest. Mech. **35**, 145–187 (1985)
24. P. Verrier, T. Waters, J.Sieber, Evolution of the L_1 halo family in the radial solar sail circular restricted three–body problem. Celest. Mech. Dyn. Astron. **120**, 373–400 (2014)
25. B.F. Villac, D.J. Scheeres, Escaping trajectories in the Hill three–body problem and applications. J. Guid. Control Dyn. **26**, 224–232 (2003)
26. C. Zagouras, V.V. Markellos, Three-dimensional periodic solutions around equilibrium points in Hill's problem. Celest. Mech. **35**, 257–267 (1985)

Computations for Minors of Weighing Matrices with Application to the Growth Problem

Christos D. Kravvaritis

1 Introduction

Weighing matrices constitute a special type of combinatorial matrices that have attracted scientific interest for many years. After giving the necessary preliminary material, we present and highlight the role of weighing matrices in two research fields. We survey the effort made for calculating principal minors of weighing matrices. Furthermore, the significance of such a study and its contribution to a well known problem in numerical analysis, the growth problem, are explored.

As a matter of fact, the renowned Hadamard matrices are a special case of weighing matrices $W(n, n-k)$ for $k = 0$. It can be proved that n is a multiple of 4. Moreover, weighing matrices constitute a special class of a broader class of real orthogonal matrices called orthogonal designs. There exist also the more generalized complex variants of the three aforementioned mathematical notions. For the shake of brevity and simplicity, only the real counterparts will be exposed and analyzed here.

Especially in the area of Numerical Analysis it is worth mentioning that weighing matrices are the only matrices known so far that exhibit growth factor close to their order, or generally moderate growth [9, 12]. This is associated with the well known growth conjecture [8]. Several strategies have been developed therefor even for specific small, non general orders, which is still a challenging issue due to the high computational costs involved in the required exhaustive searches [14, 46, 51].

C. D. Kravvaritis (✉)
Department of Mathematics, University of Athens, Athens, Greece
e-mail: ckrav@math.uoa.gr

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_19

The numerous real-life applications of weighing matrices evoke interest for the study of their structure and of properties, cf., e.g., [3, 20, 22, 39, 54, 58, 61, 68]. Besides, the consideration of their mathematical features has its own intrinsic theoretical beauty.

1.1 Notations

Throughout this work the entries of an $(0, 1, -1)$ matrix will be denoted by $(0, 1, -)$. I_n and J_n stand for the identity matrix of order n and the matrix with ones of order n, respectively. We write $A(j)$ for the absolute value of the determinant of the $j \times j$ principal submatrix in the upper left corner of the matrix A, i.e. the magnitude of the $j \times j$ leading principal minor. Similarly, $A[j]$ denotes the magnitude of the determinant of the lower right $j \times j$ principal submatrix of A.

We write $J_{b_1,b_2,\cdots,b_z}$ for the all ones matrix with diagonal blocks of sizes $b_1 \times b_1$, $b_2 \times b_2$ $\cdots$ $b_z \times b_z$, and $a_{ij}J_{b_1,b_2,\cdots,b_z}$ for the matrix, for which the elements of the block with corners
$(i + b_1 + b_2 + \cdots + b_{j-1}, i + b_1 + b_2 + \cdots + b_{i-1})$,
$(i + b_1 + b_2 + \cdots + b_{j-1}, b_1 + b_2 + \cdots + b_i)$,
$(b_1 + b_2 + \cdots + b_j, i + b_1 + b_2 + \cdots + b_{i-1})$,
$(b_1 + b_2 + \cdots + b_j, b_1 + b_2 + \cdots + b_i)$
are the integers a_{ij}. We write $(k_i - a_{ii})I_{b_1,b_2,\cdots,b_z}$ for the matrix direct sum
$(k_1 - a_{11})I_{b_1} + (k_2 - a_{22})I_{b_2} + \cdots + (k_z - a_{zz})I_{b_z}$.

2 Weighing Matrices

Definition 1 A $(0, 1, -1)$ matrix $W = W(n, n-k)$, $k = 1, 2, \ldots$, of order n satisfying $W^T W = WW^T = (n-k)I_n$ is called a *weighing matrix of order n and weight n − k* or simply a *weighing matrix*.
A $W = W(n, k)$ for which $W^T = -W$ is called a *skew-weighing matrix*.
A $W(n, n)$, $n \equiv 0 \pmod 4$, is a *Hadamard matrix of order n*.
A $W = W(n, n-k)$ for which $W^T = -W, n \equiv 0 \pmod 4$, is called a *skew-weighing matrix*.

Example 1

$$W(7,4)=\begin{bmatrix}1&1&1&1&0&0&0\\1&-&0&0&1&1&0\\1&0&-&0&-&0&1\\1&0&0&-&0&-&-\\0&1&-&0&0&1&-\\0&1&0&-&1&0&1\\0&0&1&-&-&1&0\end{bmatrix},\quad W(10,8)=\begin{bmatrix}0&1&1&1&-&0&1&1&-&1\\-&0&1&1&1&1&0&1&1&-\\1&-&0&1&1&-&1&0&1&1\\1&1&-&0&1&1&-&1&0&1\\1&1&1&-&0&1&1&-&1&0\\0&1&-&1&1&0&1&-&-&-\\1&0&1&-&1&-&0&1&-&-\\1&1&0&1&-&-&-&0&1&-\\-&1&1&0&1&-&-&-&0&1\\1&-&1&1&0&1&-&-&-&0\end{bmatrix},$$

$$W(12,5)=\begin{bmatrix}1&1&1&1&1&0&0&0&0&0&0&0\\1&1&-&-&0&1&0&0&0&0&0&0\\1&-&0&0&0&0&1&1&1&0&0&0\\1&-&0&0&0&0&-&-&0&1&0&0\\1&0&0&0&-&-&0&0&-&-&0&0\\0&1&0&0&-&-&0&0&1&1&0&0\\0&0&1&-&0&0&1&-&0&0&1&0\\0&0&1&-&0&0&-&1&0&0&0&1\\0&0&1&0&-&1&0&0&0&0&-&-\\0&0&0&1&-&1&0&0&0&0&1&1\\0&0&0&0&0&0&1&0&-&1&-&1\\0&0&0&0&0&0&0&1&-&1&1&-\end{bmatrix}.$$

Further examples of weighing matrices can be found in [20, 30, 54].

Definition 2 A $W = W(n, n-1)$, n even, with zeros on the diagonal satisfying $WW^T = (n-1)I_n$ is called a *conference matrix*. If $n \equiv 0 \pmod 4$, then $W = -W^T$ and W is called a *skew-conference matrix*. If $n \equiv 2 \pmod 4$, then $W = W^T$ and W is called a *symmetric conference matrix* and such a matrix cannot exist unless $n-1$ is the sum of two squares; thus they cannot exist for orders 22, 34, 58, 70, 78, 94.

Remark 1 Symmetric or antisymmetric conference matrices are also known in the literature as *C-matrices* [5].

Example 2

$$W(6,5)=\begin{bmatrix}0&1&1&1&1&1\\1&0&1&-&-&1\\1&1&0&1&-&-\\1&-&1&0&1&-\\1&-&-&1&0&1\\1&1&-&-&1&0\end{bmatrix}.$$

Remark 2 Sometimes a conference matrix of order n is just defined as a weighing matrix $W(n, n-1)$. For example, the matrix

$$W(4,3) = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & - & - & 1 \\ 1 & - & 0 & - \\ 1 & 1 & - & 0 \end{bmatrix}$$

satisfies this relaxed definition, but not the more strict one requiring the zero elements to be on the diagonal.

For more details and constructions of weighing matrices the reader can refer the book by Geramita and Seberry [20, 28].

Definition 3 If a Hadamard matrix H of order n can be written as $H = I + S$ where $S^T = -S$ then H is called *skew–Hadamard*. S is also a conference matrix; we call it a *skew conference matrix*.

Definition 4 Two matrices are said to be *Hadamard equivalent* or *H-equivalent* if one can be obtained from the other by a sequence of the operations:

1. interchange any pairs of rows and/or columns;
2. multiply any rows and/or columns through by -1.

Lemma 1 ([23]) *Every weighing matrix $W(n, n-1)$, with n even, is H-equivalent to a conference matrix.*

Definition 5 The entries of the first row and column of a weighing matrix can always be $+1$ (*normalized form*). Indeed, this can be achieved easily by the H-equivalence operation of multiplying columns and/or rows with -1 and leaves unaffected the properties of the matrix.

So, without loss of generality, one may assume that weighing matrices can be considered in normalized form. Weighing matrices belong to a wider class of orthogonal matrices called orthogonal designs [20].

Definition 6 An *orthogonal design* of order n and type $(s_1, s_2, \ldots, s_t)$, $s_i > 0$ integers, denoted as $OD(n; s_1, s_2, \ldots, s_t)$, on the commuting variables $x_1, x_2, \ldots, x_t$ is an $n \times n$ matrix O with entries from the set $\{0, \pm x_1, \pm x_2, \ldots, \pm x_t\}$ such that

$$OO^T = O^T O = \left(\sum_{i=1}^{t} s_i x_i^2 \right) I_n.$$

Two important properties, which follow from this definition, are that every two distinct rows or columns of O are orthogonal and each row and column of O has precisely s_i entries of the type $\pm x_i$, $i = 1, \ldots, t$. If $\sum_{i=1}^{t} s_i = n$, the OD is called full. For instance, a Hadamard matrix of order n is a full $OD(n; n)$ with entries $\{\pm 1\}$.

Example 3 As examples for ODs are given the following $OD(2; 1, 1)$, $OD(4; 2, 2)$, $OD(4; 2, 2)$, $OD(4; 1, 1, 1)$ and $OD(4; 1, 1, 1, 1)$, respectively.

$$\begin{bmatrix} x & y \\ y & -x \end{bmatrix}, \begin{bmatrix} a & b & a & b \\ -b & a & b & -a \\ -a & -b & a & b \\ -b & a & -b & a \end{bmatrix}, \begin{bmatrix} x & x & y & y \\ x & -x & y & -y \\ y & y & -x & -x \\ y & -y & -x & x \end{bmatrix}, \begin{bmatrix} a & -b & -c & 0 \\ b & a & 0 & c \\ c & 0 & a & -b \\ 0 & -c & b & a \end{bmatrix}, \begin{bmatrix} x & y & z & w \\ -y & x & w & -z \\ -z & -w & x & y \\ -w & z & -y & x \end{bmatrix}.$$

Lemma 2 ([21, 22]) *The existence of an orthogonal design of order n and type* $(s_1, \ldots, s_t)$ *is equivalent to the existence of weighing matrices* $A_1, \ldots A_t$ *of order n, where* A_i *has weight* s_i *and the matrices* A_i, $i = 1, \ldots t$, *satisfy the matrix equation*

$$XY^T + YX^T = \mathbb{O}$$

in pairs.

2.1 Hadamard Matrices

The renowned Hadamard matrices are a special class of weighing matrices $W(n, n-k)$ for $k = 0$, i.e. a $W(n, n)$, $n \equiv 0 \pmod 4$, is a Hadamard matrix. Additionally, Hadamard matrices are associated with many real-life applications, conjectures and scientific research open problems [20, 29, 34, 55, 63, 69, 75].

Definition 7 A *Hadamard matrix* $H \equiv H_n$ of order n (denoted by H_n) has entries ± 1 and satisfies $HH^T = H^T H = nI_n$.

Example 4

$$H_1 = [1], \; H_2 = \begin{bmatrix} 1 & 1 \\ 1 & - \end{bmatrix}, \; H_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & - & 1 & - \\ 1 & 1 & - & - \\ 1 & - & - & 1 \end{bmatrix}.$$

In 1893 Hadamard [27] specified quadratic matrices of orders 12 and 20 with entries ±1 having all their rows and columns mutually orthogonal. These matrices satisfied the following famous Hadamard inequality.

Theorem 1 ([27],[10, p. 496] Hadamard's Inequality) *For any matrix* $A = (a_{ij})_{1 \le i,j \le n}$ *with entries on the unit disc*

$$|\det A| \le \prod_{j=1}^{n} \left(\sum_{i=1}^{n} |a_{ij}|^2 \right)^{\frac{1}{2}} = \prod_{j=1}^{n} \|a_j\|_2 \le n^{\frac{n}{2}},$$

where a_j *denotes the* j*-th column of* A*. The equalities hold if and only if* $|a_{ij}| = 1$ *for all* i, j *and the rows of* A *are mutually orthogonal.*

However, these matrices for every order being a power of 2 have been first discovered in 1867 by Sylvester [59]. The following well known result describes the possible order n of a Hadamard matrix.

Theorem 2 ([27]) *If H is an $n \times n$ Hadamard matrix and $n > 2$, then n is a multiple of 4.*

Theorem 2 is usually proved by describing the sign patterns in three rows of a Hadamard matrix like

$$\begin{array}{cccc} \overbrace{1\dots1}^{x} & \overbrace{1\dots1}^{y} & \overbrace{1\dots1}^{z} & \overbrace{1\dots1}^{w} \\ 1\dots1 & 1\dots1 & -\cdots- & -\cdots- \\ 1\dots1 & -\cdots- & 1\dots1 & -\cdots- \end{array},$$

where x, y, z, w denote the number of columns of the respective form. Another elegant proof for Theorem 2 is given in [10].

One more recent proof for Theorem 2 can be found in [6]. Theorem 2 does not assure the existence of Hadamard matrices for every n being a multiple of 4. However, it is conjectured whether there exists a Hadamard matrix of order n for every n being a multiple of 4. This conjecture seems very realistic but is not proved yet. The smallest order, for which a Hadamard matrix has not been found, is 668 [34]. Recently a Hadamard matrix of order 428 was found [38]. Other orders smaller than 1000, for which Hadamard matrices have not been found yet, are 716, 764, 892.

3 Gaussian Elimination and the Growth Problem

Consider a linear system of the form $A \cdot x = b$, where $A = [a_{ij}] \in \mathbb{R}^{n\times n}$ is nonsingular. Gaussian elimination (GE) [9, 24, 31, 32, 65] is the simplest way to solve such a system by hand, and also the standard method for solving it on computers. GE without pivoting fails if any of the pivots is zero and it behaves worse if any pivot becomes close to zero. In this case the method can be carried out to completion but it is totally unstable and the obtained results may be totally wrong, as it is already demonstrated in a famous example by Forsythe and Moler [16].

Therefore, a search for the element with maximum absolute value is performed. If the search is done in the respective column, then we have GE with partial pivoting, else if it is done in the respective lower right submatrix we have GE with complete pivoting. Let $A^{(k)} = [a_{ij}^{(k)}]$ denote the matrix obtained after the first k pivoting operations, so $A^{(n-1)}$ is the final upper triangular matrix. A diagonal entry of that final matrix will be called a *pivot*.

Traditionally, backward error analysis for GE is expressed in terms of the *growth factor*

$$g(n, A) = \frac{\max_{i,j,k} |a_{ij}^{(k)}|}{\max_{i,j} |a_{ij}|}, \tag{1}$$

which involves all the elements $a_{ij}^{(k)}$, $k = 1, 2, \ldots, n$ that occur during the elimination. The growth factor actually measures how large the entries become during the process of elimination.

In 1991 Gould reported on matrices that exhibited, in presence of roundoff error, growth larger than n [13, 25]. These matrices were created by simulating the process of GE as an appropriate optimization problem. Thus the first part of Cryer's conjecture was shown to be false. The second part of the conjecture concerning the growth factor of Hadamard matrices still remains an open problem.

The following classic theorem illustrates the accuracy of the computed solution.

Theorem 3 (Wilkinson [31, p. 165]) *Let $A \in \mathbb{R}^{n\times n}$ and suppose GE with partial pivoting produces a computed solution $\hat{x}$ to $A \cdot x = b$. Then there exists a matrix ΔA and a constant c_{3n} such that*

$$(A + \Delta A)\hat{x} = b, \quad \|\Delta A\|_\infty \le c_{3n} n^2 g(n, A)\|A\|_\infty.$$

It is clear that the stability of GE depends on the growth factor. If $g(n, A)$ is of order 1, not much growth has taken place, and the elimination process is stable. If $g(n, A)$ is bigger than this, we must expect instability. If GE can be unstable, why is it so famous and so popular? The answer seems to be that although some matrices cause instability, these represent such an extraordinarily small proportion of the set of all matrices that they "never" arise in practice simply for statistical reasons. This explanation gives rise to a statistical approach to the growth factor and motivates the study of its behavior for random matrices [12, 66]. In practice the growth factor is usually of order 10 and therefore most numerical analysts regard the occurrence of serious element growth in GE with partial pivoting as highly unlikely in practice. So, the method can be used with confidence and constitutes a stable algorithm in practice [64].

The determination of $g(n, A)$ in general remains a mystery. Regarding the possible magnitude of the growth factor for partial pivoting, it is easy to show that $g(n, A) \le 2^{n-1}$. Wilkinson in [71, p. 289] and [72, p. 212] has reported of special form matrices attaining this upper bound. Higham and Higham characterize all matrices attaining this upper bound in the following theorem.

Theorem 4 ([33]) *All real $n \times n$ matrices A, for which $g(n, A) = 2^{n-1}$ for partial pivoting, are of the form*

$$A = DM \begin{bmatrix} T & \vdots & \theta d \\ 0 & \vdots & \end{bmatrix},$$

where $D = diag(\pm 1)$, M is unit lower triangular with $m_{ij} = -1$ for $i > j$, T is an arbitrary nonsingular upper triangular matrix of order $n - 1$, $d = (1, 2, 4, \ldots, 2^{n-1})^T$, and θ is a scalar such that $\theta := |a_{1n}| = \max_{i,j} |a_{ij}|$.

Table 1 Values of $f(n)$ for Wilkinson's bound

n	10	20	50	100	200	1000
$f(n)$	19	67	530	3300	26,000	79,00,000

Although the growth factor can be as large as 2^{n-1} for an $n \times n$ matrix, the occurrence of a growth factor even as large as n is rare. However, later than the early 1990s some applications with large growth factors have been published [15, 17, 74]. Efforts have been made to explain this phenomenon with some success [33, 65, 66], yet the matter is far from completely understood.

For complete pivoting, Wilkinson has showed in [71, pp. 282–285] that

$$g(n, A) \leq [n\, 2\, 3^{1/2} \ldots n^{1/(n-1)}]^{1/2} \equiv f(n) \sim cn^{1/2} n^{\frac{1}{4}\log n}$$

and that this bound is not attainable. The bound is a much more slowly growing function than 2^{n-1}, but it can still be quite large, cf. Table 1.

Definition 8 Matrices with the property that no exchanges are actually needed during Gaussian Elimination (GE) with complete pivoting (GECP) are called *completely pivoted (CP)* or *feasible*.

Equivalently, a matrix is CP if its rows and columns have been permuted so that GE without pivoting satisfies the requirements for complete pivoting, hence the maximum elements on each elimination step appear on the diagonal position.

For a CP matrix A we have

$$g(n, A) = \frac{\max\{p_1, p_2, \ldots, p_n\}}{|a_{11}|}, \tag{2}$$

where $p_1, p_2, \ldots, p_n$ are the pivots of A. The study of the values appearing for $g(n, A)$ and the specification of pivot patterns are referred to as *the growth problem*.

Cryer [8] defined

$$g(n) = \sup\{\, g(n, A) \mid A \in \mathbb{R}^{n \times n}\, , \mathrm{CP}\}.$$

The function $g(n)$ plays a role in the analysis of roundoff errors in GE ([16, p. 103] and [72, p. 213]). Wilkinson in [72, p. 213] noted that there were no known examples of matrices for which $g(n, A) > n$. The problem of determining $g(n)$ for various values of n is called the *growth problem*. The following results are known:

- $g(2) = 2$ (trivial)
- $g(3) = 2\frac{1}{4}$ [7, 8, 10, 62]
- $g(4) = 4$ [7, 8, 62]
- $g(5) < 5.005$ [7]

4 The Importance of Determinant Calculations

In this expository work, the principal techniques used in the majority of the works for calculating pivot structures of weighing matrices have been described. The core strategy is to evaluate principal minors, i.e. determinants, efficiently. The idea takes advantage of Lemma 3, which constitutes a powerful tool for this research, since it offers a possibility for calculating pivots in connection with minors.

So, it is obvious that the calculation of minors is important in order to study pivot structures, and moreover the growth problem for CP weighing matrices.

Lemma 3 ([8, 10, 14, 19, 51]) *Let A be a CP matrix.*

(i) *The magnitude of the pivots appearing after application of GE operations on A is given by*

$$p_j = \frac{A(j)}{A(j-1)}, \quad j = 1, 2, \ldots, n, \quad A(0) = 1. \tag{3}$$

(ii) *The maximum $j \times j$ leading principal minor of A, when the first $j - 1$ rows and columns are fixed, is $A(j)$.*

The second part of Lemma 3 assures that the maximum $j \times j$ minor appears in the upper left $j \times j$ corner of A. So, if the existence of a matrix with maximal determinant is proved for a CP weighing matrix, we can indeed assume that it always appears in the upper left corner.

Furthermore, the evaluation of principal minors of a matrix can be useful for calculating pivots by taking advantage of the following results. Lemma 4 associates leading principal minors of the upper left corners with the respective ones of the lower right corners of a generic orthogonal matrix. As a result, Corollary 1 offers the possibility of specifying pivots from the end of the pivot pattern in terms of principal minors of lower right corners of the matrix.

Lemma 4 ([10, Proposition 5.2]) *Let A be an invertible $n \times n$ matrix satisfying $AA^T = cI_n$. If $1 \le k < n$, then*

$$A(k) = c^{k-\frac{n}{2}} A[n-k].$$

Corollary 1 *If A satisfies $AA^T = cI_n$, then the* j*th pivot from the end is*

$$p_{n+1-j} = \frac{cA[j-1]}{A[j]}.$$

All in all, it is important to calculate principal minors of weighing matrices because one can compute pivots with their aid based on 3. Furthermore, the respective growth factors can be established via (2).

Generally, numerous applications in the mathematical sciences require determinants and principal minors. For example, note worthy are the detection of P matrices [26], self validating algorithms and interval matrix analysis.

The derivation of analytical formulas is useful whenever this is possible. In general it is very difficult to prove analytical formulas for the determinant of a given matrix or for its minors. However, when we have matrices with special structure, such as Hadamard matrices, Vandermonde or Hankel matrices, analytical formulae can be demonstrated.

Definition 9 Let A be an $n \times n$ matrix. The determinant of the $k \times k$ submatrix of A formed by deleting $n - k$ rows of A, and the corresponding $n - k$ columns of A, is called *principal minor* of A.

If the resulting matrix is a quadratic upper-left part of the larger matrix A (i.e., it consists of matrix elements in rows and columns from 1 to k), then the principal minor is called a *leading principal minor* (of order k) of A and is denoted by $A(k)$.

Specifically, with regard to the results presented here, Eq. (3) provides a powerful tool for computing pivots in terms of leading principal minors. Hence, the knowledge and appropriate combination of values of minors provides pivots and pivot sequences. The computations of minors offer an essential assistance in view of the specification of pivot patterns because the direct approach for evaluating all principle minors of a matrix of order n by applying LU factorizations entails high complexity of $O(2^n n^3)$ [67]. In addition, the trivial implementation of the aforementioned task includes also the exhaustive search for all possible weighing matrices of the same order before applying LU to each one of them. This cannot be achieved within a sensible and realistic time period. Hence, more sophisticated techniques should be elaborated to that end.

5 The First Four Pivots

Since pivots are strictly connected with minors (cf. Eq. (3)), we start our study with an effort of computing principal minors of skew and symmetric conference matrices. The following lemma specifies the possible values of determinants of small order. The results for orders 6 and 7 are new.

Lemma 5 ([50]) *All possible and the maximum determinant of all $n \times n$ matrices with elements ± 1 or 0, where there is at most one zero in each row and column are*

Considering the fact that the submatrices $[1]$ and $\left[\begin{smallmatrix}1 & 1\\ 1 & -1\end{smallmatrix}\right]$ can always occur in the upper lefthand corner of a CP skew and symmetric conference matrix of order $n \geq 6$, the required pivots are computed via (3).

Lemma 6 ([50]) *Let W be a CP skew and symmetric matrix, of order $n \geq 6$. Then if GE is performed on W the first two pivots are 1, and 2.*

Order	Maximum determinant	Possible determinantal values
2×2	2	0, 1, 2
3×3	4	0, 1, 2, 3, 4
4×4	16	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 16
5×5	48	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,
		19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 32, 36, 40, 48
6×6	160	160, 144, 136, 132, 130, 128, 120, 112, 108, 106, 105, 104,
		102, 100, ...
7×7	528	528, 504, 480, 468, 456, 444, 432, 420, 408, 396, 384, 372,
		366, 360, 354, 348, 342, 336, 330, 324, ...

Working in a similar fashion like before and extending the idea for higher orders, the following results determine the existence of specific submatrices in weighing matrices of small orders.

Lemma 7 ([50]) *H-equivalence operations can be used to ensure the following submatrices always occur in the upper lefthand corner of a $W(8, 7)$ and a $W(10, 9)$:*

$$B_1 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & - & 1 \\ 1 & 1 & - \end{bmatrix} \text{ or } B_2 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & - & 0 \\ 1 & 1 & - \end{bmatrix},$$

and

$$A_1 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & - & 1 & - \\ 1 & - & - & 1 \\ 1 & 1 & - & - \end{bmatrix} \text{ or } A_2 = \begin{bmatrix} 1 & 1 & 0 & - \\ 1 & - & - & - \\ 1 & - & 1 & 1 \\ 1 & 1 & - & 1 \end{bmatrix}.$$

This result can be generalized for every skew and symmetric matrix $W(n, n - 1)$. The proof is carried out by performing appropriate combinatorial techniques regarding the possible form of the first three and four rows of a $W(n, n - 1)$ for various cases of n and the existence of the respective submatrices there.

Lemma 8 ([50]) *H-equivalence operations can be used to ensure the submatrices B_1, B_2, A_1 and A_2 of Lemma 7 always occur in a skew and symmetric $W(n, n-1)$.*

Combining appropriately the results of Lemmas 8 and 3 yields the third pivot of a skew and symmetric conference matrix.

Proposition 1 ([50]) *Let W be a CP skew and symmetric conference matrix, of order $n \geq 12$ then if GE is performed on W the third pivot is 2.*

Proposition 2 presents the fourth pivot and is proved by extending the 3×3 matrices B_1 and B_2 to all possible 4×4 matrices with entries 0, ± 1 and afterwards applying appropriately Lemmas 8, 3 and Proposition 1.

Proposition 2 ([50]) *Let W be a CP skew and symmetric conference matrix, of order $n \geq 12$ then if GE is performed on W the fourth pivot is 3 or 4.*

Next, one should try to extend the 4×4 matrices A_1 and A_2 to all possible 5×5 matrices with elements $0, \ \pm 1$. It is interesting to specify the 5×5 matrices that contain the matrices A_1 or A_2 and also have the maximum possible values of the determinant (because of the CP property), which for the 5×5 matrices with specific requirements are given in Lemma 5. However, it becomes obvious that the computational time for all these exhaustive trials rises together with the rising of the order of the matrices and the algorithms cannot be implemented fast enough. Hence, more sophisticated strategies should be employed.

6 Extension of Specific Matrices with Elements 0, ±1 to $W(n, n-1)$ Matrices

Algorithm for extending a $k \times k$ matrix with elements $0, \pm 1$ to $W(n, n-1)$
Let a $k \times k$ matrix $A = [\underline{r}_1, \underline{r}_2, \ldots, \underline{r}_k]^T$. The following algorithm specifies its extension, if it exists, to a $W(n, n-1)$.

Algorithm Extend [50]

```
Step 1
read the $k \times k$ matrix $A$
Step 2
complete the first row of the matrix without loss of generality: it has exactly one 0
complete the first column of the matrix without loss of generality: it has exactly one 0
Step 3
complete (almost) the second row of the matrix without loss of generality:
                $\underline{r}_2 \cdot \underline{r}_1^T = 0$
                every row and column has exactly one zero
complete (almost) the second column of the matrix without loss of generality:
                it is orthogonal to the first column
                every row and column has exactly one zero
Step 4
Procedure Extend Rows
find all possible entries $a_{3,k+1}, a_{3,k+2}, \ldots, a_{3,n}$:
    $\underline{r}_3 \cdot \underline{r}_1^T = 0$ and $\underline{r}_3 \cdot \underline{r}_2^T = 0$
    every row and column has exactly one zero
store the results in a new matrix $B_3$ whose rows are all the possible entries
for $i = 4, \ldots, k$
  for every possible extension of the rows $\underline{r}_j, \quad j = 3, \ldots, i-1$
```

```
            find all possible entries $a_{i,k+1}, a_{i,k+2}, \dots, a_{i,n}$:
                  $\underline{r}_i$ is orthogonal with all the previous rows
                  every row and column has exactly one zero
            store the results in a new matrix $B_i$ whose rows are all the possible entries
      end
end
extend the $k$-th row of $A$ with the first row of $B_k$
extend the $k-1, \dots, 2$ rows of $A$ with the corresponding rows of
            the appropriate matrices $B_i$, $i = k-1, \dots, 3$
end {of Procedure Extend Rows}
Step 5
extend columns 3 to $k$ following a similar procedure as the one used to the rows.
Step 6
for $i = k+1, \dots, n$
      find all possible entries $a_{i,k+1}, a_{i,k+2}, \dots, a_{i,n}$:
            $\underline{r}_i$ is orthogonal with all the previous rows
            every row and column has exactly one zero
end
complete rows $k+1$ to $n$.
if columns $k+1$ to $n$ are orthogonal with all the previous columns
      $A$ is extended to $W(n, n-1)$.
```

Remark 3 In Step 3 by writing "complete almost" we mean that the second row can be completed in at most two ways up to permutation of columns. If the first row in the $k \times k$ part of the matrix contains a zero, then we complete the second row in a unique way without loss of generality. If the first row in the $k \times k$ part of the matrix doesn't contain a zero, then we complete the second row in two ways by setting the element below the 0 of the first row to 1 or -1, respectively. The same is done with the columns.

Implementation of the Algorithm Extend

We apply the algorithm for k=5, n=10 .

Steps of the algorithm

1. We start with

$$A = \begin{bmatrix} 1 & - & 0 & 1 & 1 \\ - & - & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & - \\ - & 1 & - & 1 & 1 \\ 1 & - & - & 0 & - \end{bmatrix}.$$

2. The first row and column is completed, without loss of generality, so that the property of a $W(10, 9)$ having exactly one zero in each row and column is preserved. The software package fills with zeros the rest of the entries of the required 10×10 matrix;

$$A=\begin{bmatrix}
1 & - & 0 & 1 & 1 & 1 & - & - & 1 & 1\\
- & - & 1 & 1 & 0 & 0 & & \cdots & & 0\\
1 & 1 & 1 & 1 & - & \vdots & & & & \vdots\\
- & 1 & - & 1 & 1 & & & & & \\
1 & - & - & 0 & - & 0 & & \cdots & & 0\\
- & 0 & \cdots & & & & \ddots & & & 0\\
0 & \vdots & & & & & & & & \vdots\\
- & & & & & & & & & \\
1 & & & & & & & & & \\
1 & 0 & \cdots & & & & & & & 0
\end{bmatrix}.$$

3. As before, the algorithm completes the second row in a unique way and the second column in two ways, because the element a beside the 0 of the first column below can take both values ±1;

$$A=\begin{bmatrix}
1 & - & 0 & 1 & 1 & 1 & - & - & 1 & 1\\
- & - & 1 & 1 & 0 & - & 1 & - & 1 & -\\
1 & 1 & 1 & 1 & - & 0 & & \cdots & & 0\\
- & 1 & - & 1 & 1 & 0 & & \cdots & & 0\\
1 & - & - & 0 & - & 0 & & \cdots & & 0\\
- & - & 0 & \cdots & & & & & & 0\\
0 & a & \vdots & & & \ddots & & & & \vdots\\
- & 1 & & & & & & & & \\
1 & 0 & & & & & & & & \\
1 & 1 & 0 & \cdots & & & & & & 0
\end{bmatrix}.$$

4. The algorithm takes as input this matrix A and finds all possible completions for rows 3–5 (columns 6–10), so that every row has exactly one zero, every column has at most one zero and the inner product of every two distinct rows is zero. If many ways have been found to complete rows 3–5, the algorithm keeps as a result the first solution found;

$$A=\begin{bmatrix}
1 & - & 0 & 1 & 1 & 1 & - & - & 1 & 1\\
- & - & 1 & 1 & 0 & - & 1 & - & 1 & -\\
1 & 1 & 1 & 1 & - & 1 & 0 & 1 & 1 & -\\
- & 1 & - & 1 & 1 & - & - & 1 & 1 & 0\\
1 & - & - & 0 & - & - & 1 & 1 & 1 & 1\\
- & - & 0 & \cdots & & & & & & 0\\
0 & a & \vdots & & & \ddots & & & & \vdots\\
- & 1 & & & & & & & & \\
1 & 0 & & & & & & & & \\
1 & 1 & 0 & \cdots & & & & & & 0
\end{bmatrix}.$$

5. The algorithm finds all possible completions for columns 3–5 (rows 6–10) in the same way it has done with the rows 3–5;

$$A = \begin{bmatrix} 1 & - & 0 & 1 & 1 & 1 & - & - & 1 & 1 \\ - & - & 1 & 1 & 0 & - & 1 & - & 1 & - \\ 1 & 1 & 1 & 1 & - & 1 & 0 & 1 & 1 & - \\ - & 1 & - & 1 & 1 & - & - & 1 & 1 & 0 \\ 1 & - & - & 0 & - & - & 1 & 1 & 1 & 1 \\ - & - & - & - & - & 0 & \cdots & & & 0 \\ 0 & - & - & 1 & 1 & \vdots & \ddots & & & \vdots \\ - & 1 & - & 1 & - & & & & & \\ 1 & 0 & - & 1 & - & & & & & \\ 1 & 1 & - & - & 1 & 0 & & & & 0 \end{bmatrix}.$$

6. The algorithm tries to complete,if possible, the rows 6–10 (columns 6–10) in the same way as before;

$$A = \begin{bmatrix} 1 & - & 0 & 1 & 1 & 1 & - & - & 1 & 1 \\ - & - & 1 & 1 & 0 & - & 1 & - & 1 & - \\ 1 & 1 & 1 & 1 & - & 1 & 0 & 1 & 1 & - \\ - & 1 & - & 1 & 1 & - & - & 1 & 1 & 0 \\ 1 & - & - & 0 & - & - & 1 & 1 & 1 & 1 \\ - & - & - & - & - & 1 & - & 0 & 1 & - \\ 0 & - & - & 1 & 1 & 1 & 1 & 1 & - & - \\ - & 1 & - & 1 & - & 1 & 1 & - & 0 & 1 \\ 1 & 0 & - & 1 & - & - & - & - & - & - \\ 1 & 1 & - & - & 1 & 0 & 1 & - & 1 & - \end{bmatrix}.$$

7. Finally, if matrix A could be extended, the algorithm gives the completed matrix $W(10, 9)$ and verifies whether the relationship $AA^T = 9I_{10}$ is valid.

Applying algorithm Extend subsequently for the appropriate matrices that occur as possible extensions of the matrices of smaller orders, one can prove the following proposition.

Proposition 3 ([50]) *The leading principal minors of orders 5, 6 and 7 are given for the weighing matrices $W(8, 7)$ and $W(10, 9)$ as follows.*

- $W(5) = 28$ *for a* $W(8, 7)$
- $W(5) = 48,\ 36$ *or* 30 *for a* $W(10, 9)$
- $W(6) = 144$ *or* 108 *for a* $W(10, 9)$
- $W(7) = 432$ *or* 324 *for a* $W(10, 9)$

7 A New Algorithm Extending $(0, \pm 1)$ Matrices to $W(n, n-1)$

In the literature, cf., e.g., [14, 51], effort has been made to compute the pivot structures for weighing matrices even of small orders. Although it might seem trivial to compute the pivot pattern, and thus the growth factor, of a weighing matrix of small order, actually it is not. When searching for the growth factor of a weighing matrix of order n, one should form all possible $W(n, n-k)$ and additionally specify the growth factor of each one of them. This task involves a remarkably high complexity and computational cost even for small orders. Therefore, it is important to emphasize that it is intriguing, nontrivial and challenging to determine growth factors even for weighing matrices of small orders. An essential idea for computing pivots is to take advantage of relationship (3), as it was done in the previous works for small n.

For proceeding to the order 12, after the smaller ones have been presented previously [50], it is necessary to find a way to demonstrate that specific $k \times k$ matrices with known determinant can always exist in a $W(12, 11)$. The conception is to create an algorithm, which extends a $k \times k$ $(0, +, -)$ matrix to a $W(n, n-1)$, if possible. In [50] the Algorithm Extend was proposed for this purpose. This algorithm was applied to show the pivot structures of the unique weighing matrices $W(8, 7)$ and $W(10, 9)$. The application of this algorithm for the $W(12, 11)$ encounters difficulties due to the higher order, and therefore higher complexity. The algorithm can be developed in a more sophisticated way by using the notions of parallel processing and data structures. So it can be applied for the $W(12, 11)$. Here we illustrate an improved version of Algorithm Extend, Algorithm Extend 2, and an example of its application for $W(12, 11)$. Thus, by making use of Algorithm Extend 2, one can infer if a matrix can or can't be extended to a $W(12, 11)$.

Algorithm for extending a $k \times k$ $(0, 1, -)$ **matrix to** $W(n, n-1)$
For a $k \times k$ matrix $A = [\underline{r}_1, \underline{r}_2, \dots, \underline{r}_k]^T$ the following algorithm specifies its extension, if it exists, to a $W(n, n-1)$.

```
Algorithm Extend 2  [51]
Step 1
read the $k \times k$ matrix $A$
Step 2
complete the first row of the matrix, columns $k+1, \dots, n$, without loss of generality:
        it has exactly one 0
complete the first column of the matrix, rows $k+1, \dots, n$, without loss of generality:
        it has exactly one 0
Step 3
Extend Rows $(2, k)$
```

```
*Step 4*
**Extend Columns** $(2, k)$
*Step 5*
**Extend Rows** $(k+1, n)$
**if** columns $k+1$ to $n$ are orthogonal with all the previous columns
    $A$ is extended to $W(n, n-1)$.
**end** {of Algorithm Extend 2}

**Procedure Extend Rows** $(m, z)$
**find** all possible entries $a_{m,k+1}, a_{m,k+2}, \ldots, a_{m,n}$:
    $\underline{r}_m$ is orthogonal with all the previous rows
    every row and column has exactly one zero
**store** the results in a new matrix $B_m$ whose rows are all the possible rows $\underline{r}_m$
**for** $i = m+1, \ldots, z$
  **for** every possible extension of the rows $\underline{r}_j$,   $j = m, \ldots, i-1$
    **find** all possible entries $a_{i,k+1}, a_{i,k+2}, \ldots, a_{i,n}$:
        $\underline{r}_i$ is orthogonal with all the previous rows
        every row and column has exactly one zero
    **store** the results in a new matrix $B_i$ whose rows are all the possible rows $\underline{r}_i$
  **end**
**end**
**extend** the $z$-th row of $A$ with the first row of $B_z$
**for** $i = z-1, \ldots, m$
  **complete** row $\underline{r}_i$ with the row of $B_i$, from which $\underline{r}_{i+1}$ occurs
**end**
**end** {of Procedure Extend Rows}
```

Application of the Algorithm Extend 2

In this example, the Algorithm Extend 2 is applied for k=5 and n=12.
Steps of the algorithm

1. One may start with

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & - & 1 & - & - \\ 1 & - & - & 1 & 1 \\ 1 & 1 & - & - & 1 \\ 1 & 1 & - & 1 & - \end{bmatrix}.$$

2. The first row and column are completed, without loss of generality, with entries 0, ± 1 taking into account the property of the weighing matrix $W(12, 11)$ to have exactly one zero in each row and column. The software package fills with zeros the rest of the entries of the sought 12×12 matrix;

$$A=\left[\begin{array}{cccccccccccc}
1&1&1&1&1&1&1&0&1&1&-&-\\
1&-&1&-&-&&&&&&&\\
1&-&-&1&1&&&&&&&\\
1&1&-&-&1&&&&&&&\\
1&1&-&1&-&&&&&&&\\
1&&&&&&&&&&&\\
1&&&&&&&&&&&\\
0&&&&&&&&&&&\\
-&&&&&&&&&&&\\
-&&&&&&&&&&&\\
1&&&&&&&&&&&\\
-&&&&&&&&&&&
\end{array}\right].$$

3. The algorithm takes as input this matrix A and finds all possible completions for rows 2–5 (columns 6–12), so that every row has exactly one zero, every column has at most one zero and the inner product of every two distinct rows is zero. If many ways have been found to complete rows 2–5, the algorithm keeps as a result the first solution found;

$$A=\left[\begin{array}{cccccccccccc}
1&1&1&1&1&1&1&0&1&1&-&-\\
1&-&1&-&-&-&-&-&1&1&-&0\\
1&-&-&1&1&1&-&-&0&-&-&1\\
1&1&-&-&1&-&0&-&1&-&1&-\\
1&1&-&1&-&-&-&1&-&0&-&-\\
1&&&&&&&&&&&\\
1&&&&&&&&&&&\\
0&&&&&&&&&&&\\
-&&&&&&&&&&&\\
-&&&&&&&&&&&\\
1&&&&&&&&&&&\\
-&&&&&&&&&&&
\end{array}\right].$$

4. The algorithm finds all possible completions for columns 2–5 (rows 6–12) in the same way it has done with the rows 2–5;

$$A=\left[\begin{array}{cccccccccccc}
1&1&1&1&1&1&1&0&1&1&-&-\\
1&-&1&-&-&-&-&-&1&1&-&0\\
1&-&-&1&1&1&-&-&0&-&-&1\\
1&1&-&-&1&-&0&-&1&-&1&-\\
1&1&-&1&-&-&-&1&-&0&-&-\\
1&1&-&-&-&&&&&&&\\
1&-&-&1&0&&&&&&&\\
0&1&1&1&1&&&&&&&\\
-&1&0&1&-&&&&&&&\\
-&-&-&0&1&&&&&&&\\
1&-&1&1&-&&&&&&&\\
-&0&-&1&-&&&&&&&
\end{array}\right].$$

5. The algorithm tries to complete, if possible, the rows 6–12 (columns 6–12) in the same way as before;

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & - & - \\ 1 & - & 1 & - & - & - & - & - & 1 & 1 & - & 0 \\ 1 & - & - & 1 & 1 & 1 & - & - & 0 & - & - & 1 \\ 1 & 1 & - & - & 1 & - & 0 & - & 1 & - & 1 & - \\ 1 & 1 & - & 1 & - & - & - & 1 & - & 0 & - & - \\ 1 & 1 & - & - & - & 1 & 1 & - & - & 1 & 0 & 1 \\ 1 & - & - & 1 & 0 & - & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & - & - & - & - & 1 & 1 & 1 \\ - & 1 & 0 & 1 & - & - & 1 & - & 1 & - & - & 1 \\ - & - & - & 0 & 1 & - & 1 & - & - & 1 & - & - \\ 1 & - & 1 & 1 & - & 0 & 1 & - & - & - & 1 & - \\ - & 0 & - & 1 & - & 1 & - & - & 1 & 1 & 1 & - \end{bmatrix}.$$

6. Finally, if the matrix A could be extended, the algorithm gives the completed matrix $W(12, 11)$ and verifies whether the relationship $AA^T = 11I_{12}$ is valid.

The following Proposition of Delsarte et al. [11] is a very helpful tool, which excludes many matrices by determining if they can or cannot be extended to a $W(n, n-1)$. This strategy constitutes actually a test for possible completion of a $W(n, n-1)$ matrix.

Proposition 4 ([11]) *Let A be a $W(n, n-1)$. Then A is H-equivalent with a matrix B, which has zero diagonal and satisfies*

$$BB^T = (-1)^{\frac{n+2}{2}} I.$$

In the sequel, one should examine if a matrix of order k with entries 0, ±1 can be extended to a $W(n, n-1)$. This can be done by carrying out the following steps.

1. Exchange rows and columns so that the 0's are on the diagonal;
2. Multiply columns by -1 so that all the entries of the first row are $+1$;
3. Multiply rows by -1 so that all the entries of the first column are $(-1)^{\frac{n+2}{2}}$;
4. Check if the resulting matrix C, which contains all the 0's on the diadonal, satisfies $CC^T = (-1)^{\frac{n+2}{2}} I$.

If the matrix C doesn't satisfy this relationship, it can't be completed to a $W(n, n-1)$. If the matrix C satisfies the test, then it is possible that it can be completed to a $W(n, n-1)$. This test is used in [51] for determining that some matrices definitely cannot be extended to a $W(12, 11)$. So, there is saved time by excluding these specific matrices from the application of Algorithm Extend 2 and the whole process becomes faster.

Utilizing Algorithm Extend 2 in combination with Proposition 4 yields the following result concerning the magnitude of the leading principal minors of the weighing matrix $W(12, 11)$.

Proposition 5 ([51]) *Let W be a $W(12, 11)$ weighing matrix. Then the following hold:*

1. $W(5) = 48$ or 40 or 36;
2. $W(6) = 160$ or 144 or 136 or 120;
3. $W(7) = 528$ or 440;
4. $W(8) = 1936$ or 1452.

8 An Algorithm Specifying the Existence of $k \times k(0, \pm 1)$ Submatrices in a $W(n, n-1)$

A new idea for computing $n - j$ minors of Hadamard matrices was proposed in 2001 [41]. First of all, the notion of a matrix denoted as U_j, containing all possible columns of a normalized Hadamard matrix clustered together for some number of rows j, was introduced as follows.

Definition 10 ([41]) Let $\mathbf{y}_{\beta+1}^T$ be the vectors containing the binary representation of each integer $\beta + 2^{j-1}$ for $\beta = 0, \ldots, 2^{j-1} - 1$. Replace all zero entries of $\mathbf{y}_{\beta+1}^T$ by -1 and define the $j \times 1$ vectors $\mathbf{u}_k = \mathbf{y}_{2^{j-1}-k+1}$, $k = 1, \ldots, 2^{j-1}$. U_j shall denote all the matrices with j rows and the appropriate number of columns, in which $\mathbf{u}_k$ occurs u_k times. In other words, U_j is the matrix containing all possible 2^{j-1} columns of size j with elements ± 1 starting with $+1$. So,

$$U_j = \begin{array}{ccccc} \overbrace{1\ldots1}^{u_1} & \overbrace{1\ldots1}^{u_2} & \ldots & \overbrace{1\ldots1}^{u_{2^{j-1}-1}} & \overbrace{1\ldots1}^{u_{2^{j-1}}} \\ 1\ldots1 & 1\ldots1 & \ldots & -\ldots- & -\ldots- \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ 1\ldots1 & 1\ldots1 & \ldots & -\ldots- & -\ldots- \\ 1\ldots1 & -\ldots- & \ldots & 1\ldots1 & -\ldots- \end{array} = \begin{array}{ccccc} u_1 & u_2 & \ldots & u_{2^{j-1}-1} & u_{2^{j-1}} \\ 1 & 1 & \ldots & 1 & 1 \\ 1 & 1 & \ldots & - & - \\ \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & \ldots & - & - \\ 1 & - & \ldots & 1 & - \end{array}.$$

Example 5

$$U_3 = \begin{array}{cccc} u_1 & u_2 & u_3 & u_4 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & - & - \\ 1 & - & 1 & - \end{array}, \quad U_4 = \begin{array}{cccccccc} u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 & u_8 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & - & - & - & - \\ 1 & 1 & - & - & 1 & 1 & - & - \\ 1 & - & 1 & - & 1 & - & 1 & - \end{array}.$$

The matrix U_j is important in the present study because it is used to depict a general form for the first j rows of a normalized weighing matrix.

The following algorithm determines if a $k \times k$ matrix A exists embedded within a $W(n, n-1)$, provided that the upper left $(k-1) \times (k-1)$ submatrix B of A always exists in the $W(n, n-1)$.

```
**Algorithm Exist 1** [49]
*Step 1*
**Read** the $k \times k$ matrix $A$ and the $(k-1) \times (k-1)$ matrix $B$
*Step 2*
**Create** the matrix $Z$

$$Z = \left[\begin{array}{c|c|cccccc} & & 0 & 1 & 1 & \cdots & \cdots & 1 \\ & & y_{21} & 0 & y_{23} & \cdots & \cdots & y_{2k} \\ B & U_k & y_{31} & y_{32} & 0 & y_{34} & \cdots & y_{3k} \\ & & \vdots & \vdots & & \vdots & \ddots & \vdots \\ \cline{1-1} 1\ z_2\ \cdots\ z_{k-1} & & y_{k1} & y_{k2} & y_{k3} & \cdots & y_{k,k-1} & 0 \end{array}\right], \text{ where } z_i,\ y_{ij} = \pm 1$$

**If** $B$ contains columns with 0
    they are excluded from the matrix $Z(:, n-k+1 : n)$
*Step 3*
**If** $A$ has r 0's
    **Demand** that the r columns of $Z(:, n-k+1 : n)$, in which the 0's are in
    the same position as in $A$, take the appropriate values $y_{ij}$:
        they are identical with the r columns of $A$ containing the 0's
*Step 4*
**Procedure Solve**
**For** all possible values of $z_i$,$i = 2, \ldots, k-1$
    **Form** the system of $1 + \binom{k}{2}$ equations and $2^{k-1}$ variables which results
    from counting of columns and the inner products of every two distinct rows
    **Solve** the system for all $x_i$
    **Find** the minimum values for the $x_i$ which correspond to the columns of
    $A$, given that the number of columns appearing in $Z(:, 1 : k-1)$ is $\geq 1$
    **Formulate** (if necessary) conditions and/or restrictions for the order $n$
    or for some $x_i$:
        the columns of $A$ appear (the corresponding $x_i$ are all $\geq 1$)
**End**{of Procedure Solve}
**Else**
    **Do** Procedure Solve
```

The algorithm Exist 1 can be used, for example, to confirm that the matrices B_1, B_2, A_1 and A_2 of Lemma 7 always exists in a $W(n, n-1)$.

8.1 Another Algorithm Specifying the Existence of $k \times k$ $(0, +, -)$ Submatrices in a $W(n, n-1)$

Notation We denote by $U_{k,3}$ the first three rows of the previously defined matrix U_k.

$$U_{k,3} = \begin{array}{ccccc} x_1 & x_2 & \ldots & x_{2^{k-1}-1} & x_{2^{k-1}} \\ 1 & 1 & \ldots & 1 & 1 \\ 1 & 1 & \ldots & - & - \\ 1 & 1 & \ldots & - & - \end{array}.$$

Example 6

$$U_{3,3} = \begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & - & - \\ 1 & - & 1 & - \end{array} = U_3\ , \quad U_{4,3} = \begin{array}{cccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & - & - & - & - \\ 1 & 1 & - & - & 1 & 1 & - & - \end{array}.$$

The following algorithm specifies the existence of a $k \times k$ submatrix A in a $W(n, n-1)$, given that the upper left $(k-1) \times (k-1)$ submatrix B of A always exists in a $W(n, n-1)$.

Algorithm Exist 2 [49]
Step 1
Read the $k \times k$ matrix A and the $(k-1) \times (k-1)$ matrix B
Step 2
Denote with C the upper left $3 \times (k-1)$ submatrix of B
Step 3
Create the matrix Y

$$Y_k = \left[\begin{array}{c|c|cccccc} & & 0 & 1 & 1 & 1 & \cdots & 1 \\ C & U_{k,3} & y_{21} & 0 & y_{23} & y_{24} & \cdots & y_{2k} \\ & & y_{31} & y_{32} & 0 & y_{34} & \cdots & y_{3k} \end{array}\right], \text{ where } y_{ij} = \pm 1$$

Step 4
Formulate a Lemma for the number of columns of Y_k
Step 5
Find the maximum values for the x_i which correspond to the columns of A
Step 6
Formulate (if necessary) conditions for the order n:
the columns of A appear (the corresponding x_i are all ≥ 1)

Applying algorithm Exist 2 and taking into account also the Distribution Lemma, one can prove the following result regarding the existence of 5×5 submatrices in $W(n, n-1)$ matrices.

Lemma 9 ([49]) *One of the following matrices*

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & - & 1 & - & - \\ 1 & - & - & 1 & 1 \\ 1 & 1 & - & - & 1 \\ 1 & 1 & - & 1 & - \end{bmatrix}, \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & - & 1 & - & - \\ 1 & - & - & 1 & 0 \\ 1 & 1 & - & - & 1 \\ 1 & 1 & - & 1 & - \end{bmatrix}, \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & - & 1 & - & - \\ 1 & - & - & 1 & 1 \\ 1 & 1 & - & - & 1 \\ 1 & 1 & - & 0 & - \end{bmatrix}, \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & - & 1 & - & - \\ 1 & - & - & 1 & - \\ 1 & 1 & - & - & - \\ 1 & 1 & 1 & 1 & - \end{bmatrix}, \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & - & 1 & - & - \\ 1 & - & - & 1 & - \\ 1 & 1 & - & - & 0 \\ 1 & 1 & - & 1 & - \end{bmatrix},$$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & - & 1 & - & - \\ 1 & - & - & 1 & - \\ 1 & 1 & - & - & 1 \\ 1 & 1 & 0 & - & - \end{bmatrix}, \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & - & 1 & - & - \\ 1 & - & - & 1 & 1 \\ 1 & 1 & - & - & 1 \\ 1 & 1 & - & 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 & 0 & - & 1 \\ 1 & - & - & - & - \\ 1 & - & 1 & 1 & 1 \\ 1 & 1 & - & 1 & - \\ 1 & 1 & 1 & - & - \end{bmatrix}, \begin{bmatrix} 1 & 1 & 0 & - & 1 \\ 1 & - & - & - & - \\ 1 & - & 1 & 1 & 0 \\ 1 & 1 & - & 1 & - \\ 1 & 1 & 1 & - & - \end{bmatrix}, \begin{bmatrix} 1 & 1 & 0 & - & 1 \\ 1 & - & - & - & - \\ 1 & - & 1 & 1 & 1 \\ 1 & 1 & - & 1 & 0 \\ 1 & 1 & 1 & 0 & - \end{bmatrix}$$

always exists in a $W(n, n-1)$ with $n \geq 8$.

Manipulating appropriately the outcomes of Lemma 9 in combination with relationship (3) one derives the next Theorem.

Theorem 5 ([49]) *Let W be a CP skew and symmetric conference matrix, of order $n \geq 8$ then if GE is performed on W the fifth pivot is 2 or 3 or $\frac{9}{4}$ or $\frac{10}{3}$ or $\frac{10}{4}$.*

9 Theoretical Results

The target is to calculate first the $(n-1) \times (n-1)$ minors of a skew and symmetric conference matrix of order n. The usage of a variation of a clever proof used by combinatorialists was exploited. This tool evaluates the determinant of a matrix satisfying $AA^T = (k - \lambda)I + \lambda J$, where I is the $v \times v$ identity matrix, J is the $v \times v$ matrix of ones and $k,\ \lambda$ are integers to simplify our proofs. The determinant is $k + (v-1)\lambda(k-\lambda)^{v-1}$.

Lemma 10 ([9, p. 239]) *Let $A = (k - \lambda)I_v + \lambda J_v$, where $k,\ \lambda$ are integers. Then,*

$$\det A = [k + (v-1)\lambda](k-\lambda)^{v-1} \tag{4}$$

and for $k \neq \lambda, -(v-1)\lambda$, A is nonsingular with

$$A^{-1} = \frac{1}{k^2 + (v-2)k\lambda - (v-1)\lambda^2}\{[k + (v-1)\lambda]I_v - \lambda J_v\}. \tag{5}$$

The first part of Lemma 10 is straightforward to show. The second part is a special case of the Sherman-Morrison formula, which computes the inverse of a rank-one-correction of a nonsingular matrix B as

$$(B - uv^T)^{-1} = B^{-1} + \frac{B^{-1}uv^TB^{-1}}{1 - v^TB^{-1}u},$$

where u, v are vectors and $v^TB^{-1}u \neq 1$. Indeed, (5) occurs for $B = (\kappa - \lambda)I_v$ and $u = -\lambda[1\ 1\ \ldots\ 1]^T$ and $v = [1\ 1\ \ldots\ 1]^T$.

Lemma 11 (Schur Determinant Formula [35, p. 21]) *Let* $B = \begin{bmatrix} B_1 & B_2 \\ B_3 & B_4 \end{bmatrix}$, B_1 *nonsingular. Then*

$$\det B = \det B_1 \cdot \det(B_4 - B_3B_1^{-1}B_2). \tag{6}$$

The auxiliary results of Lemmas 10 and 11 can be used for determining the $(n-1)$ minors of a skew and symmetric conference matrix of order n.

Proposition 6 ([37, 40, 44, 50]) *Let W be a CP skew and symmetric or conference matrix of order n. Then all possible minors of W of the respective orders are given by:*

1) The $(n-1) \times (n-1)$ *minors of W are* 0 *and* $(n-1)^{\frac{n}{2}-1}$.
2) The $(n-2) \times (n-2)$ *minors of W are* 0, $(n-1)^{\frac{n}{2}-2}$ *and* $2(n-1)^{\frac{n}{2}-2}$.
3) The $(n-3) \times (n-3)$ *minors of W are*

(i) 0, $2(n-1)^{\frac{n}{2}-3}$ *and* $4(n-1)^{\frac{n}{2}-3}$ *for* $n \equiv 0 \pmod 4$.
(ii) $2(n-1)^{\frac{n}{2}-3}$ *and* $4(n-1)^{\frac{n}{2}-3}$ *for* $n \equiv 2 \pmod 4$.

Applying the previous results in appropriate combination with Eq. (3) leads to Theorem 6.

Theorem 6 ([50]) *When Gaussian Elimination is applied on a CP skew and symmetric conference matrix W of order n the last two pivots are* $n-1$, *and* $\frac{n-1}{2}$.

Theorem 7 ([50]) *When Gaussian Elimination is applied on a CP* $W(n, n-k)$ *the last two pivots are (in backward order)* $n-k$ *and* $\frac{n-k}{2}$.

10 Specification of Pivot Patterns

One can proceed the study by demonstrating the pivot structure of some small weighing matrices. Later on, the respective growth factors can be calculated with substitution of the results in (1).

Lemma 12 ([40, 50]) *The unique pivot structure of* $W(6,5)$ *is* $\{1,\ 2,\ 2,\ \frac{5}{2},\ \frac{5}{2},\ 5\}$. *The pivot patterns of* $W(8,7)$ *are* $\{1,\ 2,\ 2,\ 4,\ \frac{7}{4},\ \frac{7}{2},\ \frac{7}{2},\ 7\}$ *or* $\{1,\ 2,\ 2,\ 3,\ \frac{7}{3},\ \frac{7}{2},\ \frac{7}{2},\ 7\}$. *The pivot patterns of* $W(10,9)$ *are* $\{1,2,2,3,3,4,\frac{9}{4},\frac{9}{2},\frac{9}{2},9\}$ *or* $\{1,2,2,4,3,3,\frac{9}{4},\frac{9}{2},\frac{9}{2},9\}$ *or* $\{1,2,2,3,\frac{10}{4},\frac{18}{5},\frac{9}{3},\frac{9}{2},\frac{9}{2},9\}$.

One can establish pivots from the beginning and from the end of the pivot pattern of the $W(12,11)$ by deploying appropriately the acquainted outcomes of the application of Algorithm Extend 2 with the respective input values, combined with careful substitution in (3). After the first 9 and he last 2 pivots are declared, the 10th pivot can be calculated using the property

$$p_{10} = \frac{det(W(12,11))}{\prod_{i=1_{i\neq 10}}^{12} p_i} = \frac{11^6}{1\cdot 2\cdot 2\cdot 3\cdot \frac{10}{3}\cdot \frac{17}{5}\cdot \frac{11}{17/5}\cdot \frac{11}{5/2}\cdot \frac{11}{4}\cdot \frac{11}{2}\cdot 11} \text{ or}$$

$$\frac{11^6}{1\cdot 2\cdot 2\cdot 4\cdot 3\cdot \frac{10}{3}\cdot \frac{11}{10/3}\cdot \frac{11}{3}\cdot \frac{11}{4}\cdot \frac{11}{2}\cdot 11} \text{ or}$$

$$\frac{11^6}{1\cdot 2\cdot 2\cdot 3\cdot 3\cdot 4\cdot \frac{11}{3}\cdot \frac{11}{3}\cdot \frac{11}{4}\cdot \frac{11}{2}\cdot 11} \Rightarrow p_{10} = \frac{11}{2}.$$

Lemma 13 ([51]) *The pivot patterns of the* $W(12,11)$ *are* $(1,2,2,3,\frac{10}{3},\frac{17}{5},\frac{11}{17/5},\frac{11}{5/2},\frac{11}{4},\frac{11}{2},\frac{11}{2},11)$ *or* $(1,2,2,4,3,\frac{10}{3},\frac{11}{10/3},\frac{11}{3},\frac{11}{4},\frac{11}{2},\frac{11}{2},11)$ *or* $(1,2,2,3,3,4,\frac{11}{3},\frac{11}{3},\frac{11}{4},\frac{11}{2},\frac{11}{2},11)$.

Lemma 13, the definitions of the growth factor (1) and (2) yield Theorem 8.

Theorem 8 ([51]) *The growth factor of* $W(12,11)$ *is* 11.

11 General Results for $W(n, n-k)$

In the sequel, the computation of all possible minors of orders $n-1,\ n-2,\dots$ within a weighing matrix of order n and weight $n-k$, i.e. $W(n,n-k)$, is studied for $k \geq 1$ [44]. The respective proofs are carried out by separating all possible $k\times k$ blocks in the upper left corner of a $W(n,n-k)$. Then, after considering the form of the first k rows and columns $k+1$ to n and with the aid of the representation of the aforementioned matrix U_j, Lemmas 10 and 11, one derives the requested results. Writing "all possible $k\times k$ blocks" refers to exhaustive searches over the possible entries 0, ±1 that can appear in the $k\times k$ corner and taking the specifications of the weighing matrix into account.

Proposition 7 ([40]) *Let W be a* $W(n,n-k)$, $k\geq 1$. *Then all possible* $(n-1)\times(n-1)$ *minors of W are: 0 and* $(n-k)^{\frac{n}{2}-1}$.

Working in a similar fashion, one can proceed by specifying all possible minors of order $(n-2)\times(n-2)$ of a $W(n, n-k),\ k \geq 1$. Additionally, the following structural features of a $W(n, n-k)$ should be established.

Lemma 14 ([40]) *If one specific row of a $W(n, n-k)$ is fixed, $k \geq 1$, we can always find a second row, so that the two rows have the form:*

$$\begin{array}{ccccc} \overbrace{0\ 0\ \ldots\ 0}^{j} & \overbrace{0\ \ldots\ 0}^{k-j} & \overbrace{1\ \ldots\ 1}^{k-j} & \overbrace{1\ \ldots\ 1}^{s} & \overbrace{1\ \ldots\ 1}^{s} \\ 0\ 0\ \ldots\ 0 & 1\ \ldots\ 1 & 0\ \ldots\ 0 & 1\ \ldots\ 1 & -\ \ldots\ - \end{array}, \tag{P}$$

for some j even, $0 \leq j \leq k$. This can be always achieved by performing the appropriate H-equivalence operations. Particularly for $k = 1$, the result holds trivially for $j = 0$.

Corollary 2 ([40]) *If one specific row of a $W(n, n-2)$ is fixed, we can always find a second row so that the two rows have the form*

$$\begin{array}{ccc} 0\ 0 & \overbrace{1\ \ldots\ 1}^{s} & \overbrace{1\ \ldots\ 1}^{s} \\ 0\ 0 & 1\ \ldots\ 1 & -\ \ldots\ - \end{array}.$$

This can be always achieved by performing the appropriate H-equivalent operations.

Proposition 8 ([40]) *Let W be a $W(n, n-k)$, $k \geq 1$. Then all possible $(n-2)\times(n-2)$ minors of W are: 0, $(n-k)^{\frac{n}{2}-2}$ and $2(n-k)^{\frac{n}{2}-2}$.*

Proposition 9 ([40]) *Let W be a $W(n, n-2)$. Then all possible $(n-3)\times(n-3)$ minors of W are: 0, $(n-2)^{\frac{n}{2}-3}$, $2(n-2)^{\frac{n}{2}-3}$, $3(n-2)^{\frac{n}{2}-3}$ and $4(n-2)^{\frac{n}{2}-3}$.*

Appropriate application of Algorithm Exist 1 assures that the matrices B_1 and B_2 of Lemma 7 can exist embedded in a $W(n, n-2)$.

Lemma 15 ([40]) *H-equivalence operations can be used to ensure that the submatrices B_1 and B_2 always occur in a $W(n, n-2)$ for large enough n.*

Lemma 16 ([40]) *H-equivalence operations can be used to ensure that the following submatrices always occur in a $W(n, n-2)$ for large enough n:*

$$A_1 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & - & - & 1 \\ 1 & 1 & - & - \\ 1 & - & 1 & - \end{bmatrix} \text{ or } A_2 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & - & - & 0 \\ 1 & 1 & - & - \\ 1 & - & 1 & - \end{bmatrix} \text{ or } A_3 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & - & - & 0 \\ 1 & 1 & - & - \\ 1 & 0 & 1 & - \end{bmatrix}.$$

Theorem 9 ([42, 43] The most Generalized Version of the Determinant Simplification Theorem) *Let $A = (k_i - a_{ii})I_{b_1,b_2,\cdots,b_z} + a_{ij}J_{b_1,b_2,\cdots,b_z}$, $i, j = 1,\ldots,z$. Then*

$$\det\ A = \prod_{i=1}^{z}(k_i - a_{ii})^{b_i - 1} \det\ D,$$

where

$$D = \begin{bmatrix} k_1 + (b_1 - 1)a_{11} & b_2 a_{12} & b_3 a_{13} \cdots & b_z a_{1z} \\ b_1 a_{21} & k_2 + (b_2 - 1)a_{22} & b_3 a_{23} \cdots & b_z a_{2z} \\ \vdots & \vdots & \vdots & \vdots \\ b_1 a_{z1} & b_2 a_{z2} & b_3 a_{z2} \cdots & k_z + (b_z - 1)a_{zz} \end{bmatrix}.$$

Furthermore, for the needs of the study in consideration, an algorithm for computing minors of $W(n, n-2)$ was developed in [44]. The following algorithm calculates the value of the determinant of the $(n-j) \times (n-j)$ lower right submatrix of a $W(n, n-2)$.

Let us consider the following two possible representations of $W \equiv W(n, n-2)$. The requested $(n-j) \times (n-j)$ minor of W is the determinant of the submatrix C. Any matrix $W = W(n, n-2)$ can be written according to the following two cases (it follows from an extension of the result of Corollary 2), cf. also Example 7.

Example 7

$$W(6,4) = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & - & - \\ 1 & 1 & 0 & 0 & 1 & - \\ 1 & 1 & 0 & 0 & - & 1 \\ 1 & - & 1 & - & 0 & 0 \\ 1 & - & - & 1 & 0 & 0 \end{bmatrix}.$$

Considering the above $W(6,4)$ we see that for j even there are $j/2$ 2×2 blocks with zeros in the upper left $j \times j$ corner M, while for j odd there are $(j-1)/2$ and there is also a zero entry in its lower right corner.

First Case, $j \equiv 0 \pmod 2$ (j even). $W = \begin{bmatrix} M & U_j \\ U_j^T & C \end{bmatrix}$.

M, C are $j \times j$ and $(n-j) \times (n-j)$ matrices, respectively. M has $j/2$ 2×2 blocks of zeros on the diagonal. The elements in the $(n-j) \times (n-j)$ matrix CC^T can be permuted to appear in the form

$$CC^T = (n - 2 - j - a_{ii}) I_{u_1, u_2, \cdots, u_{2^{j-1}}} + a_{ik} J_{u_1, u_2, \cdots, u_{2^{j-1}}},$$

where $(a_{ik}) = (-\underline{u}_i \cdot \underline{u}_k)$, $a_{ii} = (-\underline{u}_i \cdot \underline{u}_i) = -j$, with $\cdot$ the inner product ($\underline{u}_i$ denotes the i-th column of U_j). By the Determinant Simplification Theorem (Theorem 9)

$$\det CC^T = (n-2)^{n-2^{j-1}-j} \det D,$$

where D, of order 2^{j-1}, is given by

$$D = \begin{bmatrix} n - 2 - j u_1 & u_2 a_{12} & u_3 a_{13} \cdots & u_z a_{1z} \\ u_1 a_{21} & n - 2 - j u_2 & u_3 a_{23} \cdots & u_z a_{2z} \\ \vdots & \vdots & \vdots & \vdots \\ u_1 a_{z1} & u_2 a_{z2} & u_3 a_{z2} \cdots & n - 2 - j u_z \end{bmatrix},$$

with $z = 2^{j-1}$.

The $(n-j)\times(n-j)$ minor of W is the determinant of C, for which we have

$$\det C = ((n-2)^{n-2^{j-1}-j}\det D)^{1/2}. \tag{7}$$

Second Case, $j\equiv 1 \pmod 2$ (j odd). $W=\begin{bmatrix} M & \underline{v}\, U_j \\ \underline{v}^T & \\ U_j^T & C \end{bmatrix}$.

$M,\ C$ are $j\times j$ and $(n-j)\times(n-j)$ matrices respectively. M has $(j-1)/2$ 2×2 blocks of zeros on the diagonal and one zero element in the lower right entry. The vector $\underline{v}$ of order $j\times 1$ is of the form $[v^{(j-1)}\ 0]^T$, where $v^{(j-1)}$ is a possible column of U_{j-1}. The elements in the $(n-j)\times(n-j)$ matrix CC^T can be permuted to appear in the form

$$CC^T=\begin{bmatrix} n-1-j & \underline{y} \\ \underline{y}^T & E \end{bmatrix},$$

where $E=(n-2-j-a_{ii})I_{u_1,u_2,\cdots,u_{2j-1}}+a_{ik}J_{u_1,u_2,\cdots,u_{2j-1}}$, $(a_{ik})=(-\underline{u}_i\cdot\underline{u}_k)$, with $\cdot$ the inner product, and $\underline{y}$ is a vector of order $1\times(n-j-1)$, whose elements are obtained from the inner products of $\underline{v}$ with $\underline{u}_i$. Precisely, we have

$$\underline{y}=[\underbrace{-(\underline{v}\cdot\underline{u}_1)\ldots-(\underline{v}\cdot\underline{u}_1)}_{u_1\ times}\ \underbrace{-(\underline{v}\cdot\underline{u}_2)\ldots-(\underline{v}\cdot\underline{u}_2)}_{u_2\ times}\ \cdots\ \underbrace{-(\underline{v}\cdot\underline{u}_{2j-1})\ldots-(\underline{v}\cdot\underline{u}_{2j-1})}_{u_{2j-1}\ times}]$$

$$=[\underbrace{b_1\ldots b_1}_{u_1}\ \underbrace{b_2\ldots b_2}_{u_2}\ \cdots\ \underbrace{b_z\ldots b_z}_{u_z}],$$

where $b_i=(-\underline{v}\cdot\underline{u}_i)$ and $z=2^{j-1}$.
We want to calculate *det* CC^T with help of formula (6). So, we have

$$\det CC^T=(n-1-j)\cdot\det(E-\frac{1}{n-1-j}\underline{y}^T\underline{y}).$$

We have $\underline{y}^T\underline{y}=\gamma_{ik}J_{u_1,u_2,\ldots,u_z}$, where $\gamma_{ik}=b_ib_k$.

$$X\equiv E-\frac{1}{n-1-j}\underline{y}^T\underline{y}$$

$$=(n-2-j-a_{ii})I_{u_1,u_2,\cdots,u_{2j-1}}+a_{ik}J_{u_1,u_2,\cdots,u_{2j-1}}-\frac{1}{n-1-j}\gamma_{ik}J_{u_1,u_2,\ldots,u_z}$$

$$=(n-2-j-a_{ii})I_{u_1,u_2,\cdots,u_{2j-1}}+(a_{ik}-\frac{1}{n-1-j}\gamma_{ik})J_{u_1,u_2,\cdots,u_{2j-1}}$$

We set $\delta_{ik} = \dfrac{1}{n-1-j}\gamma_{ik}$. For the sake of simplicity we omit the subscripts $u_1, \cdots, u_{2^{j-1}}$. Hence

$$\begin{aligned}
X &= (n-2-j-a_{ii})I + (a_{ik}-\delta_{ik})J \\
&= [n-2-j-\delta_{ii}-(a_{ii}-\delta_{ii})]I + (a_{ik}-\delta_{ik})J \\
&= (\lambda_i - \varepsilon_{ii})I + \varepsilon_{ik}J,
\end{aligned}$$

where $\lambda_i = n-2-j-\delta_{ii}$ and $\varepsilon_{ik} = a_{ik} - \delta_{ik}$.
By the Determinant Simplification Theorem (Theorem 9)

$$\det X = \prod_{i=1}^{z}(\lambda_i - \varepsilon_{ii})^{u_i - 1} \det D, \tag{8}$$

where

$$D = \begin{bmatrix}
\lambda_1 + (u_1-1)\varepsilon_{11} & u_2\varepsilon_{12} & u_3\varepsilon_{13} & \cdots & u_z\varepsilon_{1z} \\
u_1\varepsilon_{21} & \lambda_2 + (u_2-1)\varepsilon_{22} & u_3\varepsilon_{23} & \cdots & u_z\varepsilon_{2z} \\
\vdots & \vdots & \vdots & & \vdots \\
u_1\varepsilon_{z1} & u_2\varepsilon_{z2} & u_3\varepsilon_{z2} & \cdots & \lambda_z + (u_z-1)\varepsilon_{zz}
\end{bmatrix}.$$

Finally

$$\det C = ((n-1-j)\det X)^{1/2}. \tag{9}$$

Remark 4 For the appropriate implementation of the algorithm the following notion is required. The most practical way to manage the variables, which represent the unknown number of columns of U_j, is to denote with $u_l^{(s)}$, $l = 1, \ldots, 2^{k-1}$, $k = 3, \ldots, j$, $s = 1, \ldots, j-2$, the number of columns starting with the same vectors of order $s+2$. For example, for $j = 5$, the matrix U_5 will be of the form (here $+$ and $-$ stand for $+1$ and -1, respectively):

$$\begin{array}{cccccccccccccccc}
\multicolumn{4}{c}{\overbrace{\hphantom{+\quad+\quad+\quad+}}^{u_1^{(1)}}} & \multicolumn{4}{c}{\overbrace{\hphantom{+\quad+\quad+\quad+}}^{u_2^{(1)}}} & \multicolumn{4}{c}{\overbrace{\hphantom{+\quad+\quad+\quad+}}^{u_3^{(1)}}} & \multicolumn{4}{c}{\overbrace{\hphantom{+\quad+\quad+\quad+}}^{u_4^{(1)}}} \\
\multicolumn{2}{c}{\overbrace{\hphantom{+\quad+}}^{u_1^{(2)}}} & \multicolumn{2}{c}{\overbrace{\hphantom{+\quad+}}^{u_2^{(2)}}} & \multicolumn{2}{c}{\overbrace{\hphantom{+\quad+}}^{u_3^{(2)}}} & \multicolumn{2}{c}{\overbrace{\hphantom{+\quad+}}^{u_4^{(2)}}} & \multicolumn{2}{c}{\overbrace{\hphantom{+\quad+}}^{u_5^{(2)}}} & \multicolumn{2}{c}{\overbrace{\hphantom{+\quad+}}^{u_6^{(2)}}} & \multicolumn{2}{c}{\overbrace{\hphantom{+\quad+}}^{u_7^{(2)}}} & \multicolumn{2}{c}{\overbrace{\hphantom{+\quad+}}^{u_8^{(2)}}} \\
u_1^{(3)} & u_2^{(3)} & u_3^{(3)} & u_4^{(3)} & u_5^{(3)} & u_6^{(3)} & u_7^{(3)} & u_8^{(3)} & u_9^{(3)} & u_{10}^{(3)} & u_{11}^{(3)} & u_{12}^{(3)} & u_{13}^{(3)} & u_{14}^{(3)} & u_{15}^{(3)} & u_{16}^{(3)} \\
+ & + & + & + & + & + & + & + & + & + & + & + & + & + & + & + \\
+ & + & + & + & + & + & + & + & - & - & - & - & - & - & - & - \\
+ & + & + & + & - & - & - & - & + & + & + & - & - & - & + & - \\
+ & + & - & - & + & + & - & - & + & + & - & + & - & + & - & - \\
+ & - & + & - & + & - & + & - & + & - & + & + & + & - & - & -
\end{array}$$

We see easily that the following relation connects the above numbers of columns:

$$u_{2l-1}^{(s+1)} + u_{2l}^{(s+1)} = u_l^{(s)}, \quad l = 1, 2, \ldots, 2^{j-1}, \quad s \geq 1. \tag{10}$$

The following algorithm calculates the value of the determinant of the $(n - j) \times (n - j)$ lower right submatrix of a $W(n, n - k)$.

Algorithm Minors [44]

```
Step 1: Read all $j \times j$ matrices $M$, which can exist in the upper left corner of a
$W(n, n - k)$
Step 2: For every matrix $M$
            Create the $j \times n$ matrix $N = [M \;\; U_j]$, if $j$ even, or $N = [M \;\; v \;\; U_j]$, if $j$
odd
Step 3: $s := 0$
        For $k = 3, 4, \ldots, j$
Step 4: Consider the first $k$ rows of $N$
          $s := s + 1$
          Set $u_l^{(s)}$ the number of columns starting with the vectors $\underline{u}_l$, $l = 1, \ldots, 2^{k-1}$
          Form the system resulting from orthogonality of rows and counting of
          columns, with unknowns $u_l^{(s)}$
          Solve the system taking into account (10)
        End {for $k = 3, \ldots, j$}
Step 5: For every acceptable solution $(u_1^{(j-2)}, \ldots, u_{2^{j-1}}^{(j-2)})$ of the system calcu-
late the
                    values of the $(n - j) \times (n - j)$ minors, using (7), or (8) and (9).
         End {for every matrix $M$}
    End {of Algorithm}
```

Application of the Algorithm Minors

We want to calculate the $n-4$ minor of a $W(n, n-2)$. After finding all possible $2^2 = 4$ matrices M, we create $N = [M \;\; U_4]$, where M is of the form

$$\begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & a \\ 1 & 1 & 0 & 0 \\ 1 & b & 0 & 0 \end{bmatrix},$$

with $a, b = \pm 1$.
For $k = 3$ the system is

$$\begin{cases} u_1^{(1)} + u_2^{(1)} + u_3^{(1)} + u_4^{(1)} = n - 4 \\ u_1^{(1)} + u_2^{(1)} - u_3^{(1)} - u_4^{(1)} = -1 - a \\ u_1^{(1)} - u_2^{(1)} + u_3^{(1)} - u_4^{(1)} = 0 \\ u_1^{(1)} - u_2^{(1)} - u_3^{(1)} + u_4^{(1)} = 0 \end{cases}$$

with solution $(u_1^{(1)}, u_2^{(1)}, u_3^{(1)}, u_4^{(1)}) = (\frac{1}{4}(n-5-a), \frac{1}{4}(n-5-a), \frac{1}{4}(n-3+a), \frac{1}{4}(n-3+a))$.
For $k = 4$ the system is

$$\begin{cases} u_1^{(2)} + u_2^{(2)} + u_3^{(2)} + u_4^{(2)} + u_5^{(2)} + u_6^{(2)} + u_7^{(2)} + u_8^{(2)} = n-4 \\ u_1^{(2)} + u_2^{(2)} + u_3^{(2)} + u_4^{(2)} - u_5^{(2)} - u_6^{(2)} - u_7^{(2)} - u_8^{(2)} = -1-a \\ u_1^{(2)} - u_2^{(2)} - u_3^{(2)} + u_4^{(2)} + u_5^{(2)} - u_6^{(2)} - u_7^{(2)} + u_8^{(2)} = 0 \\ u_1^{(2)} - u_2^{(2)} + u_3^{(2)} - u_4^{(2)} + u_5^{(2)} - u_6^{(2)} + u_7^{(2)} - u_8^{(2)} = 0 \\ u_1^{(2)} + u_2^{(2)} - u_3^{(2)} - u_4^{(2)} - u_5^{(2)} - u_6^{(2)} + u_7^{(2)} + u_8^{(2)} = 0 \\ u_1^{(2)} - u_2^{(2)} + u_3^{(2)} - u_4^{(2)} - u_5^{(2)} + u_6^{(2)} - u_7^{(2)} + u_8^{(2)} = 0 \\ u_1^{(2)} - u_2^{(2)} - u_3^{(2)} + u_4^{(2)} + u_5^{(2)} - u_6^{(2)} - u_7^{(2)} + u_8^{(2)} = -1-b \end{cases}$$

with solution $(u_1^{(2)}, u_2^{(2)}, u_3^{(2)}, u_4^{(2)}, u_5^{(2)}, u_6^{(2)}, u_7^{(2)}, u_8^{(2)}) = (\frac{1}{4}(n-5-b) - u_8^{(2)}, u_8^{(2)}, u_8^{(2)}, \frac{1}{4}(n-5-b) - u_8^{(2)}, u_8^{(2)}, \frac{1}{4}(n-3+a) - u_8^{(2)}, \frac{1}{4}(n-3+a) - u_8^{(2)}, u_8^{(2)})$.

Since $u_7^{(2)} + u_8^{(2)} = u_4^{(1)}$ and thus $u_8^{(2)} = u_4^{(1)} - u_7^{(2)} \leq u_4^{(1)}$ ($u_7^{(2)}$ is always a non negative number), the range of values for $u_8^{(2)}$ is from 0 to $u_4^{(1)}$. We now compute for all the possible values of $u_8^{(2)}$ the acceptable solutions for the remaining $u_i^{(2)}$ and calculate the requested minor from (7).

For example, for $n = 12$, we have $0 \leq u_8^{(2)} \leq \frac{1}{4}(9+a)$. For all possible values of a the upper bound is 2 ($u_8^{(2)}$ must be an integer), so for $u_8^{(2)} = 0, 1, 2$ we find the possible values for the rest of $u_i^{(2)}$ and finally apply formula (7). The resulting value for the 8×8 minor of the $W(12, 10)$, if we have 2×2 blocks with zeros on the diagonal of M, is always 400.

Proposition 10 ([37]) *Let W be a weighing matrix $W(n, n-1)$ of order $n > 6$, with zeros on the diagonal. Then the $(n-1) \times (n-1)$ minors of W are $W(n-1) = 0$.*

Proposition 11 ([37]) *Let W be a weighing matrix $W(n, n-1)$ of order $n > 6$, with zeros on the diagonal. Then the $(n-2) \times (n-2)$ minors of W are $W(n-2) = (n-1)^{\frac{n}{2}-2}$.*

We see that, when we have zeros on the diagonal, we get the lowest value from those presented in Proposition 6, , that is $W(n, n-1) = 0$. This agrees with the result of Lemma 17, as $n-1$ is odd and the submatrix C is skew-symmetric with real elements.

Furthermore, for $W(n, n-2)$ we get the lowest non-zero value $W(n-2) = (n-1)^{\frac{n}{2}-2}$. For the case $n \equiv 0 \pmod 4$, this agrees with the result of Lemma 17, as $n-2$ is even and the submatrix C is skew-symmetric with real elements.

Additionally, when we have zeros on the diagonal, we get the lowest values for $n-3$, from those presented in Proposition 6, that is $W(n-3) = 0$ for $n \equiv 0 \pmod 4)$ and $2(n-1)^{\frac{n}{2}=3}$ for $n \equiv 2 \pmod 4$. The zero value for $n \equiv 0 \pmod 4$ agrees with the result of Lemma 17. Since all the matrices found by removing an $l \times l$ submatrix, l odd, from a skew-symmetric weighing matrix of order $n \equiv 0 \pmod 4$ while preserving the skew-symmetry satisfy the previous sentence, we have that all the $(n-l) \times (n-l)$ minors are zero. Indeed, for the determinant of skew-symmetric matrices we have the following result.

Lemma 17 ([36])

1) *If n is odd and A is a skew-symmetric matrix with real elements then* $\det A = 0$. *If n is odd and the elements of the matrix A of order n are not from the field of characteristic 2 , then* $\det A = 0$.
2) *If n is even and A is a skew-symmetric matrix with real elements then* $\det A$ *is* $PF(A)^2$, *where* $PF(A)$ *is the Pfaffian of A, a polynomial in the entries of A.*

Proposition 12 ([44]) $(n-2) \times (n-2)$ *minors of a weighing matrix* $W(n, n-1)$, *where n is even, are*

$$0,\ (n-1)^{\frac{n}{2}-2},\ 2(n-1)^{\frac{n}{2}-2}.$$

In [37] was introduced a nobel approach for calculating minors of weighing matrices. It is based principally on the techniques described before for the same purpose, on appropriate partitioning of the matrices in blocks. The proofs take advantage of Lemma 11 and of a fundamental feature of orthogonal matrices presented in [60].

Lemma 18 ([37, 60]) *Given any unitary matrix*

$$U = \begin{bmatrix} A & B \\ C & D \end{bmatrix},$$

where A and D are square matrices not necessarily of the same size, then we have $|\det(A)| = |\det(D)|$.

Lemma 18 was proved in [60] by utilizing appropriate algebraic manipulations of the blocks of the unitary matrix. In [37] was proposed a novel approach based on the eigenvalues of particular products of blocks for the same purpose, which offers possibilities for further processing. Both proofs take the orthogonality of the initial matrices appropriately into account.

Corollary 3 ([37]) *Let*

$$W \equiv W(n, n-1) = \begin{bmatrix} A & B \\ C & D \end{bmatrix}.$$

be a weighing matrix partitioned as above with $l \leq \frac{n}{2}$. *Then the lower right* $(n - l) \times (n-l),\ l \geq 1$, *minor of W is*

$$W[n-l] = \det D = (n-1)^{\frac{n}{2}-l} \det A.$$

Proposition 13 ([37]) *Let W be a weighing matrix* $W(n, n-1)$ *of order* $n \geq 8$,*where n is even, with zeros on the diagonal. Then,*

$$W(n-3) = \begin{cases} 0, & \textit{for } n \equiv 0 \pmod 4, \\ 2(n-1)^{\frac{n}{2}-3}, & \textit{for } n \equiv 2 \pmod 4. \end{cases}$$

Theorem 10 ([44]) *When Gaussian Elimination is applied on a CP $W(n, n-k)$ the last two pivots are (in backward order) $n-k$ and $\frac{n-k}{2}$.*

Proposition 14 ([44]) *Let W be a CP $W(n, n-2)$, of order $n \geq 6$, then if GE is performed on W the third pivot is* 2*.*

Proposition 15 ([44]) *Let W be a CP $W(n, n-2)$, of order $n \geq 10$, then if GE is performed on W the fourth pivot is* 3 *or* 4 *or* $\frac{5}{2}$*.*

Theorem 11 ([44]) *When GE is applied on a CP $W(n, n-2)$ the last three pivots are (in backward order) $n-2$, $\frac{n-2}{2}$ and $\frac{n-2}{2}$.*

Proposition 16 ([37]) *Let W be a weighing matrix, $W(n, n-1)$, of order $n \geq 10$, with zeros on the diagonal. Then the $(n-4) \times (n-4)$ minors of W are*

$$W(n-4) = \begin{cases} 16(n-1)^{\frac{n}{2}-4}, \textit{for } n \equiv 0 \ (\textit{mod } 4), \\ 12(n-1)^{\frac{n}{2}-4}, \textit{for } n \equiv n \equiv 2 \ (\textit{mod } 4). \end{cases}$$

Proposition 17 ([37]) *Let W be a CP skew and symmetric conference matrix, $W(n, n-1)$, of order $n > 10$. Then the $(n-4) \times (n-4)$ minors of W are $W(n-4) = 16(n-1)^{\frac{n}{2}-4}$ or $W(n-4) = 12(n-1)^{\frac{n}{2}-4}$.*

Theorem 12 ([37]) *Let W be a weighing matrix $W(n, n-1)$ of order $n > 6$, where n is even and the zeros are on the diagonal. Then, the $(n-r) \times (n-r)$, $r \geq 1$, minor of W is*

$$W(n-r) = [(n-1)^{n-r-2^{r-1}} \det M]^{1/2},$$

where

$$M = \begin{bmatrix} n-1-ru_1 & u_1c_{1,2} & u_1c_{1,3} & \dots & u_1c_{1,2^{r-1}} \\ u_2c_{1,2} & n-1-ru_2 & u_2c_{2,3} & \dots & u_2c_{2,2^{r-1}} \\ \vdots & \vdots & & & \vdots \\ u_{2^{r-1}}c_{1,2^{r-1}} & u_{2^{r-1}}c_{2,2^{r-1}} & u_{2^{r-1}}c_{3,2^{r-1}} & \dots & n-1-ru_{2^{r-1}} \end{bmatrix}_{2^{r-1} \times 2^{r-1}},$$

$c_{i,j} = -\tilde{u_i}^T \cdot \tilde{u_j}$, $i, j = 1, \dots, 2^{r-1}$, $\tilde{u_j}$ are the columns of the matrix U_j of Definition 10 appearing u_j times.

For completing the proof of Theorem 12, an algorithm was devised. It operates similarly to the algorithm Minors presented before. However, it dodges the computational difficulties occurring because it requires mainly the computation of inner products of the form $c_{i,j} = -\tilde{u_i}^T \cdot \tilde{u_j}$. Indeed, for a given matrix $W(n, n-1)$ we can directly specify the vectors $\tilde{u_i}$ and the quantities u_i. Then $c_{i,j}$ are computed by simple inner products of the form $c_{i,j} = -\tilde{u_i}^T \cdot \tilde{u_j}$ requiring only $O(l)$ flops. Thus this algorithm achieves the evaluation of minors attaining lower complexity than a computing program. For example, the evaluation of $W(n-3)$ for the weighing matrix $W(24, 23)$ using the proposed algorithm demands a complexity of order 4^3 in order to evaluate the determinant of the 4×4 matrix M, while a program that uses LU factorization for the direct evaluation of the determinant of CC^T would demand a complexity of order 21^3.

After having demonstrated the above results, one can now specify explicitly pivot sequences of specific weighing matrices $W(n, n - k)$ for various n and k. These analytical computations are completed (and facilitated) by the property that the product of the pivots of a matrix is equal to its determinant. The following numerical results are derived in [44].

Lemma 19 *If GE with complete pivoting is applied on a* $W(6, 4)$ *the pivot pattern is* $(1, 2, 2, 4, 4, 6)$.

Lemma 20 *If GE with complete pivoting is applied on a* $W(8, 6)$ *the pivot patterns are* $(1, 2, 2, 4, \frac{3}{2}, 3, 3, 6)$ *or* $(1, 2, 2, \frac{5}{2}, \frac{12}{5}, 3, 3, 6)$.

Lemma 21 *If GE with complete pivoting is applied on a* $W(10, 8)$*, then the possible pivot patterns appearing are those given in Table 2.*

Lemma 22 ([44]) *If GE with complete pivoting is applied on a* $W(12, 10)$*, then the possible pivot patterns appearing are those given in Table 3.*

Theorem 13 ([44]) *The growth factors of the* $W(6, 4)$*,* $W(8, 6)$*,* $W(10, 8)$ *and* $W(12, 10)$ *are* 4*,* 6*,* 8 *and* 10*, respectively.*

The results presented above in this section, together with extensive numerical experiments performed in [44] for calculating growth factors of $W(n, n - k)$ for several n and k, give rise to the growth conjecture for $W(n, n - k)$.

Table 2 All possible pivot patterns of $W(10, 8)$

	Pivot patterns of $W(10, 8)$
1	$(1,2,2,\frac{5}{2},\frac{16}{5},4,2,4,4,8)$ or
2	$(1,2,2,\frac{5}{2},2,\frac{24}{7},2,4,4,8)$ or
3	$(1,2,2,\frac{5}{2},2,4,2,4,4,8)$ or
4	$(1,2,2,\frac{5}{2},\frac{8}{3},4,2,4,4,8)$ or
5	$(1,2,2,3,\frac{8}{3},4,2,4,4,8)$ or
6	$(1,2,2,3,\frac{14}{5},4,2,4,4,8)$ or
7	$(1,2,2,4,2,4,2,4,4,8)$ or
8	$(1,2,2,4,\frac{14}{5},4,2,4,4,8)$ or
9	$(1,2,2,4,\frac{14}{5},4,\frac{8}{3},4,4,8)$ or
10	$(1,2,2,4,\frac{16}{5},4,\frac{8}{3},4,4,8)$

Table 3 All possible pivot patterns of $W(12, 10)$

	Pivot patterns of $W(12, 10)$
1	$(1,2,2,\frac{5}{2},\frac{5}{2},4,3,4,\frac{10}{3},5,5,10)$ or
2	$(1,2,2,\frac{5}{2},\frac{5}{2},5,2,4,4,5,5,10)$ or
3	$(1,2,2,\frac{5}{2},\frac{14}{5},\frac{26}{7},\frac{40}{13},5,\frac{5}{2},5,5,10)$ or
4	$(1,2,2,\frac{5}{2},\frac{14}{5},4,\frac{45}{14},\frac{40}{9},\frac{5}{2},5,5,10)$ or
5	$(1,2,2,\frac{5}{2},3,\frac{10}{3},3,4,\frac{10}{3},5,5,10)$ or
6	$(1,2,2,\frac{5}{2},\frac{18}{5},\frac{28}{9},\frac{45}{14},\frac{40}{9},\frac{5}{2},5,5,10)$ or
7	$(1,2,2,\frac{5}{2},\frac{18}{5},\frac{65}{18},\frac{36}{13},\frac{40}{9},\frac{5}{2},5,5,10)$ or
8	$(1,2,2,3,\frac{5}{2},\frac{10}{3},3,4,\frac{10}{3},5,5,10)$ or
9	$(1,2,2,3,\frac{5}{2},4,3,\frac{40}{9},\frac{5}{2},5,5,10)$ or
10	$(1,2,2,3,\frac{8}{3},\frac{13}{4},\frac{40}{13},\frac{28}{9},\frac{15}{4},\frac{10}{3},5,5,10)$ or
11	$(1,2,2,3,\frac{8}{3},\frac{13}{4},\frac{40}{13},5,\frac{5}{2},5,5,10)$ or
12	$(1,2,2,3,\frac{8}{3},\frac{7}{2},\frac{45}{14},\frac{40}{9},\frac{5}{2},5,5,10)$ or
13	$(1,2,2,3,\frac{8}{3},4,\frac{7}{2},\frac{15}{4},\frac{10}{3},5,5,10)$ or
14	$(1,2,2,3,\frac{8}{3},4,\frac{7}{2},5,\frac{5}{2},5,5,10)$ or
15	$(1,2,2,3,3,\frac{28}{9},\frac{45}{14},\frac{40}{9},\frac{5}{2},5,5,10)$ or
16	$(1,2,2,3,3,\frac{65}{18},\frac{36}{13},\frac{40}{9},\frac{5}{2},5,5,10)$ or
17	$(1,2,2,3,3,4,\frac{10}{3},\frac{10}{3},\frac{5}{2},5,5,10)$ or
18	$(1,2,2,4,2,\frac{7}{2},\frac{45}{14},\frac{40}{9},\frac{5}{2},5,5,10)$ or
19	$(1,2,2,4,2,4,\frac{5}{2},\frac{15}{4},\frac{10}{3},5,5,10)$ or
20	$(1,2,2,4,2,4,\frac{5}{2},5,\frac{5}{2},5,5,10)$ or
21	$(1,2,2,4,\frac{9}{4},\frac{28}{9},\frac{45}{14},\frac{40}{9},\frac{5}{2},5,5,10)$ or
22	$(1,2,2,4,3,\frac{10}{3},3,\frac{10}{3},\frac{5}{2},5,5,10)$

Conjecture 1

The growth conjecture for $W(n, n-k)$

Let W be a CP $W(n, n-k)$. Reduce W by GE. Then, for large enough n,

(i) $g(n, W) = n - k$.
(ii) The three last pivots are equal to $\frac{n-k}{2}$, $\frac{n-k}{2}$, $n-k$.
(iii) Every pivot before the last has magnitude at most $n-k$.
(iv) The first three pivots are equal to $1,\ 2,\ 2$. The fourth pivot can take the values 3 or 4 or $\frac{5}{2}$.

11.1 Sharpe's Results

The first known effort for calculating minors of Hadamard matrices is estimated to be accomplished in 1907 by Sharpe [57]. The essence of these results is summarized in the next Theorem.

Theorem 14 ([57]) *Let H be a Hadamard matrix of order n. The cofactor of any element of H is $\pm n^{\frac{n}{2}-1}$, the sign being the same as the sign of that element. The second minors of H are $2n^{\frac{n}{2}-2}$ or 0, according as the complementary minor is 2 or 0. The third minors of H are $4n^{\frac{n}{2}-3}$ or 0, according as the complementary minor is 4 or 0.*

Theorem 14 actually gives all possible values of $n-j$, $j=1,2,3$, minors of Hadamard matrices according to the determinant of the respective excluded $j \times j$ matrix. Sharpe's idea, which leads to short, elegant proofs, is based on considering a special arrangement of the entries in a Hadamard matrix. It takes appropriately into account the definition $HH^T = nI_n$ by observing that when comparing any pair of rows or columns, there is always an equal number of changes and permanences of sign amongst the corresponding elements. But unfortunately, it doesn't seem to be applicable for calculating minors of orders $n-j$ for $j>3$.

11.2 An Algorithm to Find Minors of Conference Matrices

In [48] was proposed a useful method for finding the $(n-3) \times (n-3)$ minors of skew Hadamard and conference matrices. This strategy can be generalized for finding $(n-j) \times (n-j)$ minors.

Firstly, a useful result is proved, which gives the number of columns having all possible forms that can be contained in the first 3 rows. To that end, the possible upper left 3×3 corners of the conference matrix are examined and for everyone of them the distribution of the columns is determined with the aid of the notion of the matrix $U)j$.

Lemma 23 ([48] The Distribution Lemma for $W(n, n-1)$) *Let W be any $W(n, n-1)$ of order $n > 2$. Then, writing $\varepsilon = (-1)^{\frac{n+2}{2}}$ and with $a, b, c \in \{1, -1\}$ for every triple of rows containing*

$$\begin{matrix} 0 & a & b \\ \varepsilon a & 0 & c \\ \varepsilon b & \varepsilon c & 0 \end{matrix}$$

the number of columns which are

(a) $(1,1,1)^T$ *or* $(-,-,-)^T$ *is* $\frac{1}{4}(n-3-bc-\varepsilon ac-ab)$
(b) $(1,1,-)^T$ *or* $(-,-,1)^T$ *is* $\frac{1}{4}(n-3-bc+\varepsilon ac+ab)$
(c) $(1,-,1)^T$ *or* $(-,1,-)^T$ *is* $\frac{1}{4}(n-3+bc-\varepsilon ac+ab)$
(d) $(1,-,-)^T$ *or* $(-,1,1)^T$ *is* $\frac{1}{4}(n-3+bc+\varepsilon ac-ab)$.

The Distribution Lemma and the Determinant Simplification Theorem (Theorem 9) were utilized in [48] for developing an algorithm that computes the

$(n-j) \times (n-j)$ minors of a conference matrix of order n. The core idea is to consider the following representation of a conference matrix. Any $W = W(n, n-1)$ can be written as

$$W = \begin{bmatrix} M & U_j \\ \varepsilon U_j^T & C \end{bmatrix}, \tag{11}$$

where $M,\ C$ are $j \times j$ and $(n-j) \times (n-j)$ matrices respectively, with diagonal entries all 0, such that $M = \varepsilon M^T$ and $C = \varepsilon C^T$. The elements in the $(n-j)\times(n-j)$ matrix CC^T obtained by removing the first j rows and columns of the weighing matrix W can be permuted to appear in the form

$$CC^T = (n-1)I_{u_1,u_2,\cdots,u_{2^{j-1}}} + a_{ik}J_{u_1,u_2,\cdots,u_{2^{j-1}}},$$

where $(a_{ik}) = (-\underline{u}_i \cdot \underline{u}_k)$, with $\cdot$ the inner product. By the Determinant Simplification Theorem (9) it follows

$$\det CC^T = (n-1)^{n-2^{j-1}-j} \det D,$$

where D, of order 2^{j-1} is given by

$$D = \begin{bmatrix} n-1-ju_1 & u_2a_{12} & u_3a_{13} \cdots & u_za_{1z} \\ u_1a_{21} & n-1-ju_2 & u_3a_{23} \cdots & u_za_{2z} \\ \vdots & \vdots & \vdots & \vdots \\ u_1a_{z1} & u_2a_{z2} & u_3a_{z2} \cdots & n-1-ju_z \end{bmatrix}.$$

Finally, the $(n-j) \times (n-j)$ minor of a $W(n, n-1)$ is the determinant of C, for which it holds

$$\det C = ((n-1)^{n-2^{j-1}-j} \det D)^{1/2}.$$

12 Conclusions and Relevant Open Problems

This expository work presented a wide range of possibilities for computing minors of weighing matrices. These theoretical approaches have their own intrinsic theoretical algebraic beauty and challenge. Additionally they are particularly useful and applicable in view of a renowned problem in numerical analysis, the growth problem. The terminology refers to the study of the growth factor, which arises at the backward error analysis of Gaussian Elimination with complete pivoting. The growth factor appears in the upper bound of the norm of the error matrix. Hence, it constitutes a measure of stability of the method.

For studying these features, one can take advantage of the result (3) that associates pivot magnitudes with minors. The condition (constraint) that the initial matrix in this respect is CP, is actually equivalent to the impact of GE with complete pivoting on the matrix. So, even if a matrix is not CP initially, it is guaranteed that the effect of GE with complete pivoting transforms it into the requested CP form and (3) can be applied with confidence.

So the target is to devise efficient and effective approaches for calculating leading principal minors of weighing matrices. It is important to emphasize that this problem is computationally non trivial and a common implementation of the required algorithms involves high complexity for the following reasons:

- When wanting to compute the growth factor of a weighing matrix W of specific order n and weight $n - k$, one has firstly to generate all possible H-equivalent $W(n, n - k)$.
- Then GECP should be applied separately on each one of the W's. The final upper triangular matrix yields the pivot pattern and moreover the growth factor for every W with respect to Eqs. (1) and (2).
- Finally, GE should be applied again on the respective submatrices of the weighing matrices kept for calculating the determinants of each one with the standard numerical method to that end.

Summarizing, it becomes evident that fast, practical and robust algorithms should be elaborated for calculating the aforementioned minors and substitute the values in (3). At the same time, they should dodge the above numerous challenging computational difficulties. The algorithms were implemented numerically as well as in a symbolical computing environment.

We may encounter the notion of weighing matrices also in diverse disciplines associated with nonlinear analysis. For instance, weighing matrices arise from the derivation of a recursive algorithm identifying nonlinear multivariable systems [4]. In such a study, particular emphasis is laid on the design of a weighing matrix that ensures consistency of the estimated parameters with the input-output data and the noise constraints, and improves convergence properties. Sufficient conditions for local asymptotic convergence of the algorithm can be examined. The proposed algorithm achieves effectiveness demonstrated through a numerical example.

Additionally, if one manages to implement a version of algorithm Minors (or another technique with the same purpose) up to the $n - 15$ case, then it would be possible to obtain an estimation for the maximum determinant of a 15×15 matrix with entries ± 1, which is an unsolved problem so far. Related problems about determinants of $(0, \pm 1)$ matrices can be found in [5, 27, 56].

It is also interesting to mention that the values of the pivots depend on the choice of the maximum element at each elimination step, when there are at least two equal maximum entries in the respective submatrix. More pivoting strategies involving several choices of the maximum entry are described in [15, 31]. For example, it is known that the D-optimal design of order 5 can exist embedded inside an H_{16}, leading to the value of the fifth pivot $p_5 = 3$. We attempt to find the pivot pattern of an H_{16}, which is equivalent to the one constructed with the command hadamard(16) in Matlab. If we use a straightforward selection of the maximum element, i.e. to choose the *first* maximum entry of every respective lower right matrix as pivot, we get $p_5 = 2$. But if we select as pivot the *last* maximum entry we might be led to $p_5 = 3$. Hence, it is challenging to investigate furthermore this phenomenon and to determine the choices of maximum entries that lead to specific pivot values.

In conclusion, the study of the growth factor is an intriguing and important issue in Numerical Analysis because it characterizes the roundoff estimates of Gaussian Elimination and captures the stability properties of the method. Particularly, the investigation of the growth factor for Hadamard matrices led to the formulation of the open Complete Pivoting conjecture for GE [8] and to the publication of several relevant articles. It would be also interesting to study the growth factors for other matrix factorizations, e.g. as it was done recently in [70] for the modified Gram-Schmidt algorithm by deriving upper bounds for growth factors arising at the solution of least squares problems.

Research on the values of minors of Hadamard matrices is ongoing, cf., e.g., [60] and the references therein. This work provides a distribution for any minors of orders j of Hadamard matrices in relation to the minors of orders $n-j$, up to a factor $n^{j-\frac{n}{2}}$. This issue is closely related to the specification of the possible determinant values of matrices with entries ± 1. A survey on this problem and current updates and results can be found in [53], cf. also [10, 44]. Further progress on this issue can help in the study of Cryer's growth conjecture for Hadamard matrices, since Eq. (3) provides a powerful tool for computing pivots in terms of leading principal minors.

Furthermore, recent noteworthy scientific results and applications involving weighing matrices, which are given from diverse viewpoints and without necessarily any reference to their minors and the current presentation, include, for example, [58]. In this work, circulant weighing matrices are under consideration. The study and detection of their existence and their classification is characterized as a "demanding challenge for parallel optimization metaheuristics". Scientific computing, theoretical algebraic techniques and high-performance computational optimization approaches in a parallel computing environment are employed. The respective algorithms are applied on a hard circulant weighing matrix existence problem and give encouraging outcomes. More results on circulant weighing matrices can be found also in [2].

The significance of developing efficient algorithms for the computation of principal minors of a matrix is highlighted, e.g., in [26]. Further similar templates to the ones presented here for weighing matrices can be developed also for other classes belonging to the broad range of orthogonal designs as well, cf., e.g., [11, 20, 22, 23, 45]. For instance, binary Hadamard matrices [45] are just a little studied in the sense of the current presentation. Moreover, in [47] was proposed a novel and highly efficient technique for computing growth factors of Hadamard matrices, which avoids the explicit computation of the pivot pattern. It involves principally the evaluations and appropriate combinations of upper left and lower right principal minors of the matrices under consideration. A similar approach could be also developed for weighing matrices. Further determinant calculations for matrices with elements 0 and 1 are presented in [73].

The methodology presented here for weighing matrices with respect to GECP can be investigated also for other pivoting strategies within GE [52]. For example, the application of the GE method with rook pivoting [18, 31] to special classes of orthogonal matrices and with respect to the computation of minors, pivots and growth factors is unexplored.

A further interesting handling of algebraic properties of weighing matrices is done in [3]. A special type of weighing matrix, called block weighing matrix, is introduced. Motivated by questions arising in optical quantum computing, the authors demonstrate that infinite families of anticirculant and Hankel block weighing matrices can be constructed from known generic weighing matrices. An important remark of this study claims that the existence of block Hankel weighing matrices is parallel to the implementability of specific schemes for optical quantum computing. Furthermore, the consideration of block weighing matrices allows for a more refined classification of combinatorial designs. A basic open problem is to investigate the existence of specific anticirculant and Hankel block weighing matrices that can not be obtained from weighing matrices of smaller orders.

A subject of further research can be the introduction of other theoretical approaches for the evaluation of minors of weighing matrices. Subsequently, the respective computational algorithms should be developed taking into account the high complexity of the problem. For example, in [37] a method was introduced for evaluating minors of matrices. The core idea is based on algebraic eigenvalue properties. The corresponding theoretical tools and numerical techniques gave the results described previously. The algorithms achieve an improvement with a lower complexity than the methods developed before for the same purpose. The respective algorithms were designed and provided results for $W(n, n-1)$, as it was discussed before. The generalizations to $W(n, n-k)$ and other orthogonal designs are matters under consideration. A similar treatment was carried out in [60] for Hadamard matrixes.

In [1] the authors study the determinants of skew-symmetric (± 1)-matrices via a cocyclic approach. These matrices have a specific distribution of the entries ± 1. The specification of the largest possible determinant of such a matrix is an interesting problem. The approach proposed in [1] for handling this question proposes construction of special cocyclic matrices. Also upper bounds on the maximal determinant of skew-symmetric (± 1)-matrices of special orders have been improved. An intermediate useful tool is a procedure deciding whether a given ± 1 matrix is equivalent to a matrix of skew type. An analogous processing could be eventually elaborated for weighing matrices as well with respect to calculations of their minors in regard to the current presentation.

Acknowledgement The work of C.D. Kravvaritis was supported by the Alexander von Humboldt Foundation, Bonn, Germany.

References

1. V. Álvarez, J.A. Armario, M.D. Frau, F. Gudiel, Determinants of $(-1, 1)$-matrices of the skew-symmetric type: a cocyclic approach. Open Math. **13**, 16–25 (2015)
2. K.T. Arasu, K.H. Leung, S.L. Ma, A. Nabavi, D.K. Ray-Chaudhuri, Determination of all possible orders of weight 16 circulant weighing matrices. Finite Fields Appl. **12**, 498–538 (2006)
3. K.T. Arasu, S. Severini, E. Velten, Block weighing matrices. Cryptogr. Commun. **5**, 201–207 (2013)
4. Y. Becis-Aubry, M. Boutayeb, M. Darouach, A parameter estimation algorithm for nonlinear multivariable systems subject to bounded disturbances, in *Proceedings of the American Control Conference*, Denver, vol. 4 (2003), pp. 3573–3578
5. J. Brenner, L. Cummings, The Hadamard maximum determinant problem. Am. Math. Mon. **79**, 626–630 (1972)
6. J.D. Christian, B.L. Shader, Nonexistence results for Hadamard-like matrices. Electron. J. Comb. **11**, #N1 (2004)
7. A.M. Cohen, A note on pivot size in Gaussian elimination. Linear Algebra Appl. **8**, 361–368 (1974)
8. C.W. Cryer, Pivot size in Gaussian elimination. Numer. Math. **12**, 335–345 (1968)
9. B.N. Datta, *Numerical Linear Algebra and Applications*, 2nd edn. (SIAM, Philadelphia, 2010)
10. J. Day, B. Peterson, Growth in Gaussian elimination. Am. Math. Mon. **95**, 489–513 (1988)
11. P. Delsarte, J.M. Goethals, J.J. Seidel, Orthogonal matrices with zero diagonal, II. Can. J. Math. **23**, 816–832 (1971)
12. T.A. Driscoll, K.L. Maki, Searching for rare growth factors using Multicanonical Monte Carlo Methods. SIAM Rev. **49**, 673–692 (2007)
13. A. Edelman, The complete pivoting conjecture for Gaussian elimination is false. Math. J. **2**, 58–61 (1992)
14. A. Edelman, W. Mascarenhas, On the complete pivoting conjecture for a Hadamard matrix of order 12. Linear Multilinear Algebra **38**, 181–187 (1995)
15. P. Favati, M. Leoncini, A. Martiinez, On the robustness of Gaussian elimination with partial pivoting. BIT Numer. Math. **40**, 62–73 (2000)
16. G.E. Forsythe, C.B. Moler, *Computer Solution of Linear Algebraic Systems* (Prentice Hall, Englewood Cliffs, 1967)
17. L.V. Foster, Gaussian elimination with partial pivoting can fail in practice. SIAM J. Matrix Anal. Appl. **15**, 1354–1362 (1994)
18. L.V. Foster, The growth factor and efficiency of Gaussian elimination with rook pivoting. J. Comput. Appl. Math. **86**,177–194 (1997)
19. F.R. Gantmacher, *The Theory of Matrices*, vol. 1 (Chelsea, New York, 1959)
20. A.V. Geramita, J. Seberry, *Orthogonal Designs: Quadratic Forms and Hadamard Matrices* (Marcel Dekker, New York, 1979)
21. A.V. Geramita, J.S. Wallis, Orthogonal designs. III. Weighing matrices. Utilitas Math. **6**, 209–236 (1974)
22. A.V. Geramita, J.M. Geramita, J.S. Wallis, Orthogonal designs. Linear Multilinear Algebra **3**, 281–306 (1975/1976)
23. J.M. Goethals, J.J. Seidel, Orthogonal matrices with zero diagonal. Can. J. Math. **19**, 1001–1010 (1967)
24. G.H. Golub, C.E. Van Loan, *Matrix Computations*, 4th edn. (John Hopkins University Press, Baltimore, 2013)
25. N. Gould, On growth in Gaussian elimination with pivoting. SIAM J. Matrix Anal. Appl. **12**, 354–361 (1991)
26. K. Griffin, M.J. Tsatsomeros, Principal minors, Part I: A method for computing all the principal minors of a matrix. Linear Algebra Appl. **419**, 107–124 (2006)

27. J. Hadamard, Résolution d'une question relative aux déterminants. Bull. Sci. Math. **17**, 240–246 (1893)
28. M. Hall, *Combinatorial Theory* (Wiley, New York, 1986)
29. M. Harwit, N.J.A. Sloane, *Hadamard Transform Optics* (Academic, New York, 1979)
30. A.S. Hedayat, N.J.A. Sloane, J. Stufken, *Orthogonal Arrays* (Springer, New York, 1999)
31. N.J. Higham, *Accuracy and Stability of Numerical Algorithms* (SIAM, Philadelphia, 2002)
32. N.J. Higham, Gaussian elimination. WIREs Comput. Stat. **3**, 230–238 (2011)
33. N.J. Higham, D.J. Higham, Large growth factors in Gaussian elimination with pivoting. SIAM J. Matrix Anal. Appl. **10**, 155–164 (1989)
34. K.J. Horadam, *Hadamard Matrices and Their Applications* (Princeton University Press, Princeton, 2007)
35. R.A. Horn, C.R. Johnson, *Matrix Analysis* (Cambridge University Press, Cambridge, 1985)
36. E. Howard, *Elementary Matrix Theory* (Dover Publications, Mineola, 1980)
37. A. Karapiperi, M. Mitrouli, M.G. Neubauer, J. Seberry, An eigenvalue approach evaluating minors for weighing matrices $W(n, n-1)$. Linear Algebra Appl. **436**, 2054–2066 (2012)
38. H. Kharaghani, B. Tayfeh-Rezaie, A Hadamard matrix of order 428. J. Comb. Des. **13**, 435–440 (2005)
39. C. Koukouvinos, J. Seberry, Weighing matrices and their applications. J. Stat. Plan. Inference **62**, 91–101 (1997)
40. C. Koukouvinos, M. Mitrouli, J. Seberry, Growth in Gaussian elimination for weighing matrices, $W(n, n-1)$. Linear Algebra Appl. **306**, 189–202 (2000)
41. C. Koukouvinos, M. Mitrouli, J. Seberry, An algorithm to find formulae and values of minors of Hadamard matrices. Linear Algebra Appl. **330**, 129–147 (2001)
42. C. Koukouvinos, M. Mitrouli, J. Seberry, Values of minors of an infinite family of D-optimal designs and their application to the growth problem. SIAM J. Matrix Anal. Appl. **23**, 1–14 (2001)
43. C. Kravvaritis, M. Mitrouli, Determinant evaluations for weighing matrices. Int. J. Pure Appl. Math. **34**, 163–176 (2007)
44. C. Kravvaritis, M. Mitrouli, Evaluation of minors associated to weighing matrices. Linear Algebra Appl. **426**, 774–809 (2007)
45. C. Kravvaritis, M. Mitrouli, A technique for computing minors of binary Hadamard matrices and application to the growth problem. Electron. Trans. Numer. Anal. **31**, 49–67 (2008)
46. C. Kravvaritis, M. Mitrouli, The growth factor of a Hadamard matrix of order 16 is 16. Numer. Linear Algebra Appl. **16**, 715–743 (2009)
47. C. Kravvaritis, M. Mitrouli, On the complete pivoting conjecture for Hadamard matrices: further progress and a good pivots property. Numer. Algorithms **62**, 571–582 (2013)
48. C. Kravvaritis, E. Lappas, M. Mitrouli, An algorithm to find values of minors of Skew Hadamard and Conference matrices. Lect. Notes Comput. Sci. **3401**, 375–382 (2005),
49. C. Kravvaritis, M. Mitrouli, J. Seberry, Counting techniques specifying the existence of submatrices in weighing matrices. Lect. Notes Comput. Sci. **3718**, 294–305 (2005)
50. C. Kravvaritis, M. Mitrouli, J. Seberry, On the growth problem for skew and symmetric conference matrices. Linear Algebra Appl. **403**, 83–206 (2005)
51. C. Kravvaritis, M. Mitrouli, J. Seberry, On the pivot structure for the weighing matrix $W(12, 11)$. Linear Multilinear Algebra **55**, 471–490 (2007)
52. M. Olschowka, A. Neumaier, A new pivoting strategy for Gaussian elimination. Linear Algebra Appl. **240**, 131–151 (1996)
53. W.P. Orrick, B. Solomon, Spectrum of the determinant function. Indiana University (2012). http://www.indiana.edu/~maxdet/spectrum.html. Cited 26 January 2017
54. J. Seberry, Weighing Matrices. University of Wollongong (2001). http://www.uow.edu.au/~jennie/sequences.html. Cited 26 January 2017
55. J. Seberry, M. Yamada, Hadamard matrices, sequences and block designs, in *Contemporary Design Theory: A Collection of Surveys*, ed. by D.J. Stinson, J.H. Dinitz (Wiley, New York, 1992), pp. 431–560

56. J. Seberry, T. Xia, C. Koukouvinos, M. Mitrouli, The maximal determinant and subdeterminants of ±1 matrices. Linear Algebra Appl. **373**, 297–310 (2003)
57. F.R. Sharpe, The maximum value of a determinant. Bull. Am. Math. Soc. **14**, 121–123 (1907)
58. D. Souravlias, K.E. Parsopoulos, I.S. Kotsireas, Circulant weighing matrices: a demanding challenge for parallel optimization metaheuristics. Optim. Lett. **10**, 1303–1314 (2016)
59. J.J. Sylvester, Thoughts on inverse orthogonal matrices, simultaneous sign successions, and tessellated pavements in two or more colours, with applications to Newton's rule, ornamental tile-work, and the theory of numbers. Philos. Mag. **34**, 461–475 (1867)
60. F. Szöllősi, Exotic complex Hadamard matrices and their equivalence. Cryptogr. Commun. **2**, 187–198 (2010)
61. V.D. Tonchev, Generalized weighing matrices and self-orthogonal codes. Discrete Math. **309**, 4697–4699 (2009)
62. L. Tornheim, Pivot size in Gauss reduction. Technical Report, California Resources Corporation, Richmond (1964)
63. W. Trappe, L. Washington, *Introduction to Cryptography with Coding Theory*, 2nd edn. (Pearson Education, Upper Saddle River, 2006)
64. L.N. Trefethen, Three mysteries of Gaussian elimination. ACM SIGNUM Newsl. **20**, 2–5 (1985)
65. L.N. Trefethen, D. Bau III, *Numerical Linear Algebra* (SIAM, Philadelphia, 1997)
66. L.N. Trefethen, R.S. Schreiber, Average-case stability of Gaussian elimination. SIAM J. Matrix Anal. Appl. **11**, 335–360 (1990)
67. M. Tsatsomeros, L. Li, A recursive test for P-matrices. BIT Numer. Math. **40**, 404–408 (2000)
68. W. van Dam, Quantum algorithms for weighing matrices and quadratic residues. Algorithmica **34**, 413–428 (2002)
69. W.D. Wallis, A.P. Street, J.S. Wallis, *Combinatorics: Room Squares, Sum-Free Sets, Hadamard Matrices*. Lectures Notes in Mathematics, vol. 292 (Springer, New York, 1972)
70. M. Wei, Q. Liu, On growth factors of the modified Gram-Schmidt algorithm. Numer. Linear Algebra Appl. **15**, 621–626 (2008)
71. J.H. Wilkinson, Error analysis of direct methods of matrix inversion. J. Assoc. Comput. Mach. **8**, 281–330 (1961)
72. J.H. Wilkinson, *The Algebraic Eigenvalue Problem* (Oxford University Press, London, 1965)
73. J. Williamson, Determinants whose elements are 0 and 1. Am. Math. Mon. **53**, 427–434 (1946)
74. S.J. Wright, A collection of problems for which Gaussian elimination with partial pivoting is unstable. SIAM J. Sci. Comput. **14**, 231–238 (1993)
75. R.K. Yarlagadda, J.E. Hershey, *Hadamard Matrix Analysis and Synthesis: With Applications to Communications and Signal/Image Processing* (Kluwer, Boston, 1997)

Robots That Do Not Avoid Obstacles

Kyriakos Papadopoulos and Apostolos Syropoulos

1 Introduction

According to Latombe [9], “the ultimate goal of robotics is to create autonomous robots”. Farber [4] adds that

> …such robots should be able to accept high-level description of tasks and execute them without further human intervention. The input description specifies what should be done and the robot decides how to do it and performs the task. One expects robots to have sensors and actuators.

Typically, robots should be programmed so to be able to plan collision-free motions for complex bodies from some point A to another point B while having a collection of static obstacles in between. This task is called *motion planning*. Naturally, motion planning is very interesting but there are many cases where this is not even desirable. For example, a rover moving on the surface of a planet should be able to go above obstacles or to even pass through obstacles.

Dynamical systems are characterized by equations that describe their evolution. A dynamical system is called *linear* when its evolution is a linear *process*. A process is linear when a change in any variable at some initial time produces a change in some variable at some later time, however, if the initial variable changes n times, then the new variable will change n times at the later time. In other words, any change propagates without any alterations. Any system that is not linear is called a *nonlinear* dynamical system [14]. A basic characteristic of these systems is that any

K. Papadopoulos (✉)
Department of Mathematics, Kuwait University, Safat, Kuwait

A. Syropoulos
Greek Molecular Computing Group, Xanthi, Greece

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_20

change in a variable at some initial moment leads to a change to some variable at a later time, which is not proportional to the initial change. For example, the *logistic map* [12]

$$x_{n+1} = rx_n(1 - x_n),$$

where $x_n \in [0, 1]$ is the magnitude of population in generation n and x_{n+1} the magnitude of population at generation $n + 1$, is a typical example of an equation that describes a nonlinear system. In this case, the system is the population of some species and the dynamics the changes from one generation to another.

Although a robotic system can be either linear or nonlinear, it seems that nonlinear systems are more interesting in terms of applications. A robotic system is called *nonlinear* when its control is not nonlinear. In particular, a control system is called nonlinear when it contains at least one nonlinear component [15]. For example, a *soft robot* [8], that is, a robotic system that consists of several deformable spherical components, is a nonlinear robotic system [5]. Unlike (some) rigid robots, a soft robot can in general go through or above an obstacle. Consider a robot, rigid or soft, that moves on a specific path. Assume that we assign to each obstacle which is on this path a penetrability degree. Then, the degree to which the robot will not deviate from its path to avoid the obstacle will depend on this degree. If the robot can go through the obstacle or above it, then we have a nonlinear system moving on a "vague" environment. Thus one can say that the motion of a soft robot can be described also by using fuzzy "mathematics" (i.e., a very popular mathematical formulation of vagueness).

The central problem of robotics is how to go from point A to point B. As explained above, avoiding obstacles by deviating from a "predetermined" path is the "classical" way to solve this problem. However, this is not an interesting problem for us. We are interested in systems that can use an extended form of the motion planning algorithm able to describe robots tat go through or above obstacles. But first, let us examine what is the "classical" motion planning algorithm.

2 Obstacle Avoiding: An Up-to-Date Mathematical Formulation

Given a vehicle V, a starting point A (usually called an initial configuration) and an ending point B (called a final configuration), one can form the set P of all paths that V can follow, starting from A and ending in B. Clearly, one can define a number of fuzzy subsets of P, for example, the fuzzy subset of easy paths, the fuzzy subset of smooth paths, etc. Obviously, the problem is how to chose a path in order to go from A to B. This problem is called the *motion planning problem* [9].

A *motion-planing algorithm* [9] is a solution to the motion planning problem. Before giving a formal definition to this problem and to its solution, we describe these notions intuitively. The main task is to find a path starting at a point A and ending at point B. The path has to avoid *collisions* with a known set of stationary obstacles. At any given moment, a robot moving on this path is on a specific *robot configuration* (i.e., a point of this path). In order to solve this problem one needs a geometric description of both the vehicle and the space where the vehicle moves. The *configuration* q of a vehicle is a specification of the positions of all vehicle points relative to a fixed coordinate system. The *configuration space* is the space of all possible configurations.

Assume that $W \subset \mathbb{R}^3$ is the configuration space on which the vehicle moves, where $\mathbb{R}^3$ is the Euclidean space of dimension 3, and denote by $\mathcal{O} \in W$ the set of all possible obstacles that the vehicle can meet. Such obstacles will be presented in terms of neighborhoods in $\mathbb{R}^3$. The expression $\mathcal{A}(q)$ is used to denote that the vehicle is in configuration $q \in C \subseteq W$. Then,

$$C_{\text{free}} = \left\{ q \in C \,\middle|\, \mathcal{A}(q) \cap \mathcal{O} = \emptyset \right\}$$

$$C_{\text{obs}} = C / C_{\text{free}}.$$

Let q_S be the initial configuration and q_G the final configuration. Then, the *motion planning problem* is the process of finding a continuous path $p : [0, 1] \to C_{\text{free}}$, where $p(0) = q_S$ and $p(1) = q_G$.

One approaches the motion planning problem using different tools and methodologies and, thus, there are different solutions to it. For example, Lozano-Pérez [10] presented a *simple* solution, Ashiru and Czarnecki [1] discussed motion planning using genetic algorithms and Farber [4] presented a *probabilistic* solution. Most of all these approaches assume that the vehicle should always avoid obstacles, but there has not been a study of cases where the vehicle can pass through (penetrate) an obstacle.

2.1 A Mathematical Formulation

We will use Farber's [4] notation and mathematical description of robot motion planning algorithm. For topological notions like path-connected spaces, compact-open topology, etc., see [3].

Let X be a path-connected topological space and denote by PX the space of all continuous paths. PX is supplied with the compact open topology. Consider the map $\pi : PX \to X \times X$, which assigns to a path the pair $(\gamma(0), \gamma(1))$ of the so-called initial-final configurations. π is a fibration in the sense of Serre.

Definition 1 A *motion planning algorithm* is a section $s : X \times X \to PX$ of fibration, that is, $\pi \circ s = 1_{X \times X}$.

One of Farber's research goals was to predict the character of instabilities of the behavior of the robot, knowing several topological properties of the configuration space, such as its cohomology algebra. Here we will not concern ourselves with this approach. We will stick in Farber's declaration that there may exist a better mathematical notion of a configuration space, describing a partially known topological space, whose (geometric and topological) properties are being gradually revealed. We believe that fuzzy set theory is the key tool for this.

Farber introduced four numerical invariants $\mathrm{TC}_i(X)$, $i = 1, 2, 3, 4$, measuring the complexity of the problem of navigation of a robot configuration space. These invariants coincide for "good" spaces, such us for simplicial polyhedra. We will now present $\mathrm{TC}_4(X)$, for our purposes, since it is linked with random motion planning algorithms.

Definition 2 A random n-valued path σ, on a path-connected topological space X, starting at $A \in X$ and ending at $B \in X$ is given by an ordered sequence of paths $\gamma_1, \cdots, \gamma_n \in PX$ and an ordered sequence of real numbers $p_1, \cdots, p_n \in [0, 1]$, such that each $\gamma_j : [0, 1] \to X$ is a continuous path in X starting at $A = \gamma_j(0)$ and ending at $B = \gamma_j(1)$, such that $p_j \geq 0$ and $\Sigma_{i=1}^{n} \gamma_i = 1$.

The notation $P_n X$, of Farber, refers to the set of all n-valued random paths in X. This set is a factor-space of a subspace of the Cartesian product of n copies of $PX \times [0, 1]$.

Definition 3 $\mathrm{TC}_4(X)$ is defined as the minimal integer n, such that there exists an n-valued random motion planning algorithm $s : X \times X \to P_n X$.

Remark 2.1 It has been proved that $\mathrm{TC}_{n+1}(X) = \mathrm{cat}(X^n)$, for $n \geq 1$, where $\mathrm{cat}(X^n)$ is the Lusternik-Schnirelmann category [11]. These categories have been used to solve problems in nonlinear analysis (e.g., see [2]).

2.2 *Remarks on This Formulation*

No one can doubt the usefulness of Farber's approach, both in the field of Topology and in Robotics. The instabilities in the robot motion planning algorithm are linked to topological invariants and the universe where the robot moves is seen through the eyes of a topologist who sees configuration spaces. When it comes to engineering though, an interpretation of the invariant $TC_4(X)$ is tough. What does it mean for a vehicle to take a random path? Is it better to talk about a plausible path? Moreover, instead of bypassing obstacles, can we assume that a robot can go through obstacles?

In what follows, we describe a *fuzzy* motion planing problem and explain how it can be solved. These ideas are explained practically and we conclude with some questions and problems related to this approach.

3 Questioning an Even More Theoretical Approach to Motion Planning Problem

Here we ask for the possibility of investigating purely topological properties of robot motion planning algorithms via function spaces, based on the study in [6] and on the results by Farber. Considering a function space $\mathcal{F}(X, Y)$, there are several topological problems one can study. Knowing topological properties of X (or Y), what are the topological properties of $\mathcal{F}(X, Y)$ and vice versa.

Let X be an arbitrary topological space. Let $PX = \mathcal{C}([0, 1], X)$ be the function space of all continuous paths $\gamma : [0, 1] \to X$, supplied with the compact-open topology. Let $\pi : PX \to X \times X$ be the map which assigns to a path γ the pair $(\gamma(0), \gamma(1)) \in X \times X$ of the so-called "initial-final configurations". Consider the function space $\mathcal{F}(PX, X \times X)$. A motion planning algorithm is a map $s : X \times X \to PX$, such that $\pi \circ s = 1_{X \times X}$. Consider the function space $\mathcal{F}^M(X \times X, PX)$, consisting of motion planning algorithms. Notice that this is a subspace of the function space $\mathcal{F}(X \times X, PX)$.

Question 1

Farber questions under what conditions there exist motion planing algorithms which are continuous, and gives an answer through contractibility. More generally, add (the minimum number of) topological conditions on the function space $\mathcal{C}(X \times X, PX)$, so that its functions to be motion planning algorithms, and thus study topological properties of the function space $\mathcal{C}^M(X \times X, PX)$ of continuous motion planning algorithms. Here we should remark that we did not recommend X to be path-connected (which practically means that one can fully control the system by bringing it to an arbitrary state from a given state) as an initial condition.

Question 2

Start with a topological space X, as the configuration space of a mechanical system, with no explicit information about its local or global topological properties. Apply Step 0 to Step n of the construction given in [6], to the motion planning algorithms space $\mathcal{F}(X \times X, PX)$. Study the possibility for the existence of a minimal integer n "revealing as much as possible topological information about X". This will give a partial answer to Farber's question on robot motion planning algorithms, on whether there exists a way to study very complex configuration spaces which are gradually revealing their topological properties.

Question 3

Given answers to our Question 1, a further theory can be developed, studying the topological complexity of tame motion planning algorithms, in the language of function spaces (see [4])

Question 4

If a space X is path-connected, one can "fully control it", in a sense that for any two fixed points there is a path joining them. One could define a topological space, so that for any two points A and B there exists a *linear ordered topological space* (lots) starting from A and ending at B, and this would generalize path-connected spaces and furthermore motion planning algorithms.

Can one achieve this in a different way rather than refining the definition of a continuous path γ, by adding the extra property that the path γ should be also order preserving (taking in [0, 1] the natural order $<$)?

One can consider the space of all such lots on X, say PX, mapped to $X \times X$ as a fibration π, and define a section $s : X \times X \to PX$, such that $\pi \circ s = id_{X \times X}$. One could then study its Schwartz genus, as a notion of a topological complexity of X, and link notions of order theory and general topology to algebraic topological ideas.

There will be a problem if one considered an arbitrary lots. Consider for example the lots consisting of just two points can be mapped into any space

X with two points A and B and that mpa will be a homeomorphic embedding, if and only if X is T_1. One does not want this sort of "teleporting" behavior to be possible, that perhaps one wants there to be many points linking A to B along what "resembles a path". A general way to achieve this is to require that the lots to be a dense order. If one follows this route, it would be most natural to require paths to be closed subsets and the map to be a homeomorphic embedding. Alternatively, one could fix a lots L that is to work for all pairs of points in the space:

1. when $L = \{0, 1\}$ then we have a T_1 space and
2. when $L = [0, 1]$, then we have a path-connected space.

What if $Y = \mathbb{Q} \cap [0, 1]$? What if Y is the Cantor set C? What if $Y = \omega + 1$. In either cases, the "interesting" spaces are going to be totally disconnected

4 Further Topological Remarks

For a more detailed discussion, see [4]. Here we add a few more questions of topological nature.

Consider a path-connected topological space X. A random n-valued path σ, in X, which starts at point A and ends at point B, is given from a sequence of paths $\gamma_1, \gamma_2, \ldots, \gamma_n$ which belong to PX (the space of all continuous paths on X) and a sequence of real numbers $p_1, p_2, \ldots, p_n$ in [0, 1], such that every path $\gamma_j : [0, 1] \to X$ is continuous, where $\gamma_j(0) = A$ and $\gamma_j(1) = B$ and also $p_j \geq 0$ and $p_1 + p_2 + \cdots + p_n = 1$. From the third Axiom of probability theory, one induces that $\sigma = p_1\gamma_1 + p_2\gamma_2 + \cdots + p_n\gamma_n$.

Consider now the map $\pi : P_nX \to X \times X$, where P_nX denotes the set of all random n-valued paths on X. An n-valued random algorithm is a map $s : X \times X \to PX$, such that $\pi \circ s = 1_{X \times X}$.

In other words, if one considers the pair (A, B) in $X \times X$ (input), the output is an ordered probability distribution $s(A, B) = p_1\gamma_1 + p_2\gamma_2 + \cdots + p_n\gamma_n$, that is the algorithm s induces the path γ_j with probability p_j.

A first question, is which probability distributions are outputs of such motion planning algorithms. It would be of a theoretical interest to characterize probability distributions via motion planning algorithms. What about if the number of paths is not countable? If one can define such motion planning algorithm, then what kind of probability distribution can one expect as an output? This is a good point to pass into the next section, which is the approach to the motion planning problem through fuzzy logic.

5 Obstacle Avoiding: A Fuzzy Logic Approach

A fuzzy motion planning problem is a problem that asks how a vehicle can move from a point A to a point B by possibly going through/climb over/penetrate and so on, a number of obstacles, instead of avoiding them. All obstacles, which are represented mathematically by neighborhoods, are associated with a *traversal difficulty degree* that specifies how difficult it is to go over a specific obstacle. This degree is a number drawn from [0, 1] and when it is equal to 1 for a given obstacle O, this implies that O is actually not an obstacle. On the other hand, a traversal difficulty degree equal to 0 means that it is impossible to go over O, so the robot will have to find ways to avoid it.

Definition 4 A fuzzy continuous path is a map $p^{\lambda,\ell} : [0, 1] \to C$ that goes over obstacles $O_1, \ldots, O_n \in C_{\text{obs}}$, where the traversal difficulty degree of each obstacle O_i is λ_i, has a plausibility degree that equals $\lambda = \min_{i=0} \lambda_i$ and its length is ℓ.

Clearly, the smaller the value of λ is, the less plausible a specific path is.

Figure 1 depicts a terrain with some obstacles. The vehicle's task is to go from A to B. Obviously, the dotted path is one that avoids all obstacles but it is quite long. On the other hand, the straight line is a path that goes over three obstacles but it is the shortest possible path. Thus, the ideal path is the one that it will be as short as possible and as easy to traverse as possible.

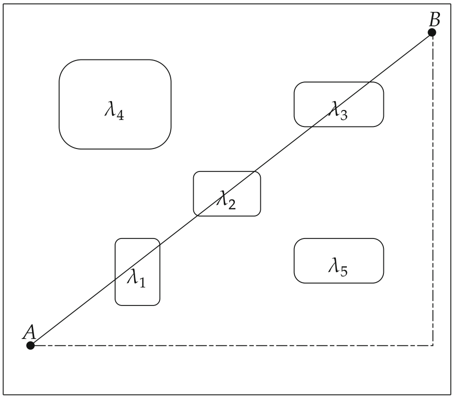


Fig. 1 The problem of moving a vehicle from A to B and two possible solutions

Definition 5 A *fuzzy n-valued path* σ, on X, starting at $A \in X$ and ending at $B \in X$ is an ordered sequence of paths $p_1^{\lambda_1,\ell_1}, p_2^{\lambda_2,\ell_2}, \cdots, p_n^{\lambda_n,\ell_n} \in PX$, where

$$\sigma = \min_{\ell_i} \max_{\lambda_i} p_i^{\lambda_i,\ell_i}, \forall i = 1, 2, \ldots, n.$$

Assume that $P_n X$ is the set of all fuzzy n-valued paths. Then, the function:

$$\pi : P_n X \to X \times X$$

maps to a fuzzy path its starting and end points.

Definition 6 An *n-valued fuzzy motion planning algorithm* is defined as the map:

$$s : X \times X \to P_n X.$$

Thus, the algorithm is a twofold process: first it identifies n distinct paths and it then chooses the most plausible one, not just someone “in random”.

Remark 5.2 The function s is a continuous section of the fibration π.

Having given the above definition of an n-valued fuzzy motion planning algorithm, we now have a clearer picture of how one can define an invariant, similar to $\text{TC}_4(X)$ but more realistic, describing its navigational complexity. Let us call such an invariant $\text{TC}_4^*(X)$. This invariant will depend on both parameters λ and ℓ of Definition 5. So, it will be sufficient to declare it as the “smallest integer n, such that an n-valued fuzzy motion planning algorithm exists”. $\text{TC}_4^*(X)$ certainly describes a wider range or properties of the configuration space. Sometimes, in real situations, it will be better to go through an obstacle, e.g. a vehicle towards water, provided that in such a way ℓ is small, even if λ is small too. A mission running out of time, for example, will put a vehicle into such a risk. In other cases, it might be better for ℓ to be big in order λ to be big, too; for instance, a short distance and a harsh obstacle might put the vehicle into a great risk or might force it to spend a sufficiently big amount of fuel, etc.

5.1 An Example

Imagine that a vehicle, like NASA's Curiosity, is on the surface of planet Mars. Assume that this vehicle can recognize obstacles and it can assess whether it is possible to go over an obstacle or not. For example, the rover might have access to an on-board databank with pictures of obstacles, which have been rated somehow (e.g., by a human expert), and using some sort of object recognition algorithm, then it can assign traversal difficulty degrees to various objects and so it can "deduce" whether a specific path is traversable or not. More generally, the vehicle can perform this action several times to find different traversable paths and to choose the best path. Of course, the system should be able to retract and make another choice since it is quite possible that some initial estimation was more vague than expected.

6 Soft Robots or Fuzzy Motion Planning Algorithms?

On the one hand each obstacle in the path of a robot can be associated with a number that will show to what extend it is possible to go through or above the obstacle but on the other hand we have soft robots that are able to go through obstacles. What is really missing here is that even for soft robots it would not be absolutely sure that one can go through a specific obstacle. Thus even for soft robots, each obstacle should be associated with a number whose value would indicate to what degree it is possible to go through it. In different words, the behavior of soft robots can be better described with the use of fuzzy set theory. Let us roughly describe how this can be realized.

First we chose the path our robot with follow. Then we assign to each obstacle an "absolute" traversal degree, as if our robot is a rigid one. Depending on the shape of the robot and how flexible it is, we modify the absolute traversal degree so to take into account the capabilities of the soft robot. The modified traversal degrees can be used to define a fuzzy motion planning algorithm. The interest thing here is that the dynamics of the robot are nonlinear and we can use fuzzy sets to described a motion planning algorithm.

7 Conclusions and Open Questions

After describing the motion planning problem problem, we briefly discussed a more "realistic" solution and commented on its unsuitability. Next, we presented a formulation of the problem that uses "vagueness" and proposed a solution that makes use of fuzzy set theory. The result is more natural as it coincides with the procedure that humans follow in order to choose the most suitable path. We then gave a first comparison of the fuzzy formulation with that one that uses soft robots.

Here we list a list of open problems which, in our own opinion, are interesting both from a theoretical perspective as well as in applications.

1. Implement the methodology given in the section "Obstacle Avoiding: a Fuzzy Logic approach" with simulation(s) and (an) experiment(s), and see how it works in practice, comparing it with a similar methodology referring to soft-robots.
2. Nonlinear analysis has been used to analyze fuzzy systems (e.g., see [7]). Also, tools used to analyze fuzzy systems have been used to analyze nonlinear systems (e.g., see [13]). The question is: Can use use both methodologies to assist us to build and test a flexible robot?

References

1. I. Ashiru, C. Czarnecki, Optimal motion planning for mobile robots using genetic algorithms, in *IEEE/IAS International Conference on Industrial Automation and Control, 1995 (I A & C'95)* (Cat. No. 95TH8005) (1995), pp. 297–300
2. F.E. Browder, Lusternik-schnirelman category and nonlinear elliptic eigenvalue problems. Bull. Am. Math. Soc. **71**(4), 644–648 (1965)
3. R. Engelking, *General Topology* (Heldermann, Berlin, 1989)
4. M. Farber, Topology of robot motion planning, in *Morse Theoretic Methods in Nonlinear Analysis and in Symplectic Topology*, ed. by P. Biran, O. Cornea, F. Lalonde. NATO Science Series II: Mathematics, Physics and Chemistry, vol. 217 (Springer, Berlin, 2006), pp. 185–230
5. Y. Fei, X. Shen, Nonlinear analysis on moving process of soft robots. Nonlinear Dyn. **73**(1), 671–677 (2013)
6. D. Georgiou, A. Megaritis, K. Papadopoulos, V. Petropoulos, A study concerning splitting and jointly continuous topologies on $C(Y, Z)$. Sci. Robot. **39**(3), 363–379 (2016)
7. D.F. Jenkins, K.M. Passino, An introduction to nonlinear analysis of fuzzy control systems. J. Intell. Fuzzy Syst. **7**, 75–103 (1999)
8. C. Laschi, B. Mazzolai, M. Cianchetti, Soft robotics: technologies and systems pushing the boundaries of robot abilities. Sci. Robot. **1**(1) (2016). https://doi.org/10.1126/scirobotics.aah3690
9. J.-C. Latombe, *Robot Motion Planning* (Springer, New York, 1991)
10. T. Lozano-Pérez, A simple motion-planning algorithm for general robot manipulators. IEEE J. Robot. Autom. **3**(3), 224–238 (1987)
11. G. Lupton, J. Scherer, Topological complexity of H-spaces. Proc. Am. Math. Soc. **141**(5), 1827–1838 (2013)
12. R.M. May, Simple mathematical models with very complicated dynamics. Nature **261**, 459–467 (1976)
13. F. Mei, Z. Man, T. Nguyen, Fuzzy modelling and tracking control of nonlinear systems. Math. Comput. Model. **33**(6), 759–770 (2001)
14. S. Sastry, *Nonlinear Systems: Analysys, Stability, and Control* (Springer, New York, 1999)
15. J.J.E. Slotine, W. Li, *Applied Nonlinear Control* (Prentice Hall, Englewood Cliffs, 1991), p. 07632

On the Exact Solution of Nonlinear Integro-Differential Equations

I. N. Parasidis and E. Providas

1 Introduction

Differential equations and integral equations are mainly employed to model physical phenomena and processes in most disciplines of science, engineering and economics. Integro-differential equations are those equations which contain both differential and integral operators. They appear in modeling many situations in areas such as mechanics [7], electromagnetic theory [3], population dynamics [4], pharmacokinetic studies [13], forestry [6] and many others [10]. The nonlinear integro-differential equations are characterized by the fact that the integrand is a nonlinear function of the unknown function and its derivatives, and that they can have many solutions which can be complex in addition to the real ones. For such equations with a high degree of complexity, usually numerical methods and iterative techniques are utilized to find an approximate solution as it can be verified in the vast literature on the subject, see for example [1, 2] and [8], and the references therein.

Exact solutions may be obtained for a class of nonlinear integro-differential equations where the integrand can be factored in kernels which are degenerate or separable. The procedure used in this case always leads to a nonlinear system of algebraic (transcendental) equations [12, 15]. This system has to be solved exactly and then each solution is properly exploited to produce the corresponding solution to the integro-differential equation in closed form. This in general is difficult and sometimes it may be even impossible to solve the nonlinear system of algebraic equations, while the whole process requires tedious algebra. Therefore, care should

I. N. Parasidis
Department of Electrical Engineering, TEI of Thessaly, Larissa, Greece
e-mail: paras@teilar.gr

E. Providas (✉)
Department of Mechanical Engineering, TEI of Thessaly, Larissa, Greece
e-mail: providas@teilar.gr

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_21

be taken to confine the number of the nonlinear algebraic equations to a minimum and to organize the solution process in an efficient way. The main benefit in pursuing the exact methods for solving nonlinear integro-differential equations, apart from producing the solution in an explicit form, is that they can provide all possible solutions compared to analytical and numerical techniques which can deliver only one approximate, or exact in some cases, solution [14].

In this paper, by using the case of Fredholm nonlinear integro-differential equations as a vehicle, we develop a method for constructing exact solutions to problems involving in general a nonlinear operator **B** defined as a perturbation of a linear correct operator $\widehat{A}$ with linear bounded functionals and nonlinear continuous functionals. The operator $\widehat{A} : X \to Y$, where $X,\ Y$ are complex Banach spaces, is correct if $R(\widehat{A}) = Y$ and its inverse $\widehat{A}^{-1}$ exists and is continuous. The proposed technique is an advancement of the extension operator method introduced by the authors for the exact solution of linear integro-differential equations [10]. The method is applicable to several types of nonlinear problems, it is easily programmable in a computer algebra system and it is suitable for large scale problems.

The rest of the paper is organized in four sections. In Sect. 2, the nonlinear Fredholm integro-differential equations are formulated in an operator form as outlined above. In Sect. 3, the extension operator method for nonlinear problems is expounded. Several example problems are considered in Sect. 4 to demonstrate the efficiency of the method. Finally, some conclusions are quoted in Sect. 5.

2 Nonlinear Integro-Differential Equations of Fredholm Type

Nonlinear Fredholm integro-differential equations of the second kind assume the general form

$$\sum_{\kappa=0}^{\nu} p_\kappa(x)u^{(\kappa)}(x) = f(x) + \int_\alpha^\beta K\left(x, t, u(t), u^{(1)}(t), \dots, u^{(\nu)}(t)\right) dt, \tag{1}$$

with the initial conditions

$$u^{(\kappa)}(\alpha) = \alpha_\kappa, \quad 0 \le \kappa \le \nu - 1, \tag{2}$$

or the boundary conditions

$$u^{(\kappa)}(\alpha) = \alpha_\kappa, \quad u^{(\iota)}(\beta) = \beta_\iota, \quad 0 \le \kappa,\ \iota \le \nu - 1, \quad 0 \le \kappa + \iota \le \nu - 1, \tag{3}$$

or nonlocal conditions, e.g. $u(\alpha) = cu(\beta)$ with c being an arbitrary constant, where $p_\kappa(x),\ \kappa = 0, 1, \dots, \nu$ and $f(x)$ are continuous functions on $[\alpha, \beta]$, $u(x)$

is the unknown function to be determined, $u^{(\kappa)}(x)$ denotes the κth derivative of $u(x)$, the integrand $K\left(x, t, u, u^{(1)}, \ldots, u^{(\nu)}\right)$ is a continuous nonlinear function of $u,\ u^{(1)},\ \ldots,\ u^{(\nu)}$ and α_κ and β_ι are given constants that determine the initial or boundary conditions. An interesting class of nonlinear integro-differential equations is that when the integrand in Eq. (1) can be factored as

$$K\left(x, t, u(t), u^{(1)}(t), \ldots, u^{(\nu)}(t)\right) = \sum_{j=1}^{n} \breve{K}_j(x,t) F_j\left(t, u(t), u^{(1)}(t), \ldots, u^{(\nu)}(t)\right), \quad (4)$$

or

$$\begin{aligned} K\left(x, t, u(t), u^{(1)}(t), \ldots, u^{(\nu)}(t)\right) &= \sum_{i=1}^{m} \bar{K}_i(x,t) P_i\left(t, u(t), u^{(1)}(t), \ldots, u^{(\nu)}(t)\right) \\ &+ \sum_{j=1}^{n} \breve{K}_j(x,t) F_j\left(t, u(t), u^{(1)}(t), \ldots, u^{(\nu)}(t)\right), \quad (5) \end{aligned}$$

where $P_i\left(t, u, u^{(1)}, \ldots, u^{(\nu)}\right),\ i = 1, \ldots, m$ and $F_j\left(t, u, u^{(1)}, \ldots, u^{(\nu)}\right),\ j = 1, \ldots, n$ are linear and nonlinear functions of $u,\ u^{(1)},\ \ldots,\ u^{(\nu)}$, respectively. The kernels $\bar{K}_i(x,t)$ and $\breve{K}_j(x,t)$ are considered to be separable and with no loss of generality we may assume

$$\begin{aligned} \bar{K}_i(x,t) &= g_i(x) h_i(t), \quad i = 1, \ldots, m\,, \\ \breve{K}_j(x,t) &= q_j(x) r_j(t), \quad j = 1, \ldots, n\,. \end{aligned} \quad (6)$$

Substituting (6) in Eq. (5) and then into Eq. (1), we have

$$\begin{aligned} \sum_{\kappa=0}^{\nu} p_\kappa(x) u^{(\kappa)}(x) = f(x) &+ \sum_{i=1}^{m} g_i(x) \int_\alpha^\beta h_i(t) P_i\left(t, u(t), u^{(1)}(t), \ldots, u^{(\nu)}(t)\right) dt \\ &+ \sum_{j=1}^{n} q_j(x) \int_\alpha^\beta r_j(t) F_j\left(t, u(t), u^{(1)}(t), \ldots, u^{(\nu)}(t)\right) dt. \quad (7) \end{aligned}$$

Furthermore, we assume that Eq. (7) is accompanied by the homogeneous initial conditions

$$u^{(\kappa)}(\alpha) = 0, \quad 0 \le \kappa \le \nu - 1, \quad (8)$$

or the homogeneous boundary conditions

$$u^{(\kappa)}(\alpha) = 0, \quad u^{(\iota)}(\beta) = 0, \quad 0 \le \kappa,\ \iota \le \nu - 1, \quad 0 \le \kappa + \iota \le \nu - 1. \quad (9)$$

Notice that nonhomogeneous conditions can be converted to homogeneous ones by making the substitution $u(x) = v(x) + z(x)$, where $z(x) \in \mathbf{C}^{\nu-1}[\alpha, \beta]$ satisfying the given conditions, and solving the problem for $v(x)$. For example, in the case of Eq. (7) with the nonhomogeneous conditions (2), we can define the function $z(x) = \sum_{\kappa=0}^{\nu-1} \frac{a_\kappa}{\kappa!}(x - \alpha)^\kappa$. Correspondingly, for Eq. (7) with, for example, the nonhomogeneous boundary conditions $u(\alpha) = \alpha_0$ and $u(\beta) = \beta_0$, we may specify the function $z(x) = \alpha_0 \frac{\beta-x}{\beta-\alpha} + \beta_0 \frac{x-\alpha}{\beta-\alpha}$.

Let us now define the linear correct operator $\widehat{A} : \mathbf{C}[\alpha, \beta] \to \mathbf{C}[\alpha, \beta]$ by

$$\begin{aligned} \widehat{A}u(x) &= \sum_{\kappa=0}^{\nu} p_\kappa(x)u^{(\kappa)}(x), \\ D(\widehat{A}) &= \{u \in \mathbf{C}^\nu[\alpha, \beta] : u^{(\kappa)}(\alpha) = 0,\ 0 \leq \kappa \leq \nu - 1\}, \end{aligned} \tag{10}$$

if initial conditions are specified or

$$\begin{aligned} D(\widehat{A}) = \{u \in \mathbf{C}^\nu[\alpha, \beta] : u^{(\kappa)}(\alpha) = 0,\ u^{(\iota)}(\beta) = 0, \\ 0 \leq \kappa,\ \iota \leq \nu - 1,\ 0 \leq \kappa + \iota \leq \nu - 1\}, \end{aligned} \tag{11}$$

when boundary conditions are prescribed. Additionally, let the linear bounded functionals $\psi_i \in (\mathbf{C}^\nu[\alpha, \beta])^*$ and the continuous nonlinear functionals $\phi_j : \mathbf{C}^\nu[\alpha, \beta] \to \mathbf{C}$ defined as follows

$$\begin{aligned} \psi_i(u) &= \int_\alpha^\beta h_i(t)P_i\left(t, u(t), u^{(1)}(t), \ldots, u^{(\nu)}(t)\right) dt, \quad i = 1, \ldots, m, \\ \phi_j(u) &= \int_\alpha^\beta r_j(t)F_j\left(t, u(t), u^{(1)}(t), \ldots, u^{(\nu)}(t)\right) dt, \quad j = 1, \ldots, n, \end{aligned} \tag{12}$$

Finally, designate $\Psi = col(\psi_1, \ldots, \psi_m)$, $\Phi = col(\phi_1, \ldots, \phi_n)$, $g = (g_1(x), \ldots, g_m(x))$ and $q = (q_1(x), \ldots, q_n(x))$ and let $f(x) \in \mathbf{C}[\alpha, \beta]$. We may now write the νth-order integro-differential equation of Fredholm type in Eq. (7) under the homogeneous initial (8) or boundary conditions (9) in the operator form

$$\mathbf{B}u(x) = \widehat{A}u(x) - g\Psi(u) - q\Phi(u) = f(x), \quad D(\mathbf{B}) = D(\widehat{A}), \tag{13}$$

where $\mathbf{B} : \mathbf{C}[\alpha, \beta] \to \mathbf{C}[\alpha, \beta]$ is defined as a perturbation of the linear correct operator $\widehat{A}$ with the vectors of linear functionals Ψ and the nonlinear functionals Φ. For completeness we quote that the operator $\mathbf{B}$ is an extension of the minimal operator A_0 defined by

$$A_0u = \widehat{A}u, \quad D(A_0) = \{u \in D(\widehat{A}) : \Psi(u) = 0,\ \Phi(u) = 0\}, \tag{14}$$

as it is stated in [9] and [5].

3 Extension Operator Method for Nonlinear Problems

In a recent work [10], the authors of the present article considered the linear operator $B : X \to Y$,

$$Bu = \widehat{A}u - g\Psi(u), \quad D(B) = D(\widehat{A}), \tag{15}$$

where X, Y and Z are complex Banach spaces, $\widehat{A} : X \to Y$ is a linear correct operator with $D(\widehat{A}) \subset Z \subseteq X$, $\Psi = col(\psi_1, \ldots, \psi_m)$ is a vector of complex-valued bounded linear functionals on Z, $g = (g_1, \ldots, g_m)$ is a vector with $g_i \in Y$, $i = 1, \ldots, m$ and $u \in D(\widehat{A})$, and provided the exact solution for the linear problem

$$Bu = \widehat{A}u - g\Psi(u) = f, \quad f \in Y. \tag{16}$$

In particular, it has been shown [10, Theorem 1, p. 476] that the linear operator B is correct if and only if

$$\det W = \det\Big[I_m - \Psi(\widehat{A}^{-1}g)\Big] \neq 0, \tag{17}$$

where I_m stands for the $m \times m$ identity matrix. Moreover, for any $f \in Y$ the unique solution of Eq. (16) is given by

$$u = B^{-1}f = \widehat{A}^{-1}f + \widehat{A}^{-1}gW^{-1}\Psi(\widehat{A}^{-1}f). \tag{18}$$

The procedure used for deriving Eq. (18) is based on the knowledge of the solution of the simpler linear problem

$$\widehat{A}u = f, \quad f \in Y. \tag{19}$$

Here, we advance the method presented in [10] to handle nonlinear problems.

Let X, Y and Z be complex Banach spaces and $\widehat{A} : X \to Y$ be a linear correct operator with $D(\widehat{A}) \subset Z \subseteq X$. Consider first the nonlinear operator $\mathbf{B} : X \to Y$ defined by

$$\mathbf{B}u = \widehat{A}u - q\Phi(u), \quad D(\mathbf{B}) = D(\widehat{A}), \tag{20}$$

where $\Phi = col(\phi_1, \ldots, \phi_n)$ is a vector of complex-valued continuous nonlinear functionals on Z, $q = (q_1, \ldots, q_n)$ is a vector of elements $q_j \in Y$ and $u \in D(\widehat{A})$. Without any loss of generality, we may assume that each of the sets $\{\phi_j\}$ and $\{q_j\}$ is linearly independent; otherwise, we could reduce the number of their elements.

Theorem 1 *Let* $\mathbf{B} : X \to Y$ *be the nonlinear operator in Eq. (20). Then the exact solution to the problem*

$$\mathbf{B}u = \widehat{A}u - q\Phi(u) = f, \quad f \in Y, \tag{21}$$

is given by

$$u = \widehat{A}^{-1} f + \widehat{A}^{-1} q \mathbf{b}^*, \tag{22}$$

for every vector $\mathbf{b}^* = \Phi(u)$ *that solves the nonlinear algebraic (transcendental) system of n equations*

$$\mathbf{b} = \Phi\left(\widehat{A}^{-1} f + \widehat{A}^{-1} q \mathbf{b}\right). \tag{23}$$

Proof Since the operator $\widehat{A}$ is correct, there exists the inverse operator $\widehat{A}^{-1}$. Applying this inverse operator on both sides of Eq. (21) we get

$$u = \widehat{A}^{-1} f + \widehat{A}^{-1} q \Phi(u). \tag{24}$$

Employing the vector Φ on Eq. (24), we obtain

$$\Phi(u) = \Phi\left(\widehat{A}^{-1} f + \widehat{A}^{-1} q \Phi(u)\right). \tag{25}$$

Setting $\mathbf{b} = \Phi(u)$, we acquire the nonlinear system of n equations in n unknowns in Eq. (23). Let $\mathbf{b}^*$ be a solution of this system satisfying $\mathbf{b}^* = \Phi(u)$. Substitution of $\mathbf{b}^*$ into Eq. (24) produces Eq. (22). □

Consider next the more general nonlinear operator $\mathbf{B} : X \to Y$ defined by

$$\mathbf{B}u = \widehat{A}u - g\Psi(u) - q\Phi(u), \quad D(\mathbf{B}) = D(\widehat{A}), \tag{26}$$

where X, Y and Z are complex Banach spaces and $\widehat{A} : X \to Y$ is a linear correct operator with $D(\widehat{A}) \subset Z \subseteq X$, $\Psi = col(\psi_1, \ldots, \psi_m)$ is a vector of bounded linear functionals $\psi_i : Z \to \mathbf{C}$, $\Phi = col(\phi_1, \ldots, \phi_n)$ is a vector of continuous nonlinear functionals $\phi_j : Z \to \mathbf{C}$, $g = (g_1, \ldots, g_m) \in Y^m$ and $q = (q_1, \ldots, q_n) \in Y^n$. We may assume without any loss of generality that each of the four sets $\{\psi_i\}$, $\{g_i\}$, $\{\phi_j\}$ and $\{q_j\}$ is linearly independent; otherwise, we could diminish the number of their corresponding elements.

Theorem 2 *Let* $\mathbf{B} : X \to Y$ *be the nonlinear operator in Eq. (26). Then the exact solution to the problem*

$$\mathbf{B}u = \widehat{A}u - g\Psi(u) - q\Phi(u) = f, \quad f \in Y, \tag{27}$$

is provided by

$$u = \widehat{A}^{-1} f + \widehat{A}^{-1} g \mathbf{a}^* + \widehat{A}^{-1} q \mathbf{b}^*, \tag{28}$$

where the set of the vectors $\mathbf{a}^* = \Psi(u)$ *and* $\mathbf{b}^* = \Phi(u)$ *is a solution of the* $m + n$ *nonlinear algebraic (transcendental) equations*

$$W\mathbf{a} - Q\mathbf{b} = \Psi\left(\widehat{A}^{-1} f\right), \tag{29}$$

$$\mathbf{b} = \Phi\left(\widehat{A}^{-1} f + \widehat{A}^{-1} g\mathbf{a} + \widehat{A}^{-1} q\mathbf{b}\right), \tag{30}$$

with the $m \times m$ *matrix* $W = I_m - \Psi\left(\widehat{A}^{-1} g\right)$ *and the* $m \times n$ *matrix* $Q = \Psi\left(\widehat{A}^{-1} q\right)$.

Proof Applying the inverse operator $\widehat{A}^{-1}$ on both sides of Eq. (27), we get

$$u - \widehat{A}^{-1} g\Psi(u) - \widehat{A}^{-1} q\Phi(u) = \widehat{A}^{-1} f. \tag{31}$$

Acting by the vector of the linear functionals Ψ on both sides of Eq. (31) and using its linearity properties, we have

$$\begin{aligned}
&\Psi\left(u - \widehat{A}^{-1} g\Psi(u) - \widehat{A}^{-1} q\Phi(u)\right) = \Psi\left(\widehat{A}^{-1} f\right),\\
&\Psi(u) - \Psi\left(\widehat{A}^{-1} g\Psi(u)\right) - \Psi\left(\widehat{A}^{-1} q\Phi(u)\right) = \Psi\left(\widehat{A}^{-1} f\right),\\
&\Psi(u) - \Psi\left(\widehat{A}^{-1} g\right)\Psi(u) - \Psi\left(\widehat{A}^{-1} q\right)\Phi(u) = \Psi\left(\widehat{A}^{-1} f\right),\\
&\left[I_m - \Psi\left(\widehat{A}^{-1} g\right)\right]\Psi(u) - \Psi\left(\widehat{A}^{-1} q\right)\Phi(u) = \Psi\left(\widehat{A}^{-1} f\right),
\end{aligned} \tag{32}$$

or in a compact form

$$W\Psi(u) - Q\Phi(u) = \Psi\left(\widehat{A}^{-1} f\right), \tag{33}$$

where we have set $W = I_m - \Psi\left(\widehat{A}^{-1} g\right)$ and $Q = \Psi\left(\widehat{A}^{-1} q\right)$. Likewise, implementation of the vector of the nonlinear functionals Φ on Eq. (31) yields

$$\Phi(u) = \Phi\left(\widehat{A}^{-1} f + \widehat{A}^{-1} g\Psi(u) + \widehat{A}^{-1} q\Phi(u)\right). \tag{34}$$

Setting $\mathbf{a} = \Psi(u)$ and $\mathbf{b} = \Phi(u)$ into Eqs. (33) and (34), we obtain the nonlinear system of $m + n$ equations in $m + n$ unknowns in Eqs. (29) and (30). Let $\mathbf{a}^* = \Psi(u)$, $\mathbf{b}^* = \Phi(u)$ be a compatible solution of this nonlinear system of equations. Substitution then into Eq. (31) produces Eq. (28). □

Theorem 2 is a general theorem that involves the exact solution of a system of a total of $m + n$ nonlinear algebraic (transcendental) equations. This, in cases of large problems can be a challenge or an impossible task even for the best available computer algebra systems. The following three corollaries deliver the solution of Eq. (27) more efficiently since the system of the nonlinear equations to be solved is reduced to n equations in n unknowns.

Corollary 1 *If the* $m \times m$ *matrix* W *in Eq. (29) is nonsingular, i.e.*

$$det\ W = det\left[I_m - \Psi(\widehat{A}^{-1} g)\right] \neq 0, \tag{35}$$

then the exact solution of Eq. (27) may be obtained conveniently by

$$u = \widehat{A}^{-1}f + \widehat{A}^{-1}gW^{-1}\Psi(\widehat{A}^{-1}f) + \left[\widehat{A}^{-1}q + \widehat{A}^{-1}gW^{-1}Q\right]\mathbf{b}^{*}, \tag{36}$$

where the vector $\mathbf{b}^{*} = \Phi(u)$ *is a solution of the system of the n nonlinear algebraic (transcendental) equations*

$$\mathbf{b} = \Phi\left(\widehat{A}^{-1}f + \widehat{A}^{-1}gW^{-1}\Psi(\widehat{A}^{-1}f) + \left[\widehat{A}^{-1}q + \widehat{A}^{-1}gW^{-1}Q\right]\mathbf{b}\right). \tag{37}$$

Proof Working as for the proof of Theorem 2, we can arrive at Eqs. (31) and (33). Suppose that Eq. (35) holds true. This means that the matrix $W = I_m - \Psi\left(\widehat{A}^{-1}g\right)$ is nonsingular and consequently Eq. (33) can be solved uniquely with respect to the vector $\Psi(u)$ to get

$$\Psi(u) = W^{-1}\left[\Psi\left(\widehat{A}^{-1}f\right) + Q\Phi(u)\right]. \tag{38}$$

Substituting Eq. (38) into Eq. (31), we obtain

$$u = \widehat{A}^{-1}f + \widehat{A}^{-1}gW^{-1}\Psi(\widehat{A}^{-1}f) + \left[\widehat{A}^{-1}q + \widehat{A}^{-1}gW^{-1}Q\right]\Phi(u). \tag{39}$$

Acting by the vector Φ on both sides of Eq. (39), we acquire

$$\Phi(u) = \Phi\left(\widehat{A}^{-1}f + \widehat{A}^{-1}gW^{-1}\Psi(\widehat{A}^{-1}f) + \left[\widehat{A}^{-1}q + \widehat{A}^{-1}gW^{-1}Q\right]\Phi(u)\right). \tag{40}$$

Setting $\mathbf{b} = \Phi(u)$, we get the nonlinear system in Eq. (37). Let $\mathbf{b}^{*} = \Phi(u)$ be an admissible solution vector of this nonlinear system. Substitution of $\mathbf{b}^{*}$ into Eq. (39) provides Eq. (36). □

Remark 1 Observe that Eq. (36), the solution of the nonlinear problem (27), can be written equivalently in the following explicit form

$$\begin{aligned} u = {} & \widehat{A}^{-1}f + \widehat{A}^{-1}gW^{-1}\Psi(\widehat{A}^{-1}f) \\ & + \sum_{j=1}^{n}\left(\widehat{A}^{-1}q_j + \widehat{A}^{-1}gW^{-1}\Psi(\widehat{A}^{-1}q_j)\right)b_j^{*}. \end{aligned} \tag{41}$$

In addition, notice that Eq. (27) by removing the nonlinear terms degenerates to the linear problem (16), namely

$$Bu = \widehat{A}u - g\Psi(u) = f, \quad f \in Y, \tag{42}$$

whose exact solution is given by

$$u_f = \widehat{A}^{-1}f + \widehat{A}^{-1}gW^{-1}\Psi(\widehat{A}^{-1}f). \tag{43}$$

If in Eq. (42), f is replaced by a $q_j \in Y,\ j = 1, \dots, n$ then the associated solution is provided by

$$u_{q_j} = \widehat{A}^{-1} q_j + \widehat{A}^{-1} g W^{-1} \Psi(\widehat{A}^{-1} q_j). \tag{44}$$

Consequently, the solution (36) of the original nonlinear Eq. (27) can alternatively assume the form

$$u = u_f + \sum_{j=1}^{n} u_{q_j} b_j^*. \tag{45}$$

Corollary 2 *Suppose that* $\det W = 0$ *but the rank of the* $m \times (m+n)$ *matrix* $[W \ \ -Q]$ *in Eq. (29) is* $\operatorname{rank}[W \ \ -Q] = m$*. Further, let us assume without any loss of generality that the first* $k < m$ *columns of* W *and the first* $r = m - k$ *columns of* Q *are linearly independent and express Eq. (29) as follows*

$$V \begin{pmatrix} \mathbf{a}_k \\ \mathbf{b}_r \end{pmatrix} + U \begin{pmatrix} \mathbf{a}_r \\ \mathbf{b}_l \end{pmatrix} = \Psi(\widehat{A}^{-1} f), \tag{46}$$

where $l = n - r$*, the* $m \times m$ *matrix* $V = [W_k \ \ -Q_r]$ *is nonsingular, the* $m \times n$ *matrix* $U = [W_r \ \ -Q_l]$*,* $\mathbf{a}_k = col(a_1, \dots, a_k)$*,* $\mathbf{a}_r = col(a_{k+1}, \dots, a_m)$*,* $\mathbf{b}_r = col(b_1, \dots, b_r)$ *and* $\mathbf{b}_l = col(b_{r+1}, \dots, b_n)$*. Then, the solution of Eq. (27) is provided aptly in closed form by*

$$u = \widehat{A}^{-1} f + \widehat{A}^{-1} g \begin{pmatrix} \mathbf{a}_k \\ \mathbf{a}_r^* \end{pmatrix} + \widehat{A}^{-1} q \begin{pmatrix} \mathbf{b}_r \\ \mathbf{b}_l^* \end{pmatrix}, \tag{47}$$

with

$$\begin{pmatrix} \mathbf{a}_k \\ \mathbf{b}_r \end{pmatrix} = V^{-1} \left(\Psi(\widehat{A}^{-1} f) - U \begin{pmatrix} \mathbf{a}_r^* \\ \mathbf{b}_l^* \end{pmatrix} \right), \tag{48}$$

for every solution $\mathbf{a}_r^* = \Psi_r(u)$ *and* $\mathbf{b}_l^* = \Phi_l(u)$ *of the nonlinear system of* n *equations*

$$\begin{pmatrix} \mathbf{b}_r \\ \mathbf{b}_l \end{pmatrix} = \Phi \left(\widehat{A}^{-1} f + \widehat{A}^{-1} g \begin{pmatrix} \mathbf{a}_k \\ \mathbf{a}_r \end{pmatrix} + \widehat{A}^{-1} q \begin{pmatrix} \mathbf{b}_r \\ \mathbf{b}_l \end{pmatrix} \right), \tag{49}$$

where

$$\begin{pmatrix} \mathbf{a}_k \\ \mathbf{b}_r \end{pmatrix} = V^{-1} \left(\Psi(\widehat{A}^{-1} f) - U \begin{pmatrix} \mathbf{a}_r \\ \mathbf{b}_l \end{pmatrix} \right). \tag{50}$$

Proof Repeating the same steps as in the proof of Theorem 2 we obtain Eqs. (31) and (33). Since $rank\,[W \;\; -Q] = m$ there are exist m columns of the $m \times (m+n)$ matrix $[W \;\; -Q]$ which are linearly independent. We assume without any loss of generality that the first $k < m$ columns of W and the first $r = m-k < n$ columns of Q are linearly independent; otherwise, we could change the order of the $m = k+r$ and $n = r+l$ columns, respectively. We partition Eq. (33) as follows

$$\begin{gathered}
\begin{bmatrix} W_k & W_r \end{bmatrix} \Psi(u) - \begin{bmatrix} Q_r & Q_l \end{bmatrix} \Phi(u) = \Psi\left(\widehat{A}^{-1} f\right), \\
\begin{bmatrix} W_k & -Q_r \end{bmatrix} \begin{pmatrix} \mathbf{a}_k \\ \mathbf{b}_r \end{pmatrix} + \begin{bmatrix} W_r & -Q_l \end{bmatrix} \begin{pmatrix} \mathbf{a}_r \\ \mathbf{b}_l \end{pmatrix} = \Psi\left(\widehat{A}^{-1} f\right), \\
V \begin{pmatrix} \mathbf{a}_k \\ \mathbf{b}_r \end{pmatrix} + U \begin{pmatrix} \mathbf{a}_r \\ \mathbf{b}_l \end{pmatrix} = \Psi(\widehat{A}^{-1} f), \qquad (51)
\end{gathered}$$

where we have set

$$\Psi(u) = \mathbf{a} = \begin{pmatrix} \mathbf{a}_k \\ \mathbf{a}_r \end{pmatrix}, \quad \Phi(u) = \mathbf{b} = \begin{pmatrix} \mathbf{b}_r \\ \mathbf{b}_l \end{pmatrix}, \qquad (52)$$

and $V = [W_k \;\; -Q_r]$ is the $m \times m$ matrix containing the m linearly independent columns, while $U = [W_r \;\; -Q_l]$ is the $m \times n$ matrix that consists of the remaining n columns of $[W \;\; -Q]$. Accordingly, Eq. (31) is written

$$u = \widehat{A}^{-1} f + \widehat{A}^{-1} g \begin{pmatrix} \mathbf{a}_k \\ \mathbf{a}_r \end{pmatrix} + \widehat{A}^{-1} q \begin{pmatrix} \mathbf{b}_r \\ \mathbf{b}_l \end{pmatrix}, \qquad (53)$$

where by means of Eq. (51)

$$\begin{aligned}
\begin{pmatrix} \mathbf{a}_k \\ \mathbf{b}_r \end{pmatrix} &= V^{-1} \left(\Psi(\widehat{A}^{-1} f) - U \begin{pmatrix} \mathbf{a}_r \\ \mathbf{b}_l \end{pmatrix} \right) \\
&= \begin{bmatrix} \left(V^{-1}\right)_k \\ \left(V^{-1}\right)_r \end{bmatrix} \left(\Psi(\widehat{A}^{-1} f) - U \begin{pmatrix} \mathbf{a}_r \\ \mathbf{b}_l \end{pmatrix} \right). \qquad (54)
\end{aligned}$$

Applying the vector Φ on both sides of (53), we get

$$\begin{pmatrix} \mathbf{b}_r \\ \mathbf{b}_l \end{pmatrix} = \Phi \left(\widehat{A}^{-1} f + \widehat{A}^{-1} g \begin{pmatrix} \mathbf{a}_k \\ \mathbf{a}_r \end{pmatrix} + \widehat{A}^{-1} q \begin{pmatrix} \mathbf{b}_r \\ \mathbf{b}_l \end{pmatrix} \right), \qquad (55)$$

which is the nonlinear system in Eq. (49) with $\mathbf{a}_k$, $\mathbf{b}_r$ given by (50) and $\mathbf{a}_r$, $\mathbf{b}_l$ being the n unknowns. Let $\mathbf{a}_r^*$, $\mathbf{b}_l^*$ be a solution of this nonlinear system compatible with (52). Substituting into Eqs. (54) and (53) we obtain Eq. (47). □

Corollary 3 *In Theorem 2, if $m = n$ then the matrix Q in Eq. (29) becomes a square matrix. If this matrix is nonsingular, i.e.*

$$det\ Q = det\left[\Psi(\widehat{A}^{-1}q)\right] \neq 0, \tag{56}$$

then the exact solution to the problem (27) is given readily by

$$u = \widehat{A}^{-1}\left[g + qQ^{-1}W\right]\mathbf{a}^* + \widehat{A}^{-1}\left[f - qQ^{-1}\Psi(\widehat{A}^{-1}f)\right], \tag{57}$$

where the vector $\mathbf{a}^* = \Psi(u)$ is a solution to the system of m nonlinear algebraic (transcendental) equations

$$\begin{aligned} &Q^{-1}[W\mathbf{a} - \Psi(\widehat{A}^{-1}f)] = \\ &\qquad \Phi\left(\widehat{A}^{-1}\left[g + qQ^{-1}W\right]\mathbf{a} + \widehat{A}^{-1}\left[f - qQ^{-1}\Psi(\widehat{A}^{-1}f)\right]\right). \end{aligned} \tag{58}$$

Proof By employing the inverse operator $\widehat{A}^{-1}$ on both sides of Eq. (27) we get

$$u - \widehat{A}^{-1}g\Psi(u) - \widehat{A}^{-1}q\Phi(u) = \widehat{A}^{-1}f. \tag{59}$$

Acting by the vector Ψ on both sides of Eq. (59), we have

$$\begin{aligned} \Psi\left(\widehat{A}^{-1}q\right)\Phi(u) &= \left[I_m - \Psi\left(\widehat{A}^{-1}g\right)\right]\Psi(u) - \Psi\left(\widehat{A}^{-1}f\right), \\ Q\Phi(u) &= W\Psi(u) - \Psi\left(\widehat{A}^{-1}f\right), \end{aligned} \tag{60}$$

where we have put $Q = \Psi(\widehat{A}^{-1}q)$ and $W = I_m - \Psi\left(\widehat{A}^{-1}g\right)$. By hypothesis $det\ Q \neq 0$ and hence

$$\Phi(u) = Q^{-1}\left[W\Psi(u) - \Psi\left(\widehat{A}^{-1}f\right)\right]. \tag{61}$$

Application of the vector Φ on Eq. (59) yields

$$\Phi(u) = \Phi\left[\widehat{A}^{-1}g\Psi(u) + \widehat{A}^{-1}q\Phi(u) + \widehat{A}^{-1}f\right]. \tag{62}$$

Substituting Eq. (61) into Eq. (62) we obtain

$$\begin{aligned} &Q^{-1}\left[W\Psi(u) - \Psi\left(\widehat{A}^{-1}f\right)\right] = \\ &\qquad \Phi\left(\widehat{A}^{-1}\left[g + qQ^{-1}W\right]\Psi(u) + \widehat{A}^{-1}\left[f - qQ^{-1}\Psi(\widehat{A}^{-1}f)\right]\right). \end{aligned} \tag{63}$$

Setting $\mathbf{a} = \Psi(u)$, we obtain the system of m nonlinear equations (58). Let $\mathbf{a}^* = \Psi(u)$ be a vector that solves this nonlinear system. Substitution of $\mathbf{a}^*$ along with Eq. (61) into Eq. (59) yields Eq. (57). □

Finally, we quote the next corollary that provides a sufficient condition for the solvability of the nonlinear problem (27).

Corollary 4 *If the rank* $\left[W \;\; -Q \;\; \Psi(\widehat{A}^{-1}f)\right] > rank\,[W \;\; -Q]$ *then Eq. (27) has no solutions.*

Proof If the rank of the augmented matrix $\left[W \;\; -Q \;\; \Psi(\widehat{A}^{-1}f)\right]$ is greater than that of $[W \;\; -Q]$, it is known that no solution exists to the linear system in Eq. (29). This is to say that there are not any $\mathbf{a}$ and $\mathbf{b}$ which satisfy simultaneously both Eq. (29) and Eq. (30). Therefore, it is concluded that the nonlinear equation (27) does not possess any solution. □

4 Example Problems

In this section we solve some selected initial value problems and boundary value problems with nonlinear integro-differential equations of Fredholm type of the second kind to reveal the capabilities of the method proposed and to highlight all concepts postulated in the theory in Sect. 3.

Problem 1 Consider the boundary value problem with nonlocal boundary conditions

$$\begin{aligned} &u''(x) - 70x^2 \int_0^1 t^2u^3(t)dt = 89x^2 + 24x, \\ &u(0) = -u(1), \quad u'(0) = -u'(1), \quad x \in [0, 1]. \end{aligned} \tag{64}$$

Imitating the procedure outlined in Sect. 2, the problem can be put in the operator form

$$\begin{aligned} &\mathbf{B}u(x) = \widehat{A}u(x) - q\Phi(u) = f(x), \\ &D(\mathbf{B}) = \{u \in \mathbf{C}^2[0, 1] : u(0) = -u(1),\; u'(0) = -u'(1)\}. \end{aligned} \tag{65}$$

where the operators $\mathbf{B}$, $\widehat{A} : \mathbf{C}[0, 1] \to \mathbf{C}[0, 1]$ with

$$\begin{aligned} &\widehat{A}u(x) = u''(x), \quad D(\widehat{A}) = D(\mathbf{B}), \\ &\Phi(u) = \phi(u) = \int_0^1 t^2u^3(t)dt, \\ &q = q(x) = 70x^2, \\ &f(x) = 89x^2 + 24x. \end{aligned} \tag{66}$$

The linear operator $\widehat{A}$ is correct and its inverse, see e.g. [11], is given by

$$\widehat{A}^{-1}f(x) = \int_0^x (x-t)f(t)dt - \frac{1}{2}\int_0^1 \left(x - t + \frac{1}{2}\right) f(t)dt, \text{ for all } f(x) \in \mathbf{C}[0, 1]. \tag{67}$$

The nonlinear functional $\phi : \mathbf{C}[0, 1] \to \mathbf{C}$ is continuous and hence Theorem 1 can be applied. By using the inverse operator $\widehat{A}^{-1}$, we get

$$\begin{aligned}\widehat{A}^{-1}q(x) &= \frac{70}{24}(2x^4 - 4x + 1),\\ \widehat{A}^{-1}f(x) &= \frac{1}{24}(178x^4 + 96x^3 - 500x + 113),\end{aligned} \tag{68}$$

and after substituting into Eq. (23), we obtain a third-degree polynomial equation that possesses one real root, namely $b^* = -\frac{89}{70}$, which when is put in Eq. (22) yields the exact real solution to the problem (64), viz.

$$u(x) = 4x^3 - 6x + 1. \tag{69}$$

Problem 2 From [15], solve the second order integro-differential equation of Fredholm type with initial conditions,

$$\begin{aligned}&u''(x) - \frac{1}{2}\int_{-1}^1 (xt + x^2t^2)(u(t) - u^2(t))dt = \frac{19}{35}x^2 + \frac{11}{15}x + 2,\\ &u(0) = u'(0) = 1, \quad x \in [-1, 1].\end{aligned} \tag{70}$$

Setting

$$v(x) = u(x) - x - 1, \tag{71}$$

carries Eq. (70) into

$$\begin{aligned}&v''(x) + \frac{x}{2}\int_{-1}^1 t(2t+1)v(t)dt + \frac{x^2}{2}\int_{-1}^1 t^2(2t+1)v(t)dt\\ &\qquad + \frac{x}{2}\int_{-1}^1 tv^2(t)dt + \frac{x^2}{2}\int_{-1}^1 t^2v^2(t)dt = \frac{12}{35}x^2 + \frac{2}{5}x + 2,\\ &v(0) = v'(0) = 0, \quad x \in [-1, 1],\end{aligned} \tag{72}$$

with homogeneous initial conditions. Let the operator $\mathbf{B} : \mathbf{C}[-1, 1] \to \mathbf{C}[-1, 1]$ be defined by

$$\begin{aligned}\mathbf{B}v(x) &= \widehat{A}v(x) - g\Psi(v) - q\Phi(v) = f(x),\\ D(\mathbf{B}) &= \{v \in \mathbf{C}^2[-1, 1] : v(0) = v'(0) = 0\},\end{aligned} \tag{73}$$

where

$$
\begin{aligned}
&\widehat{A}v(x)=v''(x), \quad D(\widehat{A})=D(\mathbf{B}),\\
&\Psi(v)=\begin{pmatrix}\psi_1(v)\\ \psi_2(v)\end{pmatrix}=\begin{pmatrix}\int_{-1}^{1} t(2t+1)v(t)dt\\ \int_{-1}^{1} t^2(2t+1)v(t)dt\end{pmatrix},\\
&\Phi(v)=\begin{pmatrix}\phi_1(v)\\ \phi_2(v)\end{pmatrix}=\begin{pmatrix}\int_{-1}^{1} tv^2(t)dt\\ \int_{-1}^{1} t^2v^2(t)dt\end{pmatrix},\\
&g=\begin{pmatrix}g_1(x) & g_2(x)\end{pmatrix}=\begin{pmatrix}-\frac{x}{2} & -\frac{x^2}{2}\end{pmatrix},\\
&q=\begin{pmatrix}q_1(x) & q_2(x)\end{pmatrix}=\begin{pmatrix}-\frac{x}{2} & -\frac{x^2}{2}\end{pmatrix},\\
&f(x)=\frac{12}{35}x^2+\frac{2}{5}x+2.
\end{aligned}
\tag{74}
$$

It is known that the operator $\widehat{A}$ is correct with its inverse being

$$
\widehat{A}^{-1}f(x)=\int_0^x (x-t)f(t)dt, \quad \text{for all } f\in\mathbf{C}[-1,1],
\tag{75}
$$

while the linear functionals $\psi_i,\ i=1,2$ are bounded and the nonlinear functionals $\phi_j,\ j=1,2$ are continuous on $\mathbf{C}[-1,1]$ and thus we can compute the matrices W and Q as follows

$$
\begin{aligned}
&\widehat{A}^{-1}g=\begin{pmatrix}\widehat{A}^{-1}g_1 & \widehat{A}^{-1}g_2\end{pmatrix}=\begin{pmatrix}-\frac{x^3}{12} & -\frac{x^4}{24}\end{pmatrix},\\
&\widehat{A}^{-1}q=\begin{pmatrix}\widehat{A}^{-1}q_1 & \widehat{A}^{-1}q_2\end{pmatrix}=\begin{pmatrix}-\frac{x^3}{12} & -\frac{x^4}{24}\end{pmatrix},\\
&\widehat{A}^{-1}f=\frac{x^4}{35}+\frac{x^3}{15}+x^2,\\
&W=\begin{pmatrix}1-\psi_1(\widehat{A}^{-1}g_1) & -\psi_1(\widehat{A}^{-1}g_2)\\ -\psi_2(\widehat{A}^{-1}g_1) & 1-\psi_2(\widehat{A}^{-1}g_2)\end{pmatrix}=\begin{pmatrix}\frac{31}{30} & \frac{1}{42}\\ \frac{1}{21} & \frac{85}{84}\end{pmatrix},\\
&Q=\begin{pmatrix}\psi_1(\widehat{A}^{-1}q_1) & \psi_1(\widehat{A}^{-1}q_2)\\ \psi_2(\widehat{A}^{-1}q_1) & \psi_2(\widehat{A}^{-1}q_2)\end{pmatrix}=\begin{pmatrix}-\frac{1}{30} & -\frac{1}{42}\\ -\frac{1}{21} & -\frac{1}{84}\end{pmatrix},\\
&\Psi(\widehat{A}^{-1}f)=\begin{pmatrix}\psi_1(\widehat{A}^{-1}f)\\ \psi_2(\widehat{A}^{-1}f)\end{pmatrix}=\begin{pmatrix}\frac{3098}{3675}\\ \frac{328}{735}\end{pmatrix}.
\end{aligned}
\tag{76}
$$

Since $det\ W=\frac{3098}{3528}\neq 0$, we employ Corollary 1. By means of (76), we construct the nonlinear system of two equations in two unknowns in Eq. (37). This system has the exact solution $\mathbf{b}^*=col(0,\ \frac{2}{7})$. From Eq. (36), we get the solution $v(x)=x^2$ for the transformed problem (72), and subsequently from (71) the exact solution to the original problem (70), specifically

$$u(x) = x^2 + x + 1. \tag{77}$$

Problem 3 Consider the following initial value problem implicating a second order integro-differential equation of Fredholm type

$$\begin{aligned}&u'' - \frac{96}{7}x\int_0^1 u(t)dt - \int_0^1 tu'(t)dt - 35x^2\int_0^1 u^2(t)dt = -\frac{71}{3}x^2 - 2x + \frac{7}{12},\\ &u(0) = u'(0) = 0, \quad x \in [0, 1].\end{aligned} \tag{78}$$

Working as in Sect. 2, the operator $\mathbf{B} : \mathbf{C}[0, 1] \to \mathbf{C}[0, 1]$ is defined by

$$\begin{aligned}&\mathbf{B}u(x) = \widehat{A}u(x) - g\Psi(u) - q\Phi(u) = f(x),\\ &D(\mathbf{B}) = \{u \in \mathbf{C}^2[0, 1] : u(0) = u'(0) = 0\},\end{aligned} \tag{79}$$

where

$$\begin{aligned}&\widehat{A}u(x) = u''(x), \quad D(\widehat{A}) = D(\mathbf{B}),\\ &\Psi(u) = \begin{pmatrix}\psi_1(u)\\ \psi_2(u)\end{pmatrix} = \begin{pmatrix}\int_0^1 u(t)dt\\ \int_0^1 tu'(t)dt\end{pmatrix},\\ &\Phi(u) = \phi(u) = \int_0^1 u^2(t)dt,\\ &g = \begin{pmatrix}g_1(x) & g_2(x)\end{pmatrix} = \begin{pmatrix}\frac{96x}{7} & 1\end{pmatrix},\\ &q = q(x) = 35x^2,\\ &f(x) = -\frac{1}{12}(284x^2 + 24x - 7).\end{aligned} \tag{80}$$

By employing the inverse operator $\widehat{A}^{-1}f(x) = \int_0^x (x-t)f(t)dt$ and by observing that $\psi_i,\ i = 1, 2$ are bounded and ϕ is continuous on $\mathbf{C}^1[0, 1]$, we construct the matrices W and Q, viz.

$$\begin{aligned}&\widehat{A}^{-1}g = \begin{pmatrix}\widehat{A}^{-1}g_1 & \widehat{A}^{-1}g_2\end{pmatrix} = \begin{pmatrix}\frac{16x^3}{7} & \frac{x^2}{2}\end{pmatrix},\\ &\widehat{A}^{-1}q = \frac{35}{12}x^4,\\ &\widehat{A}^{-1}f = -\frac{1}{12}\left(\frac{71}{3}x^4 + 4x^3 - \frac{7}{2}x^2\right),\end{aligned}$$

$$\begin{aligned}
W &= \begin{bmatrix} 1-\psi_1(\widehat{A}^{-1}g_1) & -\psi_1(\widehat{A}^{-1}g_2) \\ -\psi_2(\widehat{A}^{-1}g_1) & 1-\psi_2(\widehat{A}^{-1}g_2) \end{bmatrix} = \begin{bmatrix} \frac{3}{7} & -\frac{1}{6} \\ -\frac{12}{7} & \frac{2}{3} \end{bmatrix}, \\
Q &= \begin{bmatrix} \psi_1(\widehat{A}^{-1}q) \\ \psi_2(\widehat{A}^{-1}q) \end{bmatrix} = \begin{bmatrix} \frac{7}{12} \\ \frac{7}{3} \end{bmatrix}, \\
\Psi(\widehat{A}^{-1}f) &= \begin{pmatrix} \psi_1(\widehat{A}^{-1}f) \\ \psi_2(\widehat{A}^{-1}f) \end{pmatrix} = \begin{pmatrix} -\frac{137}{360} \\ -\frac{49}{30} \end{pmatrix}.
\end{aligned} \tag{81}$$

Since $det\, W = 0$ but $rank\,[W - Q] = 2$ we implement the Corollary 2. Accordingly, we assemble the matrices V and U in Eq. (46), namely

$$V = \begin{bmatrix} \frac{3}{7} & -\frac{7}{12} \\ -\frac{12}{7} & -\frac{7}{3} \end{bmatrix}, \quad U = \begin{bmatrix} -\frac{1}{6} \\ \frac{2}{3} \end{bmatrix}, \tag{82}$$

from Eq. (50) we find

$$\begin{pmatrix} a_1 \\ b \end{pmatrix} = \begin{pmatrix} \frac{84a_2+7}{216} \\ \frac{71}{105} \end{pmatrix}, \tag{83}$$

and by substituting in Eq. (49), we get the nonlinear equation

$$507888a_2^2 + 58296a_2 - 1101889 = 0. \tag{84}$$

This equation has two real roots $a_2^* = \frac{17}{12}$ and $a_2^* = -\frac{64817}{42324}$. Putting each of them into Eqs. (48) and (47), we get the two solutions of the nonlinear problem (78) in the explicit form

$$u_1(x) = x^3 + x^2, \quad u_2(x) = -\frac{17147}{10581}x^3 - \frac{5016}{10581}x^2. \tag{85}$$

Problem 4 Find the exact solution of the following first order integro-differential equation of Fredholm type

$$\begin{aligned}
&u' - 12x^2 \int_0^1 u(t)dt - x \int_0^1 u^2(t)dt = -10x^2 + \frac{29}{30}x + 1, \\
&u(0) = 0, \quad x \in [0, 1].
\end{aligned} \tag{86}$$

We begin by formulating (86) in the operator form

$$\begin{aligned}
&\mathbf{B}u(x) = \widehat{A}u(x) - g\Psi(u) - q\Phi(x) = f(x), \\
&D(\mathbf{B}) = \{u(x) \in \mathbf{C}^1[0,1] : u(0) = 0\},
\end{aligned} \tag{87}$$

where the operator $\mathbf{B} : \mathbf{C}[0, 1] \to \mathbf{C}[0, 1]$ and

$$\begin{aligned}
&\widehat{A}u(x) = u'(x), \quad D(\widehat{A}) = D(\mathbf{B}),\\
&\Psi(u) = \psi(u) = \int_0^1 u(t)dt, \quad \Phi(u) = \phi(u) = \int_0^1 u^2(t)dt,\\
&g = g(x) = 12x^2, \quad q = q(x) = x,\\
&f(x) = -10x^2 + \frac{29}{30}x + 1. \qquad (88)
\end{aligned}$$

It is known that the inverse operator $\widehat{A}^{-1} f(x) = \int_0^x f(t)dt$ for all $f \in \mathbf{C}[0, 1]$ while ψ is bounded and ϕ is continuous on $\mathbf{C}[0, 1]$, and therefore we can easily compute

$$W = 1 - \Psi(\widehat{A}^{-1}g) = 0, \quad Q = \Psi(\widehat{A}^{-1}q) = \frac{1}{6}, \quad \Psi(\widehat{A}^{-1}f) = -\frac{31}{180}. \qquad (89)$$

Because $m = n = 1$, $det\, W = 0$ and $det\, Q = \frac{1}{6} \neq 0$, Corollary 3 applies. Substituting into Eq. (58), we get the nonlinear equation

$$120a^2 - 46a - 45 = 0, \qquad (90)$$

which possesses two real roots, $a^* = \frac{5}{6}$ and $a^* = \frac{-9}{20}$. Putting these roots into Eq. (57), we obtain in closed form the two solutions to the nonlinear problem (86), viz.

$$u_1(x) = x^2 + x, \quad u_2(x) = -\frac{77}{15}x^3 + x^2 + x. \qquad (91)$$

Problem 5 As a last example, consider the first order integro-differential initial value problem of Fredholm type

$$\begin{aligned}
&u' - 12x^2 \int_0^1 u(t)dt - (6x - 2)\int_0^1 u^2(t)dt = 3x^2 - 2x + 1,\\
&u(0) = 0, \quad x \in [0, 1]. \qquad (92)
\end{aligned}$$

Let the nonlinear operator $\mathbf{B} : \mathbf{C}[0, 1] \to \mathbf{C}[0, 1]$ and formulate the problem (92) as follows

$$\begin{aligned}
&\mathbf{B}u(x) = \widehat{A}u(x) - g\Psi(u) - q\Phi(x) = f(x),\\
&D(\mathbf{B}) = \{u \in \mathbf{C}^1[0, 1] : u(0) = 0\}, \qquad (93)
\end{aligned}$$

where

$$
\begin{aligned}
&\widehat{A}u(x) = u'(x), \quad D(\widehat{A}) = D(\mathbf{B}),\\
&\Psi(u) = \psi(u) = \int_0^1 u(t)dt, \quad \Phi(u) = \phi(u) = \int_0^1 u^2(t)dt,\\
&g = g(x) = 12x^2, \quad q = q(x) = 6x - 2,\\
&f(x) = 3x^2 - 2x + 1.
\end{aligned} \tag{94}
$$

By making use of the inverse operator $\widehat{A}^{-1} f(x) = \int_0^x f(t)dt$, for all $f \in \mathbf{C}[0, 1]$, and the fact that ψ is bounded and ϕ is continuous on $\mathbf{C}[0, 1]$, we have

$$
W = 1 - \Psi(\widehat{A}^{-1}g) = 0, \quad Q = \Psi(\widehat{A}^{-1}q) = 0, \quad \Psi(\widehat{A}^{-1}f) = \frac{5}{12}. \tag{95}
$$

Observe that $rank\left[W \; -Q \;\; \Psi(\widehat{A}^{-1}f)\right] > rank\,[W \; -Q]$ and hence Eq. (92) does not possess any solutions by Corollary 4.

5 Conclusions

The extension operator method has been presented for constructing exact solutions to a class of initial and boundary value problems involving a nonlinear operator **B** defined as a perturbation of a linear correct operator $\widehat{A}$ with linear bounded functionals and nonlinear continuous functionals.

The method has been applied for solving exactly nonlinear integro-differential equations of Fredholm type with separable kernels and has been proved to be very efficient.

It can also be employed equally well for the exact solution of linear and nonlinear integral equations, differential equations with loads and difference equations.

References

1. K.E. Atkinson, A survey of numerical methods for solving nonlinear integral equations. J. Integral Methods Appl. **4**(1), 15–46 (1992)
2. A.H. Bhrawy, E. Tohidi, F. Soleymani, A new Bernoulli matrix method for solving high-order linear and nonlinear Fredholm integro-differential equations with piecewise intervals. Appl. Math. Comput. **219**, 482–497 (2012)
3. F. Bloom, *Ill-posed Problems for Integrodifferential Equations in Mechanics and Electromagnetic Theory* (SIAM, Philadelphia, 1981)
4. J.M. Cushing, *Integrodifferential Equations and Delay Models in Population Dynamics* (Springer, Berlin, 1977)

5. A.A. Dezin, Nonstandard problems. Math. Notes **41**(3), 205–210 (1987). https://doi.org/10.1007/BF01158250
6. M.A. Kraemer, L.V. Kalachev, Analysis of a class of nonlinear integro-differential equations arising in a forestry application. Q. Appl. Math. **61**(3), 513–535 (2003)
7. L.P. Lebedev, I.I. Vorovich, *Functional Analysis in Mechanics* (Springer, Berlin, 2003)
8. A. Molabahrami, Direct computation method for solving a general nonlinear Fredholm integro-differential equation under the mixed conditions: degenerate and non-degenerate kernels. J. Comput. Anal. Appl. **282**, 34–43 (2015)
9. R.O. Oinarov, I.N. Parasidis, Correct extensions of operators with finite defect in Banach spaces. Izv. Akad. Kaz. SSR. **5**, 42–46 (1988)
10. I.N. Parasidis, E. Providas, Extension operator method for the exact solution of integro-differential equations, in *Contributions in Mathematics and Engineering*, ed. by P.M. Pardalos, T.M. Rassias (Springer, Berlin, 2016), pp. 473–496. https://doi.org/10.1007/978-3-319-31317-7_23
11. I.N. Parasidis, P. Tsekrekos, Some quadratic correct extensions of minimal operators in Banach spaces. Oper. Matrices **4**, 225–243 (2010)
12. A.D. Polyanin, A.I. Zhurov, Exact solutions to some classes of nonlinear integral, integro-functional, and integro-differential equations. Dokl. Math. **77**(2), 315–319 (2008)
13. P. Veng-Pedersen, J.A. Widness, L.M. Pereira, C. Peters, R.L. Schmidt, L.S. Lowe, Kinetic evaluation of nonlinear drug elimination by a disposition decomposition analysis. Application to the analysis of the nonlinear elimination kinetics of erythropoietin in adult humans. J. Pharm. Sci. **84**, 760–767 (1995). https://doi.org/10.1002/jps.2600840619
14. A.M. Wazwaz, A comparison study between the modified decomposition method and the traditional methods for solving nonlinear integral equations. Appl. Math. Comput. **181**, 1703–1712 (2006)
15. A.M. Wazwaz, *Linear and Nonlinear Integral Equations* (Springer, Berlin, 2011)

Qualitative, Approximate and Numerical Approaches for the Solution of Nonlinear Differential Equations

Eugenia N. Petropoulou and Michail A. Xenos

1 Introduction

During a standard undergraduate course in differential equations (DEs), ordinary (ODEs) or partial (PDEs), we are taught several methods in order to find their exact solutions, i.e. their solutions in terms of elementary or special functions, or even in the form of power series in their independent variable(s), provided that the corresponding coefficients can be uniquely determined. All these methods refer mostly to linear equations and a "small" amount of specific classes of nonlinear equations. Unfortunately, most differential equations describing real life problems are nonlinear and the vast majority of these cannot be solved explicitly. In these cases, we are obliged to use other, more advanced techniques widely used in research, some of which are also taught in some undergraduate and many postgraduate courses. In several cases not even these methods can give satisfactory results and the need for new methods is required.

The aim of this chapter is to describe some of these more advanced techniques which can be characterized as (a) qualitative, (b) approximate or (c) numerical. Qualitative methods are employed in order to obtain information about the qualitative characteristics of the solution of a DE. As a consequence, they provide answers to questions such as:

- Do they exist solutions in specific spaces of interest? If yes, are they unique?
- Is the solution bounded? Can we find explicit bounds for it?

E. N. Petropoulou
Department of Civil Engineering, University of Patras, Patras, Greece
e-mail: jenpetr@upatras.gr

M. A. Xenos (✉)
Department of Mathematics, University of Ioannina, Ioannina, Greece
e-mail: mxenos@cc.uoi.gr

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_22

- Do they exist positive/negative solutions?
- Do they exist periodic solutions?
- Can we describe the limiting behavior of the solution in the neighborhood of infinity, or in the neighborhood of specific points of interest?
- In the case of ODEs, can we describe their trajectories?

Such kind of topics as the ones raised by the previous questions are covered in various books. Indicatively we mention [6, 40, 43] and [57] for ODEs and [71–73] for PDEs. Especially the last three questions, are connected with what is called the analysis of dynamics of an ODE, which includes bifurcation analysis and the theory of chaos.

Approximate methods may be considered as the next best thing after the methods which calculate explicitly the solution of a DE. Their aim is to analytically calculate several approximations of the true solution of the DE under consideration, each of which is better than the previous one. One typical example of such a method is Picard's iteration technique, which is included in almost all standard textbooks regarding ODEs. Another class of approximate methods are the perturbation methods, the differential transform method, the Adomian decomposition method, as well as the homotopy analysis method (HAM). We shall confine ourselves to perturbation methods, which are covered in various books such as [3, 8, Lectures 17 and 18], [50, Chapter 2] and [56], as well as the HAM, for which a thorough presentation can be found in [46]. For the Adomian decomposition method we refer to [1, 2], whereas for the differential transform method, we refer to the recent book [29].

Finally, numerical methods are our last hope when none of the aforementioned two categories of methods can provide us with satisfactory results. Numerical methods calculate numerically the approximate solution of a DE and they are based on discretizing the physical domain, i.e. the domain where the dependent and independent variables are defined. There is a large number of numerical methods. One class of these are the Runge-Kutta (RK) methods, which are used as a reference methodology for comparison with other analytical or numerical approaches. RK methods were initially introduced a century ago but still are a golden-standard approach for the numerical solution of ODEs. Several scientists used the RK method to numerically solve the Duffing oscillator and the van der Pol equation [14, 55, 68]. Another class of these methods are the finite differences methods (FDM), which are based on replacing the derivatives appearing in a DE with finite difference approximations. In this way, instead of the DE under consideration, it is necessary to solve a finite system of algebraic equations, which is called the corresponding numerical scheme. See for example [45]. Dal used the FDM for the numerical solution of the van der Pol oscillator with small fractional damping [19]. Two similar in philosophy, but different in the implementation techniques, are the finite volumes (FVM) and the finite elements (FEM) methods. With the help of FEM we are able to construct an approximate solution of the initial value problem under consideration. The Duffing equation was studied with the use of the FEM in [18, 63]. Finally, another class of numerical methods are the spectral methods. See for example [11].

Apart from these numerical methods, there also exist numerical techniques based on nonstandard finite differences schemes, which tend to minimize or even vanish problems associated with the numerical instabilities of a numerical scheme. We shall not deal with such kind of schemes here and instead we refer to [53] and [54].

In many cases a problem must be "attacked" with a combination of methods. For example, we may use specific theorems describing the qualitative properties of the trajectories of an ODE, but then we'll probably apply a FDM in order to actually visualize some of these trajectories. Also, in many cases approximate techniques are combined with numerical ones. For example, the differential transform method was recently combined with the method of steps in [65, 66], in order to solve differential and functional differential equations with delay. Many methods may also fall in more than one of the aforementioned categories. For example, the method that we'll present in Sect. 5, is a functional-analytic technique (FAT) developed in [59, 60] and [58], which gives information regarding the existence and uniqueness of bounded solutions of an ODE in a specified Banach space of analytic functions defined in the open unit disc. In this sense, it can be regarded as a qualitative method. However, it can be implemented with the aid of a computer in order to calculate the solution of the ODE in a far greater domain. In this sense, it can also be regarded as an approximate technique.

In this chapter, some of the aforementioned methods will be presented by applying them to the nonlinear ODE

$$x'' + a\left(x^2 - 1\right)x' + \beta x + \gamma x^3 = f(t), \quad \text{where} \quad x = x(t), \tag{1}$$

subject to the initial conditions

$$x(0) = X_0, \quad x'(0) = X_1. \tag{2}$$

Equation (1) is referred to as a Duffing-van der Pol equation, since for $\beta = 1$, $\gamma = 0$ and $f(t) \equiv 0$ reduces to the well-known van der Pol equation, whereas for $a = 0$ and $f(t) \equiv 0$, it reduces to a special form of the famous Duffing equation. Equation (1) is a basic model for self-excited oscillations arising in various problems, including nonlinear vibrations [64] and nonlinear analysis of structures [67]. As mentioned in [33], it is also "important in problems of "stall" instability in the operation of air compressors and industrial fans or centrifugal pumps". With respect to the oscillatory problems that (1) describes, the term $a\left(x^2 - 1\right)x'$ represents the nonlinear damping effect, the term $f(t)$ stands for the applied external force, whereas the term γx^3 represents the nonlinearity of the oscillation. Due to its practical applications, we'll mostly confine ourselves, with respect to the included graphs, to values of the parameters a, β and γ which give rise to periodic solutions of (1).

Qualitative methods for the study of (1) are employed in Sect. 2, whereas (1)–(2) is solved using approximate techniques in Sect. 3, numerical techniques in Sect. 4 and a functional-analytic technique in Sect. 5. More precisely, in Sect. 2.1 some elementary special cases are treated and various results are given regarding the integrability, the periodic character and the chaotic behavior of (1). In Sect. 2.2, (1)–(2) is connected with a Green function. In Sect. 3.1, the classical perturbation method is applied, whereas in Sect. 3.2, the initial value problem (IVP) (1)–(2) is solved using HAM. In Sect. 4.1, (1)–(2) is solved using the explicit RK method of fourth order, whereas in Sect. 4.2 a standard FDM method is used. In Sect. 4.3, a Galerkin FEM method is applied. Finally, in Sect. 5, a functional-analytic method is implemented for the solution of (1)–(2), not only for $t \in \mathbb{R}$ but also for $t \in \mathbb{C}$. Section 6 contains a discussion of the presented methods regarding their advantages and limitations.

This chapter has an expository character. However, it also includes some new results, such as the ones included in Sect. 5, as well as the application of multi-parameter perturbation techniques to (1)–(2) included in Sect. 3.1.

2 Qualitative Results

2.1 Dynamic Properties

As already mentioned in the previous section, Eq. (1) reduces to van-der Pol or Duffing equation for a specific choice of $(a, \beta, \gamma, f(t))$. It also includes of course the simple harmonic oscillator, for $a = \gamma = 0$ and $f(t) \equiv 0$. In this case (1) becomes

$$x'' + \beta x = 0, \quad \text{where} \quad x = x(t), \tag{3}$$

and the solution of (3), (2) is easily found to be

$$x(t) = \begin{cases} X_1 t + X_0 & \text{for} \quad \beta = 0 \\ X_0 \cosh(t\sqrt{-\beta}) + \frac{X_1}{\sqrt{-\beta}} \sinh(t\sqrt{-\beta}) & \text{for} \quad \beta < 0 \\ X_0 \cos(t\sqrt{\beta}) + \frac{X_1}{\sqrt{\beta}} \sin(t\sqrt{\beta}) & \text{for} \quad \beta > 0 \end{cases} . \tag{4}$$

Thus,

$$x'(t) = \begin{cases} X_1 & \text{for} \quad \beta = 0 \\ X_0\sqrt{-\beta} \sinh(t\sqrt{-\beta}) + X_1 \cosh(t\sqrt{-\beta}) & \text{for} \quad \beta < 0 \\ -X_0\sqrt{\beta} \sin(t\sqrt{\beta}) + X_1 \cos(t\sqrt{\beta}) & \text{for} \quad \beta > 0 \end{cases} . \tag{5}$$

By eliminating t from (4) and (5) in the case when $\beta \neq 0$ it is straightforward to find

$$\beta x^2 + \left(x'\right)^2 = \beta X_0^2 + X_1^2. \tag{6}$$

Relation (6) describes the trajectories of (3) in the phase plane (x, x'). Thus, the trajectories of (3) are ellipses when $\beta > 0$ and hyperbolas when $\beta < 0$.

Relation (6) can also be obtained directly from (3) without explicitly knowing $x(t)$ in the following way: Multiply (3) by x' to obtain

$$x''x' + \beta x x' = 0$$

and then integrate the preceding equation with respect to t. This gives

$$\frac{\left(x'\right)^2}{2} + \beta \frac{x^2}{2} = c_1 \Rightarrow \left(x'\right)^2 + \beta x^2 = c_2, \tag{7}$$

where $c_2 = 2c_1$ is an arbitrary constant. Taking into consideration (2) we find of course $c_2 = X_1^2 + \beta X_0^2$.

Relation (7) (or (6)) is a first integral of (3). A first integral of (1) can easily be found also in the case when $a = 0$ and $f(t) \equiv 0$, i.e. in then case when (1) reduces to a Duffing equation. In this case (1) becomes

$$x'' + \beta x + \gamma x^3 = 0$$

from which we find as before

$$x''x' + \beta x x' + \gamma x^3 x' = 0 \Rightarrow \left(x'\right)^2 + \beta x^2 + \gamma \frac{x^4}{2} = c,$$

where $c = X_1^2 + \beta X_0^2 + \gamma \frac{X_0^4}{2}$ after taking (2) into account.

However, finding a first integral of (1) in the general case is not an easy task. Actually, there is not much progress in this direction for (1), mainly because it fails to pass the Painlevé test, which states that (see [69, p. 9]) “An ODE in the complex domain is said to be of Painlevé type if the only movable singularities its solution can exhibit are poles.” Lately though, there exist a few results in this direction at least for some sets of values for (a, β, γ). In [15] it was proved that equation

$$x'' + \left(a_1 + \beta_1 x^2\right) x' - \gamma_1 x + x^3 = 0, \quad \text{where} \;\; x = x(t), \tag{8}$$

which is very similar to (1), possesses the first integral

$$x' + \frac{1}{\beta_1} x + \frac{\beta_1}{3} x^3 = I_1 e^{-\frac{3}{\beta_1} t}, \tag{9}$$

where I_1 an arbitrary constant, for

$$a_1 = \frac{4}{\beta_1}, \quad \gamma_1 = -\frac{3}{\beta_1^2}.$$

Moreover, for the above choice of parameters, the authors provide the general solution of (8). For the specific choice $I_1 = 0$, (9) gives easily

$$x = -\frac{\sqrt{3}}{\beta_1\sqrt{t_0 e^{\frac{2}{\beta_1}t} - 1}}, \quad t_0 > 1,$$

which is a particular solution of (8).

More recently, in [23] the more general than (1), with respect to the left hand side, equation

$$x'' + \left(a_2 + \beta_2 x^m\right)x' - \gamma_2 x + \delta_2 x^n = 0, \quad \text{where} \quad x = x(t),$$

was considered and it was proved that under the parametric conditions

$$a_2 = \frac{\delta_2}{\beta_2} - \frac{\gamma_2\beta_2}{\delta_2}, \quad n = m + 1$$

it admits the first integral

$$y' + \frac{\delta_2}{\beta_2}y + \frac{\beta_2}{m+1}y^{m+1} = I_2 e^{-\frac{\delta_2(m+1)}{\beta_2}w},$$

where $y = xt^{1/m}$, $w = -\frac{\beta_2}{\delta_2 m}\ln t$ and I_2 an arbitrary constant. It worths mentioning that in [22], the first integral of

$$x'' + \left(a_3 + \beta_3 x^m\right)x' - \gamma_3 x + \delta_3 x^{m+1} + \delta_4 x^n = 0, \quad \text{where} \quad x = x(t),$$

was obtained under certain parametric conditions.

Finally, it should be mentioned that the singularity analysis in complex t of (1) can provide evidence for its integrability or not. In [10], the non-integrability of (1) for $\beta = 1$ and $f(t) = \delta\cos(\omega t)$ was investigated by numerically studying the analytic properties of its solution for complex t.

Another approach for studying the trajectories of (1), for $f(t) \equiv 0$, is be rewriting it in the form of the system

$$\left.\begin{array}{l} x_1 = x \\ x_2 = x' \end{array}\right\} \overset{(1)}{\Rightarrow} \left.\begin{array}{l} x_1' = x_2 = f_1(x_1, x_2) \\ x_2' = ax_2\left(1 - x_1^2\right) - \beta x_1 - \gamma x_1^3 = f_2(x_1, x_2) \end{array}\right\}, \tag{10}$$

where $x_1 = x_1(t)$, $x_2 = x_2(t)$ and performing a standard phase plane analysis. The equilibrium points of (10) are obtained as the solutions of the algebraic system

$$\left.\begin{array}{l} f_1(x_1, x_2) = 0 \\ f_2(x_1, x_2) = 0 \end{array}\right\}.$$

If $\gamma = 0$ or $\beta/\gamma > 0$, the only equilibrium point of (10) is $(0, 0)$, whereas for $\beta/\gamma < 0$, (10) has the three equilibrium points $(0, 0)$, $\left(\pm\sqrt{-\frac{\beta}{\gamma}}, 0\right)$. The corresponding Jacobian matrix is

$$J(x_1, x_2) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -2ax_1x_2 - \beta - 3\gamma x_1^2 & a - a{x_1}^2 \end{pmatrix}.$$

Thus, by computing the eigenvalues of $J(x_1, x_2)$ at $(0, 0)$ and $\left(\pm\sqrt{-\frac{\beta}{\gamma}}, 0\right)$, the corresponding equilibrium points can be characterized (see for example [43, Chapter 2]). It can be easily found that

- $(0, 0)$ is
  - a saddle point for $\beta < 0$,
  - a node for $0 < \beta < \frac{a^2}{4}$, stable if $a < 0$ and unstable if $a > 0$,
  - a spiral point for $\beta > \frac{a^2}{4}$, stable if $a < 0$ and unstable if $a > 0$.
- $\left(\pm\sqrt{-\frac{\beta}{\gamma}}, 0\right)$ are
  - spiral points for $\Delta = a^2\left(1 + \frac{\beta}{\gamma}\right)^2 + 8\beta < 0$, stable when $a\left(1 + \frac{\beta}{\gamma}\right) < 0$ and unstable when $a\left(1 + \frac{\beta}{\gamma}\right) > 0$,
  - saddle points for $\Delta > 0$ and $\beta > 0$,
  - nodes for $\Delta > 0$ and $\beta < 0$, stable when $a\left(1 + \frac{\beta}{\gamma}\right) < 0$ and unstable when $a\left(1 + \frac{\beta}{\gamma}\right) > 0$.

The characterization of the equilibrium points (10) gives a first understanding of how its trajectories are formed in the phase plane (x_1, x_2). For example, periodic solutions are expected in the cases where the equilibrium points of (10) are characterized as spiral points. Periodic solutions of (10) are also predicted by specific standard theorems which can be found in almost all standard textbooks on nonlinear ODEs. One such theorem is the following:

Theorem 1 ([43, p. 299]) *The ODE*

$$x'' + h_1(x, x')x' + h_2(x) = 0, \quad x = x(t),$$

where h_1, h_2 are continuous, has at least one periodic solution under the following conditions:

(i) $\exists\, c > 0$ *such that* $h_1(x, y) > 0$ *when* $\sqrt{x^2 + y^2} > c$
(ii) $h_1(0, 0) < 0$
(iii) $h_2(0) = 0$, $h_2(x) > 0$ *when* $x > 0$ *and* $h_2(x) < 0$ *when* $x < 0$
(iv) $\int_0^x h_2(u)du \to \infty$, $x \to \infty$.

Another similar theorem is the following:

Theorem 2 ([43, p. 306]) *The ODE*

$$x'' + h_3(x)x' + h_4(x) = 0, \quad x = x(t),$$

where h_3, h_4 are continuous, has a unique periodic solution if

(i) $H(x) = \int_0^x h_3(u)du$ *is an odd function*
(ii) $H(x)$ *is zero only at* $x = 0$, $x = \pm c$ *for some* $c > 0$
(iii) $H(x) \to \infty$ *as* $x \to \infty$, *monotonically for* $x > c$.
(iv) $h_4(x)$ *is an odd function and* $h(x) > 0$ *for* $x > 0$.

It is a simple exercise to show that both Theorems 1 and 2 hold for (1), in the case when $f(t) \equiv 0$, for $a, \beta, \gamma > 0$.

Apart from these first, almost elementary results regarding the trajectories of (10), Eq. (1) exhibits a far more interesting and exotic dynamic behavior. It is considered after all, as one of the simplest nonlinear ODEs for which strange attractors, limit cycles and chaos have been observed. Indeed we can see some examples in Fig. 1, where the trajectories of (10) are depicted for specific values of (a, β, γ). The appearance of a limit cycle is obvious from the graphs in some cases. However, proving it rigorously is not an easy task. We shall not discuss such kind of topics here. Instead, we refer to [7, 42, 44, 51, 52, 61, 70, 74, 78, 79], where topics concerning the occurrence of chaos, strange attractors, limit cycles and bifurcations of ODEs of the form (1) were studied.

2.2 The Green Function

One of the most useful tools for the study and construction of the solution of a DE is its corresponding Green function. But what is it a Green function? In order to give a rough description, let's consider the non-homogeneous, linear ODE of order n of the general form:

$$\underbrace{y^{(n)}(t) + a_{n-1}(t)y^{(n-1)}(t) + \ldots + a_1(t)y'(t) + a_0(t)y(t)}_{Ly} = h(t), \tag{11}$$

where $h(t)$ and $a_i(t)$, $i = 1, \ldots, n-1$ known functions, accompanied with suitable initial at $t = t_0$ or boundary at $t = t_1$ and $t = t_2$ conditions. If the solution of (11) can be written in the integral form

$$y(t) = \int_I G(t, \tau)h(\tau)d\tau, \tag{12}$$

where $I = [t_0, t]$ or $I = [t_1, t_2] \subset \mathbb{R}$, the function $G(t, \tau)$ is called the Green function of the corresponding initial or boundary value problem. Of course there exists a rigorous mathematical definition for Green function (see for example [28, Chapter 2]), but we do not need it for the purposes of this chapter.

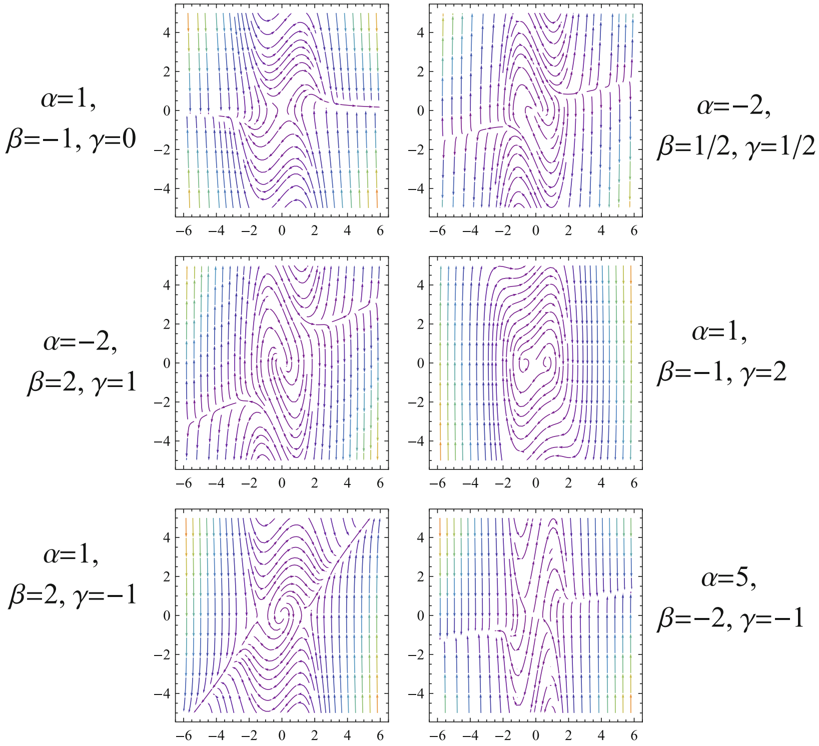


Fig. 1 Trajectories of (10) for various values of (a, β, γ)

The solution of (11) in the form (12) is quite useful, since it provides us with a way of finding the solution of (11) for any choice of $h(t)$ without resolving it every time. We just need to compute the integral on the right hand side of (12), even numerically if it cannot be expressed in terms of elementary or special functions.

Apart from being an important mathematical "object", Green function is also a very popular tool among engineers. It is associated with the boundary element method in numerical analysis and it has been used in the study of problems of practical interest, such as problems of structures, hydrology, fluid mechanics and elasticity (indicatively see [9, 26, 34, 75, 77]). With respect to strictly mathematical problems, the Green function has been associated mostly with positive solutions of boundary value problems. There is a large number of mathematical papers with results connected with Green function. It suffices to mention that MathSciNet search engine machine gives (today) approximately 4000 results for papers having the words "Green function" at their title.

There are several ways of constructing the Green function for a linear problem. Generally speaking, the Green function $G(t, \tau)$ for (11) satisfies the ODE

$$LG(t, \tau) = \delta(t - \tau), \tag{13}$$

where δ the Dirac delta function and the initial or boundary conditions that accompany (11). For further details and techniques see [3, 20, Lecture 16] or [50, Chapter 5]. For reasons of demonstration, we'll construct the Green function for (1)–(2) in the case when $a = \gamma = 0$ and $\beta > 0$, i.e. for the classical problem consisting of

$$x'' + \beta x = f(t), \quad x = x(t) \tag{14}$$

and the homogeneous initial conditions

$$x(0) = x'(0) = 0, \tag{15}$$

which describes forced, undamped oscillations with no initial displacement or velocity.

Instead of using (13), we'll use a simpler way to construct the Green function of (14)–(15) by actually solving it and then rewriting the formula for its solution in the form (12). Of course this way may not work with more complicated ODEs. The corresponding to (14) homogeneous equation is (3) and its general solution is

$$x_h(t) = c_1 \cos(t\sqrt{\beta}) + c_2 \sin(t\sqrt{\beta}),$$

where c_1, c_2 arbitrary constants. Using the method of variation of parameters, we seek a particular solution of (14) of the form

$$x_p(t) = c_1(t) \cos(t\sqrt{\beta}) + c_2(t) \sin(t\sqrt{\beta}),$$

where

$$\left.\begin{array}{l} c_1'(t)\cos(t\sqrt{\beta}) + c_2'(t)\sin(t\sqrt{\beta}) = 0 \\ -\sqrt{\beta}c_1'(t)\sin(t\sqrt{\beta}) + \sqrt{\beta}c_2'(t)\cos(t\sqrt{\beta}) = f(t) \end{array}\right\}.$$

Solving the preceding algebraic with respect to $c_1'(t)$ and $c_2'(t)$ system we find

$$c_1'(t) = -\frac{1}{\sqrt{\beta}} f(t)\sin(t\sqrt{\beta}), \quad c_2'(t) = \frac{1}{\sqrt{\beta}} f(t)\cos(t\sqrt{\beta})$$

from where we obtain

$$c_1(t) = -\frac{1}{\sqrt{\beta}} \int_0^t f(\tau)\sin(\tau\sqrt{\beta})d\tau, \quad c_2(t) = \frac{1}{\sqrt{\beta}} \int_0^t f(\tau)\cos(\tau\sqrt{\beta})d\tau.$$

As a consequence, the general solution of (14) is

$$x(t) = x_h(t) + x_p(t) = c_1\cos(t\sqrt{\beta}) + c_2\sin(t\sqrt{\beta})$$

$$+\frac{1}{\sqrt{\beta}}\sin(t\sqrt{\beta})\int_0^t f(\tau)\cos(\tau\sqrt{\beta})d\tau - \frac{1}{\sqrt{\beta}}\cos(t\sqrt{\beta})\int_0^t f(\tau)\sin(\tau\sqrt{\beta})d\tau$$

which after taking (15) into consideration becomes

$$x(t) = \frac{1}{\sqrt{\beta}} \int_0^t f(\tau)[\sin(t\sqrt{\beta}) cos(\tau\sqrt{\beta}) d\tau - \cos(t\sqrt{\beta}) \sin(\tau\sqrt{\beta})] d\tau$$

$$\Rightarrow x(t) = \int_0^t \frac{1}{\sqrt{\beta}} f(\tau) sin[(t-\tau)\sqrt{\beta}] d\tau. \tag{16}$$

Thus, the Green function of (14)–(15) is

$$G_1(t,\tau) = \frac{1}{\sqrt{\beta}} \sin[(t-\tau)\sqrt{\beta}].$$

From (16), we can immediately find the solution of (14)–(15) in two special cases of interest:

- when $f(t) = f_1(t) = \delta \cos(\omega t)$, $\omega \neq \pm\sqrt{\beta}$ and
- when $f(t) = f_2(t) = \delta \cos(t\sqrt{\beta})$.

In both cases, we have a periodic external force, but in the second case, resonance is present. In the first case it is

$$x_1(t) = \frac{\delta}{\sqrt{\beta}} \int_0^t \cos(\omega\tau) \sin[(t-\tau)\sqrt{\beta}] d\tau = \frac{\delta}{\beta - \omega^2} \left[\cos(\omega t) - \cos(t\sqrt{\beta})\right],$$

whereas in the second case it is

$$x_2(t) = \frac{\delta}{\sqrt{\beta}} \int_0^t \cos(\tau\sqrt{\beta}) \sin[(t-\tau)\sqrt{\beta}] d\tau = \frac{\delta}{2\sqrt{\beta}} t \sin(t\sqrt{\beta}).$$

Let's return to (1)–(2) and consider the auxiliary linear problem

$$\begin{gathered} x'' = f(t), \quad x = x(t) \\ x(0) = x'(0) = 0. \end{gathered} \tag{17}$$

As in the case of (14)–(15), we can construct the Green function for (17) and we find it to be

$$G_2(t,\tau) = t - \tau.$$

It is now a simple exercise (similar to exercise 16.7 p. 125 of [3]) to prove the following:

Proposition 1 *The function $x(t)$ is a solution of (1)–(2) if and only if*

$$x(t) = X_0 + X_1 t + \int_0^t G_2(t,\tau) F(\tau, x, x') d\tau, \tag{18}$$

where $F(t,x,x') = f(t) + a(1-x^2)x' - \gamma x^3 - \beta x$ *and* $G_2(t,\tau)$ *is the Green function of* (17).

Proof

(⇐) Suppose (18) holds. We'll show that (18) satisfies (1)–(2). Differentiating (18) twice with respect to t we find

$$x'(t) = X_1 + G_2(t,t)F(t,x,x') + \int_0^t \frac{\partial G_2(t,\tau)}{\partial t} F(\tau,x,x')d\tau$$

$$\Rightarrow x'(t) = X_1 + \int_0^t F(\tau,x,x')d\tau \tag{19}$$

$$\Rightarrow x''(t) = F(t,x,x'),$$

which is (1). Moreover, from (18) and (19) we find the initial conditions (2).

(⇒) Suppose (1)–(2) holds. Then

$$\int_0^t x''(\tau)d\tau = \int_0^t F(\tau,x,x')d\tau \Rightarrow x'(t) - X_1 = \int_0^t F(\tau,x,x')d\tau$$

$$\Rightarrow \int_0^t x'(\tau)d\tau = \int_0^t X_1 d\tau + \int_0^t \int_0^t F(\tau,x,x')d\tau d\tau$$

$$\Rightarrow x(\tau) - X_0 = X_1 t + \int_0^t (t-\tau)F(\tau,x,x')d\tau,$$

which is (18).

Proposition 1 enables us to rewrite (1)–(2) as an integrodifferential equation of Volterra type. Several fixed point theorems may be applied to (18), in order to ensure the existence and/or uniqueness of solutions of (18) and consequently of (1)–(2) in specific spaces. Especially in the case when $a = 0$, (1)–(2) is equivalent to the integral equation

$$x(t) = X_0 + X_1 t + \int_0^t G_2(t,\tau)F_1(\tau,x)d\tau,$$

where $F_1(t,x) = f(t) - \gamma x^3 - \beta x.$

Apart from the fact that (18) is a convenient form to rewrite (1)–(2) in order to obtain some qualitative type of results, it also gives an alternative way to obtain the solution of (1)–(2) numerically. Instead of applying a numerical method to the IVP (1)–(2), we may apply a numerical method to the integrodifferential equation (18). This is expected to decrease the errors when using an FDM method as discussed in Sect. 4.2. Moreover, (18) is called the "weak form" of (1)–(2). Having the weak form of a problem for a DE is the first step in the attempt to solve it using a FEM method.

3 Approximate Techniques

3.1 Classical Perturbation Method

Perturbation techniques are a very popular tool for approximately solving a DE providing that a small positive parameter $0 < \epsilon << 1$ appears in the DE. Sometimes, when no such parameter appears, it is artificially enforced at the DE, in such a way so that the new DE with the ϵ, to be reduced to the initial DE for a specific value of ϵ. The fact that a parameter must be present in the DE under consideration is one limitation of the perturbation methods. Another limitation is that this parameter must be much smaller than 1.

When we apply the classical perturbation technique to an ODE of the form

$$G(\epsilon; t, y, y', y'', \ldots, y^n) = 0, \tag{20}$$

where $y = y(t)$ and $0 < \epsilon << 1$, we seek for solutions of (20) of the form

$$y(t) = \sum_{m=0}^{\infty} y_m(t)\epsilon^m, \tag{21}$$

where the coefficients $y_m(t)$ are functions to be determined. The series (21) is called a perturbation series. The first coefficient $y_0(t)$ is called the leading order term, whereas the terms $y_m(t)\epsilon^m$, $m \neq 0$ are called the mth correction of the solution of (20). Usually, it suffices to consider just the first few (two or three) terms of the perturbation series in order to obtain a good approximation of the solution of (20). In this case the solution of (20) is given by the approximate expression

$$y_a(t) = y_0(t) + y_1(t)\epsilon + y_2(t)\epsilon^2 + \ldots + y_i(t)\epsilon^i + O\left(\epsilon^{i+1}\right)$$

$$\Rightarrow y_a(t) \simeq y_0(t) + y_1(t)\epsilon + y_2(t)\epsilon^2 + \ldots + y_i(t)\epsilon^i.$$

In the case when the approximate solution $y_a(t)$ converges to the exact solution of (20) at some well-defined rate as $\epsilon \to 0$, the approximation is said to be uniform and the perturbation is successful.

In order to compute the coefficients of (21), we substitute (21) to (20) and equate the coefficients of the ϵ^m appearing in both sides of the new equation. In this way, instead of solving (20), we need to solve a series (usually two or three) of new ODEs, in each of which the unknown function is $y_m(t)$, $m = 0, 1, 2, \ldots$. In order for the classical perturbation to be successful, the ODE with unknown the leading order term $y_0(t)$ must satisfy the following criteria:

- It should be the same with (20) for $\epsilon = 0$. Actually the ODE
$$G(0; t, y, y', y'', \dots, y^n) = 0$$
is called the unperturbed equation.
- It should be easier than (20).
- It should be of the same order as (20).
- We should be able to find its solution in closed form.

Such kind of problems, where the classical perturbation method is successful, are characterized as regular. If the preceding criteria are not fulfilled, then the classical perturbation technique fails and other methods should be used. Special care is needed in the cases when a power of ϵ is multiplied with $y^{(n)}(t)$. In these cases, the ODE for $y_0(t)$ is not of the same order as (20) and singular perturbation techniques must be employed.

In the IVP (1)–(2), three parameters appear, namely a, β and γ. Let's take $0 < a << 1$ and assume that (1)–(2) has a solution of the form

$$x(t) = x_0(t) + ax_1(t) + a^2x_2(t) + O\left(a^3\right). \tag{22}$$

Substituting (22) into (1) and equating the coefficients of the a^m, $m = 0, 1, 2$ appearing in both sides of the new equation, we find after some manipulations that the leading order term satisfies the IVP

$$\begin{gathered} x_0''(t) + \beta x_0(t) + \gamma x_0^3(t) = f(t), \\ x_0(0) = X_0, \quad x_0'(0) = X_1. \end{gathered} \tag{23}$$

If we choose $0 < \beta << 1$ and assume that (1)–(2) has a solution of the form

$$x(t) = \tilde{x}_0(t) + \beta\tilde{x}_1(t) + \beta^2\tilde{x}_2(t) + O\left(\beta^3\right),$$

we find as before that the leading order term satisfies the IVP

$$\begin{gathered} \tilde{x}_0'' + a\left(\tilde{x}_0^2 - 1\right)\tilde{x}_0' + \gamma\tilde{x}_0^3 = f(t), \\ \tilde{x}_0(0) = X_0, \quad \tilde{x}_0'(0) = X_1. \end{gathered} \tag{24}$$

Finally, if we choose $0 < \gamma << 1$ and assume that (1)–(2) has a solution of the form

$$x(t) = \hat{x}_0(t) + \gamma\hat{x}_1(t) + \gamma^2\hat{x}_2(t) + O\left(\gamma^3\right),$$

we find in the same way that the leading order term satisfies the IVP

$$\begin{gathered} \hat{x}_0'' + a\left(\hat{x}_0^2 - 1\right)\hat{x}_0' + \beta\hat{x}_0 = f(t), \\ \hat{x}_0(0) = X_0, \quad \hat{x}_0'(0) = X_1. \end{gathered} \tag{25}$$

All three problems (23)–(25) consist of nonlinear ODEs (the corresponding unperturbed ones) which cannot be solved analytically in closed form for all values of a, β and γ. Thus, the classical perturbation technique fails for (1)–(2). However, by assuming all three parameters very small, we may seek solutions of the form

$$x(t) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} x_{i,j,k}(t) a^i \beta^j \gamma^k. \tag{26}$$

Such a series is a three-parameter perturbation series and the corresponding method is a three-parameter perturbation method. Assuming i, j, k taking only the first two values, i.e. 0 and 1, we seek for solutions of (1)–(2) of the form:

$$\begin{aligned} x(t) = \sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1} x_{i,j,k}(t) a^i \beta^j \gamma^k + \ldots = x_{0,0,0}(t) + a x_{1,0,0}(t) + \beta x_{0,1,0}(t) \\ + \gamma x_{0,0,1}(t) + a\beta x_{1,1,0}(t) + a\gamma x_{1,0,1}(t) + \beta\gamma x_{0,1,1}(t) + a\beta\gamma x_{1,1,1}(t) + \ldots . \end{aligned} \tag{27}$$

Substituting (27) into (1)–(2), keeping only terms of $a^i \beta^j \gamma^k$ as those appearing in (27) and equating the coefficients of the corresponding terms appearing in both sides of the new equations, we end up with the IVPs of Table 1.

Table 1 IVPs for $x_{i,j,k}(t)$ of (27)

IVP1	$x''_{0,0,0}(t) = f(t)$, $x_{0,0,0}(0) = X_0,\ x'_{0,0,0}(0) = X_1$
IVP2	$x''_{1,0,0}(t) = x'_{0,0,0}(t) - x^2_{0,0,0}(t)x'_{0,0,0}(t)$, $x_{1,0,0}(0) = 0,\ x'_{1,0,0}(0) = 0$
IVP3	$x''_{0,1,0}(t) = -x_{0,0,0}(t)$, $x_{0,1,0}(0) = 0,\ x'_{0,1,0}(0) = 0$
IVP4	$x''_{0,0,1}(t) = -x^3_{0,0,0}(t)$, $x_{0,0,1}(0) = 0,\ x'_{0,0,1}(0) = 0$
IVP5	$x''_{1,1,0}(t) = -x_{1,0,0}(t) - x^2_{0,0,0}(t)x'_{0,1,0}(t)$ $+x'_{0,1,0}(t) - 2x_{0,0,0}(t)x_{0,1,0}(t)x'_{0,0,0}(t)$, $x_{1,1,0}(0) = 0,\ x'_{1,1,0}(0) = 0$
IVP6	$x''_{1,0,1}(t) = -x^2_{0,0,0}(t)x'_{0,0,1}(t) - 3x^2_{0,0,0}(t)x_{1,0,0}(t)$ $+x'_{0,0,1}(t) - 2x_{0,0,0}(t)x_{0,0,1}(t)x'_{0,0,0}(t)$, $x_{1,0,1}(0) = 0,\ x'_{1,0,1}(0) = 0$
IVP7	$x''_{0,1,1}(t) = -x_{0,0,1}(t) - 3x^2_{0,0,0}(t)x_{0,1,0}(t)$, $x_{0,1,1}(0) = 0,\ x'_{0,1,1}(0) = 0$
IVP8	$x''_{1,1,1}(t) = -x_{1,0,1}(t) - 3x^2_{0,0,0}(t)x_{1,1,0}(t) - x^2_{0,0,0}(t)x'_{0,1,1}(t)$ $+x'_{0,1,1}(t) - 6x_{0,0,0}(t)x_{0,1,0}(t)x_{1,0,0}(t) - 2x_{0,0,1}(t)x_{0,1,0}(t)x'_{0,0,0}(t) -$ $2x_{0,0,0}(t)x_{0,1,1}(t)x'_{0,0,0}(t) - 2x_{0,0,0}(t)x_{0,0,1}(t)x'_{0,1,0}(t) - 2x_{0,0,0}(t)x_{0,1,0}(t)x'_{0,0,1}(t)$, $x_{1,1,1}(0) = 0,\ x'_{1,1,1}(0) = 0$

Table 2 Solutions of the IVP1–IVP8 for $X_0 = 0 = X_1$ and $f(t) = \cos t$

$x_{0,0,0}(t) = 1 - \cos t$
$x_{1,0,0}(t) = \frac{1}{36}\,[6t + 9\sin t - 9\sin(2t) + \sin(3t)]$
$x_{0,1,0}(t) = \frac{1}{2}\Big(2 - t^2 - 2\cos t\Big)$
$x_{0,0,1}(t) = \frac{1}{72}\Big[245 - 90t^2 - 270\cos t + 27\cos(2t) - 2\cos(3t)\Big]$
$x_{1,1,0}(t) = \frac{1}{648}\Big[-1059t + 36t^3 - 1296t\cos t + 81t\cos(2t)$ $+3240\sin t - 648t^2\sin t - 567\sin(2t) + 81t^2\sin(2t) + 56\sin(3t)\Big]$
$x_{1,0,1}(t) = \frac{1}{172800}\Big[-1296420t + 14400t^3 - 1036800t\cos t + 64800t\cos(2t)$ $+3232800\sin t - 432000t^2\sin t - 612000\sin(2t) + 54000t^2\sin(2t)$ $+96800\sin(3t) - 8175\sin(4t) + 384\sin(5t)]$
$x_{0,1,1}(t) = \frac{1}{1296}\Big[40936 - 7065t^2 + 378t^4 - 42768\cos t + 3888t^2\cos t + 1944\cos(2t)$ $+486t\sin(2t) - 243t^2\cos(2t) - 112\cos(3t) - 15552t\sin t\Big]$
$x_{1,1,1}(t) = \frac{1}{74649600}\Big[-7770555660t + 385632000t^3 - 20113920t^5 - 9839232000t\cos t$ $+1082160000t\cos(2t) + 559872000t^3\cos t - 9331200t^3\cos(2t) - 136166400t\cos(3t)$ $+3385800t\cos(4t) + 22184064000\sin t - 4100544000t^2\sin t + 136857600t^4\sin t$ $+901238400t^2\cos t\sin t - 3380508000\sin(2t) - 5443200t^4\sin(2t) + 439014400\sin(3t)$ $+847872\sin(5t) - 41817600t^2\sin(3t]) - 20980575\sin(4t) + 1765800t^2\sin(4t)\Big]$

Notice that the ODEs involved in the IVPs 1–8 are linear with respect to their unknown functions $x_{i,j,k}(t)$ and thus, can be solved explicitly. Let's find in this way, the approximate solution of (1)–(2) in the case when $X_0 = 0 = X_1$ and $f(t) = \cos t$. In this case, the solutions of the IVPs 1–8 are given in Table 2. Thus, the approximate solution of (1)–(2) is (27) for $x_{i,j,k}(t)$ $i, j, k = 0, 1$ as in Table 2. This solution is depicted in Fig. 2, with dashed line, for $a = 0.002$, $\beta = 0.008$ and $\gamma = 0.003$. At the same figure, the corresponding numerical solutions of (1)–(2) obtained with the fourth order RK and FEM are also depicted. We can easily verify in this case, that the approximate solution obtained by the three-parameter classical perturbation method is in very good agreement with the corresponding numerical ones, although we considered very few terms of the perturbation series (26).

In Fig. 3, we see with dashed line again, the graph of (27) in the case when $X_0 = 0 = X_1$ and $f(t) = \cos t$ and the parameters a, β and γ are larger than before, namely for $a = 0.02$, $\beta = 0.08$ and $\gamma = 0.03$. The corresponding fourth order RK numerical solution of (1)–(2) in this case is also depicted with a continuous line. We notice that now the two solutions are in good agreement up to $t \simeq 4.5$ and after that point, the approximate solution slowly but steadily, is drawn away from the corresponding numerical one. This may be due to the fact that we have considered only very few terms of the perturbation series (26). Another reason for this diver-

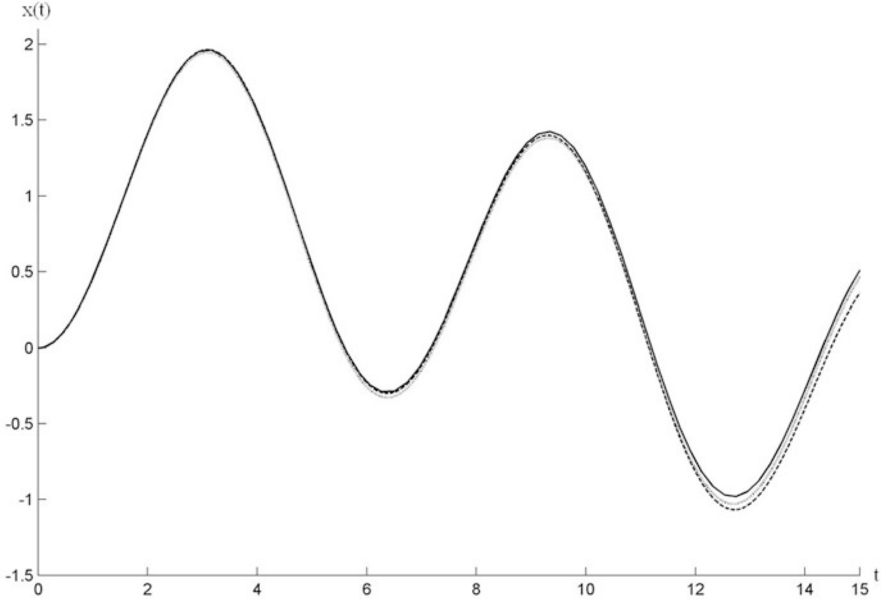


Fig. 2 Graph of (27) (dashed line), numerical solution with RK4 (continuous line) and numerical solution with FDM (dotted line) of (1)–(2), for $X_0 = X_1 = 0$, $f(t) = \cos t$, $a = 0.002$, $\beta = 0.008$ and $\gamma = 0.003$

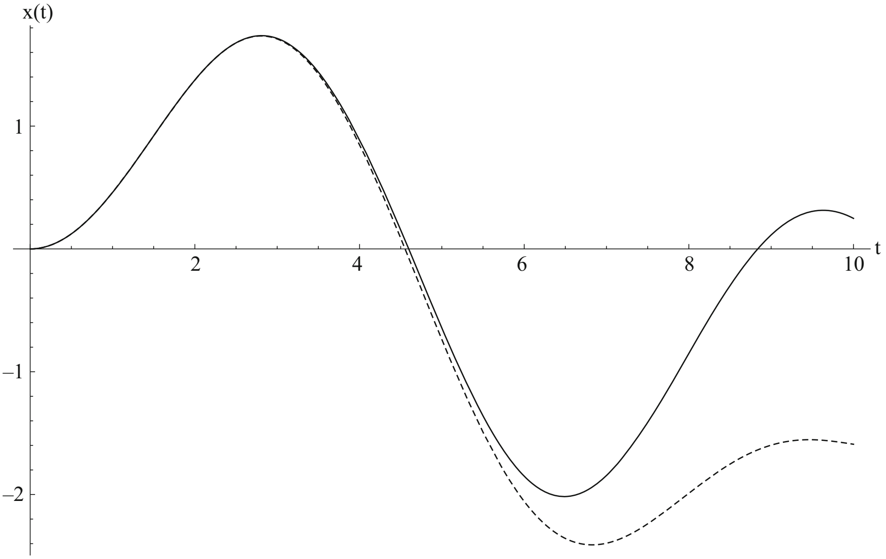


Fig. 3 Graph of (27) (dashed line) and numerical solution (continuous line) of (1)–(2), for $X_0 = X_1 = 0$, $f(t) = \cos t$, $a = 0.02$, $\beta = 0.08$ and $\gamma = 0.03$

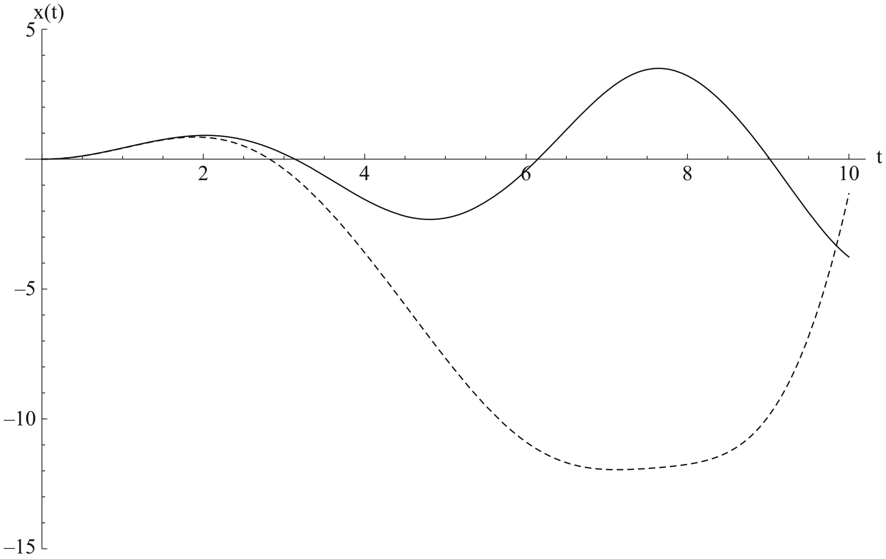


Fig. 4 Graph of (27) (dashed line) and numerical solution (continuous line) of (1)–(2), for $X_0 = X_1 = 0$, $f(t) = \cos t$, $a = 0.02$, $\beta = 1$ and $\gamma = 0.03$

gence, is the appearance of terms like $t^m \cos(\phi_1 t)$ or $t^n \sin(\phi_2 t)$ in the approximate solution. Such kind of terms are called secular terms and they may increase rapidly with t, regardless of the fact that a, β, γ, $\cos(\phi_1 t)$ and $\sin(\phi_2 t)$ are small. One way to deal with this problem is to "refine" the perturbation method in such a way so that secular terms do not appear in the solution. This may be done for example with the Lindstedt-Poincaré method (see for example [50, Chapter 2] and [56]).

Now let's "forget" for a moment that all parameters were assumed much smaller than 1 and consider the case when $a = 0.02$, $\beta = 1$ and $\gamma = 0.03$ again for $X_0 = X_1 = 0$, $f(t) = \cos t$. In this case, the graph (dashed line) of (27) is shown in Fig. 4. With the continuous line the corresponding fourth order RK numerical solution is also depicted at the same figure. It's obvious that the classical three-parameter perturbation method for (1)–(2) fails "triumphantly" and this is due to the violation of the assumption that β must be much smaller than 1.

However, not all parameters need to be simultaneously small. Let' assume $\beta = 1$, $0 < a, \gamma << 1$ and seek solutions of (1)–(2) of the form

$$x(t) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} x_{i,j}(t) a^i \gamma^j + \ldots = x_{0,0}(t) + a x_{1,0}(t) + \gamma x_{0,1}(t) + a\gamma x_{1,1}(t) + \ldots, \tag{28}$$

i.e. we apply a two-parameter classical perturbation method. As before, we are led to the IVPs of Table 3.

For $f(t) = \cos t$ and $X_0 = 0 = X_1$, the graph of (28) of (1)–(2) for the aforementioned values of parameters is shown in Fig. 5 with dashed line. With the continuous line we graph again the corresponding fourth order RK

Table 3 IVPs for $x_{i,j}(t)$ of (28)

IVP1	$x''_{0,0}(t) + x_{0,0}(t) = f(t)$, $x_{0,0}(0) = X_0,\ x'_{0,0}(0) = X_1$
IVP2	$x''_{1,0}(t) + x_{1,0}(t) = x'_{0,0}(t) - x^2_{0,0}(t)x'_{0,0}(t)$, $x_{1,0}(0) = 0,\ x'_{1,0}(0) = 0$
IVP3	$x''_{0,1}(t) + x_{0,1}(t) = -x^3_{0,0}(t)$, $x_{0,1}(0) = 0,\ x'_{0,1}(0) = 0$
IVP4	$x''_{1,1}(t) + x_{1,1}(t) = -3x^2_{0,0}(t)x_{1,0}(t) - x^2_{0,0}(t)x'_{0,1}(t)$ $+x'_{0,1}(t) - 2x_{0,0}(t)x_{0,1}(t)x'_{0,0}(t)$ $x_{1,1}(0) = 0,\ x'_{1,1}(0) = 0$

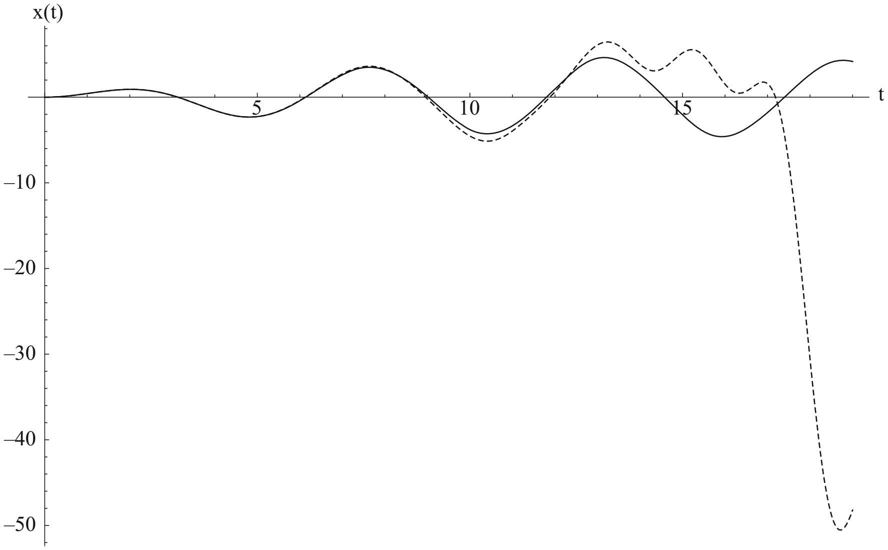


Fig. 5 Graph of (28) (dashed line) and numerical solution (continuous line) of (1)–(2), for $X_0 = X_1 = 0$, $f(t) = \cos t$, $a = 0.02$, $\beta = 1$ and $\gamma = 0.03$

numerical solution of (1)–(2). It is obvious, that now the two-parameter classical perturbation method gives an approximate solution which is in good agreement with the corresponding numerical solution up to $t \simeq 13$, but after that point the approximate solution diverges from the numerical one. This is again due to the few terms of (28) considered or/and due to the appearance of secular terms in the approximate solution.

It would be natural to think that a two-parameter perturbation method with perturbation parameters (a, β) or (β, γ) could be successfully applied to (1)–(2). However, this is not true, because in these cases the IVP for the leading order term $x_{0,0}(t)$ is nonlinear and cannot be solved analytically, at least not for all values of a, β and γ.

3.2 Homotopy Analysis Method

The homotopy analysis method is similar to perturbation methods, in the sense that instead of the original problem, a series of simpler linear problems are solved. (Remember that when using a perturbation method, the simpler problems for the leading order term and the first, second, etc corrections may still be nonlinear.) However, HAM does not require the existence of a small parameter in the equation under consideration and thus is more flexible and can be used in a greater amount of problems compared to perturbation methods. It has already been successfully applied to a variety of problems such as problems of nonlinear oscillations, deformation of beams, boundary layer flows in fluid mechanics, solitary waves, etc.

As its name suggests, HAM is based on the notion of homotopy. In topology, a homotopy is a map $H : X \times I \to Y$, where $I = [0, 1]$, such that $H(t, 0) = g_0(t)$ and $H(t, 1) = g_1(t)$, where $g_0, g_1 : X \to Y$ (see [32, p. 1]). The basic idea of HAM for ODEs can be summarized as follows:

Given a nonlinear ODE

$$\mathscr{A}[y(t)] = 0, \tag{29}$$

where $\mathscr{A}$ a nonlinear differential operator and $y(t)$ the unknown function, find a homotopy $\mathscr{H}[\Phi(t; q), q]$, $q \in [0, 1]$ and an initial approximation $y_0(t)$ of $y(t)$, in such a way that

$$\mathscr{H}[\Phi(t; q), q]|_{q=0} = 0$$

has as solution the function $y_0(t)$ and

$$\mathscr{H}[\Phi(t; q), q]|_{q=1} = 0$$

has as solution, the function $y(t)$, i.e. the solution of (29). The parameter q is often referred to as embedding parameter.

In most cases, (29) is accompanied by initial or/and boundary conditions. Thus, $y_0(t)$ is a suitable function (the word suitable will be soon clarified) satisfying the accompanying conditions.

One suitable homotopy proposed by He in [30] is the following:

$$\mathscr{H}_1[\Phi(t; q), q] = (1 - q)\mathscr{L}_1[\Phi(t; q) - y_0(t)] + q\mathscr{A}[\Phi(t; q)], \tag{30}$$

where $\mathscr{L}_1$ is a linear operator “inspired” by (29) with the property

$$\mathscr{L}_1[g] = 0 \ \text{ when } \ g = 0. \tag{31}$$

Another, more general, suitable homotopy proposed by Liao (see for example [46]) is the following:

$$\mathscr{H}_2[\Phi(t; q), q] = (1 - q)\mathscr{L}_2[\Phi(t; q) - y_0(t)] - q\hbar H(t)\mathscr{A}[\Phi(t; q)], \tag{32}$$

where $\mathscr{L}_2$ is again a linear operator "inspired" by (29) with the property (31), $\hbar$ is a nonzero auxiliary parameter also called convergence control parameter and $H(t)$ is a nonzero auxiliary function. Notice that (30) can be obtained from (32) for $\hbar = -1$ and $H(t) \equiv 1$. Observe also that for $i = 1, 2$ it is

$$\mathscr{H}_i[\Phi(t;q), q]|_{q=0} = 0 \Rightarrow \mathscr{L}_i\ [\Phi(t;q) - y_0(t)]|_{q=0} = 0 \overset{(31)}{\Rightarrow} \Phi(t;0) = y_0(t) \tag{33}$$

and

$$\mathscr{H}_i[\Phi(t;q), q]|_{q=1} = 0 \Rightarrow \mathscr{A}\ [\Phi(t;q)]|_{q=1} = 0, \tag{34}$$

which is (29). This means that as q varies from 0 to 1, the solution $\Phi(t;q)$ of $\mathscr{H}_i[\Phi(t;q), q] = 0$, $i = 1, 2$ varies from $y_0(t)$ to $y(t)$.

Since $\mathscr{H}_1$ is a specific case of $\mathscr{H}_2$, we shall further proceed with explaining HAM as proposed by Liao, see for example [46]. In order to further connect $\Phi(t;q)$ with $y(t)$, let's use Taylor's theorem to write

$$\Phi(t;q) = \Phi(t;0) + \sum_{k=1}^{\infty} \frac{1}{k!} \left.\frac{\partial^k[\Phi(t;q)]}{\partial q^k}\right|_{q=0}^{q^k}$$

or after introducing the notation $y_k(t) = \frac{1}{k!} \left.\frac{\partial^k[\Phi(t;q)]}{\partial q^k}\right|_{q=0}$ ($y_k(t)$ are called k-th order deformation derivatives)

$$\Phi(t;q) = \Phi(t;0) + \sum_{k=1}^{\infty} y_k(t) q^k \overset{(33)}{=} y_0(t) + \sum_{k=1}^{\infty} y_k(t) q^k. \tag{35}$$

Assuming that:

(A1) $\mathscr{H}_2[\Phi(t;q), q] = 0$ has a solution for all $q \in [0, 1]$,
(A2) $y_k(t)$ exist for all $k = 1, 2, 3, \ldots$ and
(A3) the power series on the right hand side of (33) converges for $q = 1$,

we can write

$$y(t) = y_0(t) + \sum_{k=1}^{\infty} y_k(t), \tag{36}$$

which is obtained from (35) for $q = 1$ after taking into consideration (34) and (29). Now the "only thing" left is a way to find $y_k(t)$. It can be shown [46, Chapter 3], that these are obtained by solving the linear ODEs, called high-order deformation equations

$$\mathscr{L}_2[y_k(t) - \chi_k y_{k-1}(t)] = \hbar H(t) \mathscr{R}_k[y_0(t), \ldots y_{k-1}(t), t] \tag{37}$$

where

$$\chi_k = \begin{cases} 0, & k = 1 \\ 1, & k = 2, 3, \ldots \end{cases},$$

$$\mathscr{R}_k[y_0(t), \ldots, y_{k-1}(t), t] = \frac{1}{(k-1)!} \left. \frac{\partial^{k-1}[\mathscr{A}(\Phi(t;q))]}{\partial q^{k-1}} \right|_{q=0} \tag{38}$$

and $y_k(t)$ satisfy suitable initial or/and boundary conditions. Since $y_0(t)$ is known, $y_1(t)$ can be obtained from (37) for $k = 1$, then $y_2(t)$ for $k = 2$ and so on. It can be proved that as long as the series in the right hand side of (36) converges, where $y_k(t)$ are obtained from (37), (36) is a solution of (29).

Up to now, we have said nothing about how $\mathscr{L}_2$, $\hbar$ and $H(t)$ are chosen. Generally speaking, they should be chosen in such a way that assumptions (A1)–(A3) are satisfied. Thus, HAM provides us with great freedom on how to choose all these. However, for practical reasons we need to follow some fundamental rules. First of all, we must choose a set of base functions, for example $S_B = \{e_i(t), \quad i = 0, 1, 2, \ldots\}$, which will "guide" us throughout the implementation of HAM. The first rule, called the rule of solution expression, states that the solution $y(t)$ of (29) should be represented by $e_i(t)$, i.e. we should be able to write $y(t)$ as a series of $e_i(t)$, i.e.

$$y(t) = \sum_{i=0}^{\infty} c_i e_i(t), \tag{39}$$

where c_i, $i = 0, 1, 2, \ldots$ coefficients. The same rule tells us how to choose $y_0(t)$ and $\mathscr{L}_2$: $y_0(t)$ can be any function expressed as a combination of some $e_i(t)$ satisfying the conditions accompanying (29) and $\mathscr{L}_2$, should be such that the solution of (37), i.e. $y_k(t)$ to be able to be expressed as a combination of $e_i(t)$. Furthermore, $H(t)$ should be chosen in such a way so that the solutions $y_k(t)$, $k = 1, 2, \ldots$ of (37) to be again expressed as a combination of $e_i(t)$. At this stage, $H(t)$ may not be uniquely determined.

The second rule, called the rule of coefficient ergodicity, states that each coefficient $c_{k,i}$ of $y_k(t)$ should be able to be modified as the order of the approximation tends to infinity. This rule, together with the first rule, gives us a unique $H(t)$ in many cases. The third and last rule, called the rule of solution existence, states that $y_0(t)$, $H(t)$ and $\mathscr{L}_2$ should be chosen in such way that all Eq. (37) can be solved and their solutions are in closed form.

In practice, the implementation of HAM involves many calculations, since the ODEs (37) must be solved for as many k as possible. However, this is not a problem nowadays that programs like Mathematica or Maple are available. The whole procedure is facilitated if S_B contains relatively simple functions. There also exist some packages for automatic derivation of HAM solutions for nonlinear periodic oscillators (see [48] and [49]).

Finally, how do we choose $\hbar$? Suppose $y_{ap}(t) = y_0(t) + \sum_{k=1}^{K} y_k(t)$ is an approximation of the solution $y(t)$ of (29) obtained by HAM. This $y_{ap}(t)$ depends also on $\hbar$. One simple way to determine $\hbar$ is the following: If the order of (29) is n, find $y_{ap}^{(n)}(t_0)$ for some t_0 and plot it against $\hbar$. This graph is called $\hbar$-curve. Assuming (36) convergent, it should be the solution of (29). Thus, $y_{ap}^{(n)}(t_0)$ should always have the same value. This means that we should be able to see a horizontal line segment at the $\hbar$-curve. The region of $\hbar$ for which this horizontal line segment appears is called the valid region of $\hbar$ and if we choose a value of $\hbar$ from this region, we're quite sure that (36) converges. The quantity $y_{ap}^{(n+1)}(t_0)$ may have a physical meaning in certain applications. Moreover, there may exist more than one important quantities with physical meaning. Thus, we can plot all of them against $\hbar$ and find more $\hbar$-curves. The more $\hbar$-curves we plot, the clearer should be how to choose the value of $\hbar$. Another way to choose $\hbar$ is by finding the optimum $\hbar$ for which the residual or the average residual error between $y_{ap}(t)$ and $y(t)$ is minimum, but we'll confine ourselves here to the use of $\hbar$-curves. For some recent developments on HAM we refer to [31] and [47].

In [16], He's HAM was applied to an equation similar to (1), namely equation

$$x'' + a_1 x + a_3 x^3 = a_2 \left(1 - x^2 - x'^2\right) x', \quad x = x(t),$$

where $a_i, i = 1, 2, 3$ are constant parameters. Also, in [17], the limit cycle of (1) for $f(t) \equiv 0$ and $\beta = 1$ was studied using Liao's HAM.

In order to demonstrate Liao's HAM, we'll apply it to (1)–(2) in the simple case when $a = 0$, $\beta = 1$, $f(t) \equiv 0$, $X_0 = 1$, $X_1 = 0$ and $\gamma = 100$, i.e. for the IVP

$$x'' + x + 100x^3 = 0, \quad x = x(t), \tag{40}$$

$$x(0) = 1, \quad x'(0) = 0. \tag{41}$$

The only reason to choose such a large value for γ is in order to show that HAM is not at all affected by it. Actually the same problem for any value of γ (and other more general) is already solved by HAM in various papers and also in [46, p. 171]. Moreover, in order (a) to keep S_B simple and (b) to gain further information regarding the frequency ω of the oscillations described by (40)–(41), we first of all make the simple change of variables:

$$\tau = \omega t, \quad x(t) = x\left(\frac{\tau}{\omega}\right) = u(\tau). \tag{42}$$

Then (40)–(41) becomes:

$$\omega^2 u'' + u + 100u^3 = 0, \quad u = u(\tau), \tag{43}$$

$$u(0) = 1, \quad u'(0) = 0. \tag{44}$$

It has been found in several papers that a good set of base functions that can describe oscillation problems is the set

$$S_B = \{\cos(m\tau) \;\; m = 1, 2, 3, \ldots\}.$$

Having this in mind, as well as (43), we choose the corresponding linear operator to be

$$\mathscr{L}_3[\Phi(\tau; q)] = \omega_0^2 \left[\Phi''(\tau; q) + \Phi(\tau; q)\right]$$

where the $'$ denotes differentiation with respect to t and ω_0 is a coefficient to be determined. Also, taking this S_B into consideration, we choose the initial approximation of $u(\tau)$ to be

$$u_0(\tau) = \cos\tau, \tag{45}$$

which of course satisfies (44). Finally, we choose $\mathscr{A}$ to be

$$\mathscr{A}[\Phi(\tau; q), \Omega(q)] = \Omega^2(q)\Phi''(\tau; q) + \Phi(\tau; q) + 100\Phi^3(\tau; q),$$

where as already mentioned

$$\Phi(\tau; q) = u_0(\tau) + \sum_{k=1}^{\infty} u_k(\tau) q^k \tag{46}$$

and in an analogous way

$$\Omega(q) = \omega_0 + \sum_{k=1}^{\infty} \omega_k q^k, \tag{47}$$

with $u_k(\tau) = \frac{1}{k!} \left.\frac{\partial^k[\Phi(\tau;q)]}{\partial q^k}\right|_{q=0}$ and $\omega_k = \frac{1}{k!} \left.\frac{\partial^k[\Omega(q)]}{\partial q^k}\right|_{q=0}$.

With the aid of $\mathscr{A}$ we can find $\mathscr{R}_k$ defined by (38). Actually it is

$$\mathscr{R}_k = \frac{1}{(k-1)!} \left.\frac{\partial^{k-1}\left[\Omega^2(q)\Phi''(\tau; q) + \Phi(\tau; q) + 100\Phi^3(\tau; q)\right]}{\partial q^{k-1}}\right|_{q=0}$$

$$= \frac{1}{(k-1)!} \left.\left[\frac{\partial^{k-1}}{\partial q^{k-1}}[\Omega^2(q)\Phi''(\tau; q)] + \frac{\partial^{k-1}}{\partial q^{k-1}}[\Phi(\tau; q)] + 100\frac{\partial^{k-1}}{\partial q^{k-1}}[\Phi^3(\tau; q)]\right]\right|_{q=0}$$

or after taking into consideration Leibnitz rule for the derivative of a product of functions

$$\mathscr{R}_k = \frac{1}{(k-1)!}\left[\frac{\partial^{k-1}[\Phi(\tau;q)]}{\partial q^{k-1}}\right.$$

$$+100\sum_{m=0}^{k-1}\sum_{\ell=0}^{m}\binom{k-1}{m}\binom{m}{\ell}\frac{\partial^{\ell}\Phi(\tau;q)}{\partial q^{\ell}}\frac{\partial^{m-\ell}\Phi(\tau;q)}{\partial q^{m-\ell}}\frac{\partial^{k-1-m}\Phi(\tau;q)}{\partial q^{k-1-m}}$$

$$+\sum_{m=0}^{k-1}\sum_{\ell=0}^{m}\binom{k-1}{m}\binom{m}{\ell}\frac{\partial^{\ell}\Omega(q)}{\partial q^{\ell}}\frac{\partial^{m-\ell}\Omega(q)}{\partial q^{m-\ell}}\left.\left.\frac{\partial^{k-1-m}\Phi''(\tau;q)}{\partial q^{k-1-m}}\right]\right|_{q=0}$$

$$= u_{k-1}(\tau)+100\sum_{m=0}^{k-1}\sum_{\ell=0}^{m}u_{\ell}(\tau)u_{m-\ell}(\tau)u_{k-1-m}(\tau)+\sum_{m=0}^{k-1}\sum_{\ell=0}^{m}\omega_{\ell}\omega_{m-\ell}(\tau)u''_{k-1-m}(\tau).$$

Choosing $H(t) \equiv 1$, (37) becomes:

$$\omega_0^2\left(u_1''+u_1\right) = \hbar\mathscr{R}_1[u_0(\tau),\tau], \tag{48}$$

$$\omega_0^2\left(u_k''+u_k\right) = \hbar\mathscr{R}_k[u_0(\tau),\ldots,u_{k-1}(\tau),\tau], \quad k=2,3,\ldots \tag{49}$$

and we choose the accompanying conditions to be

$$u_k(0)=u_k'(0)=0, \quad k=1,2,3,\ldots. \tag{50}$$

Thus,

$$u(\tau)=u_0(\tau)+\sum_{k=1}^{\infty}u_k(\tau) \tag{51}$$

is the solution of the IVP (42)–(43) and the frequency of the oscillations is given by

$$\omega=\omega_0+\sum_{k=1}^{\infty}\omega_k. \tag{52}$$

This is indeed guaranteed by the following:

Theorem 3 *As long as the series* (51) *converges, it is a solution of* (42)–(43)*, where* $u_0(\tau)$ *is given by* (45) *and* $u_k(\tau)$*,* $k = 1, 2, 3, \ldots$ *are the solutions of* (48)*,* (50) *and* (49)*,* (50)*.*

Proof First of all, (51) satisfies (43), due to (50) since

$$u(0)=u_0(0)=1 \;\text{ and }\; u'(0)=u_0'(0)=0.$$

It remains to show that (51) satisfies (42) as well. In order to prove this, suppose that (51) converges. Then

$$\lim_{m\to\infty} u_m(\tau) = 0. \tag{53}$$

Notice that

$$\hbar \sum_{k=1}^{m} \mathscr{R}_k = \sum_{k=1}^{m} \mathscr{L}_3 \left[u_k(\tau) - \chi_k u_{k-1}(\tau)\right] = \mathscr{L}_3 \left[u_1(\tau)\right] + \sum_{k=2}^{m} \mathscr{L}_3 \left[u_k(\tau) - u_{k-1}(\tau)\right]$$

$$\Rightarrow \hbar \sum_{k=1}^{m} \mathscr{R}_k = \mathscr{L}_3 \left[u_1(\tau) + \sum_{k=2}^{m} \left[u_k(\tau) - u_{k-1}(\tau)\right]\right] = \mathscr{L}_3 \left[u_m(\tau)\right]$$

$$\Rightarrow \hbar \sum_{k=1}^{\infty} \mathscr{R}_k = 0, \tag{54}$$

due to (53) and the linearity of $\mathscr{L}_3$. Moreover,

$$\sum_{k=1}^{\infty} \mathscr{R}_k = \sum_{k=1}^{\infty} u_{k-1}(\tau) + 100 \sum_{k=1}^{\infty} \sum_{m=0}^{k-1} \sum_{\ell=0}^{m} u_\ell(\tau) u_{m-\ell}(\tau) u_{k-1-m}(\tau)$$

$$+ \sum_{k=1}^{\infty} \sum_{m=0}^{k-1} \sum_{\ell=0}^{m} \omega_\ell \omega_{m-\ell}(\tau) u''_{k-1-m}(\tau)$$

$$= \sum_{n=0}^{\infty} u_n(\tau) + 100 \sum_{n=0}^{\infty} \sum_{m=0}^{n} \sum_{\ell=0}^{m} u_\ell(\tau) u_{m-\ell}(\tau) u_{n-m}(\tau)$$

$$+ \sum_{n=0}^{\infty} \sum_{m=0}^{n} \sum_{\ell=0}^{m} \omega_\ell \omega_{m-\ell}(\tau) u''_{n-m}(\tau)$$

$$\Rightarrow \sum_{k=1}^{\infty} \mathscr{R}_k = u(\tau) + 100 u^3(\tau) + \omega^2 u''(\tau). \tag{55}$$

Combining (54), (55) and taking into account that $\hbar \neq 0$, we end up with (42), which completes the proof.

By solving iteratively (48), (50) and (49), (50), we find (51) and as a consequence $x(t)$. When solving (48), (50), a term of the form $t \sin t$ appears in its solution. The appearance of such a term disobeys the rule of solution expression, since this term doesn't belong to the S_B we chose. In order for this rule to be satisfied we force the coefficient of this term to be 0. This gives us the algebraic equation:

$$\frac{38h}{\omega_0^2} - \frac{h}{2} = 0,$$

from which ω_0 is found to be

$$\omega_0 = 2\sqrt{19}.$$

The term $t \sin t$ appears also in $u_2(\tau)$. Thus, we find in the same way the linear now algebraic equation

$$\frac{1875h^2}{92416} - \frac{h\omega_1}{2\sqrt{19}} = 0$$

from which ω_1 is found to be

$$\omega_1 = \frac{1875h}{2432\sqrt{19}}.$$

Similarly we find for ω_2

$$\frac{343044375h^3}{17081434112} + \frac{1875h^2}{92416} - \frac{h\omega_2}{2\sqrt{19}} = 0 \Rightarrow \omega_2 = \frac{1875h(2927312h + 2957312)}{7192182784\sqrt{19}}, \quad \text{etc.}$$

For this simple case it suffices to calculate only $u_k(\tau)$, $k = 1, 2, 3$ to obtain satisfactory results. This is shown at Fig. 6, where the solution of (40)–(41) (with dotted line), obtained by HAM for $\hbar = -1.2$ is depicted. At the same figure, the corresponding fourth order RK numerical solution of (40)–(41) is shown with a continuous line. We can easily verify that the HAM solution is in very good agreement with the corresponding numerical one. The value of $\hbar$ was chosen after obtaining the $\hbar$-curves for $u''(0)$ and $u''''(0)$ which are shown at Fig. 7. For $\hbar = -1.2$ the frequency is $\omega = \omega_0 + \omega_1 + \omega_2 + \ldots \simeq 8.54542$, whereas the exact frequency in this case is known to be $\omega_{exact} = 8.53359$. The agreement between the exact and the approximate by HAM frequency improves as we increase k, but it may also be improved by choosing another value of $\hbar$ within the valid region. If we choose $\hbar = -1$, Fig. 6 remains the same, but the approximate by HAM frequency becomes $\omega \simeq 8.53913$ which is in better agreement with ω_{exact}.

When $\hbar \in [-1, 0)$, the series (52) converges in all $\mathbb{R}$ (see [46, p. 172]). For other cases, the series (52) and/or (51) may converge in smaller regions of $\mathbb{R}$. However, their convergence region as well as their speed of convergence can be improved by using the Padé technique. For this technique combined with HAM we refer to [46, p. 64].

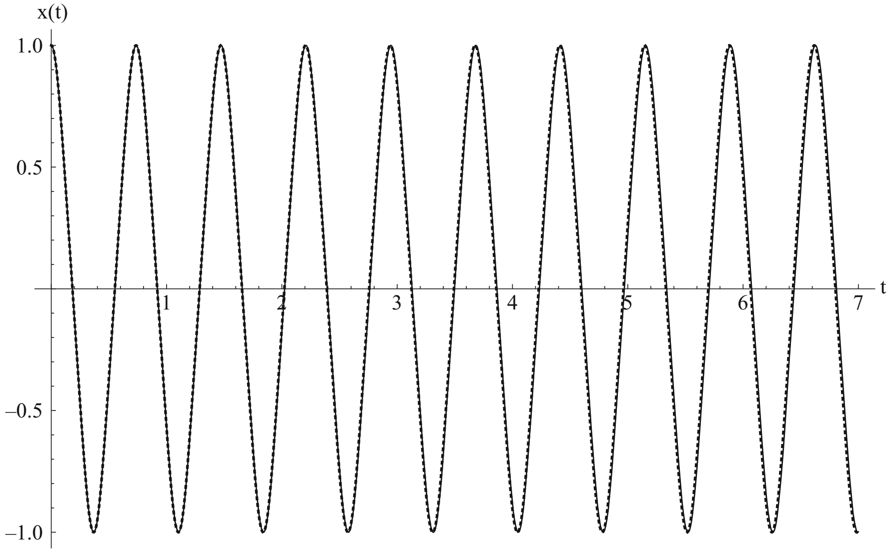


Fig. 6 Graph of $x(t)$ of (40)–(41) obtained by HAM (dotted line) and corresponding numerical solution (continuous line)

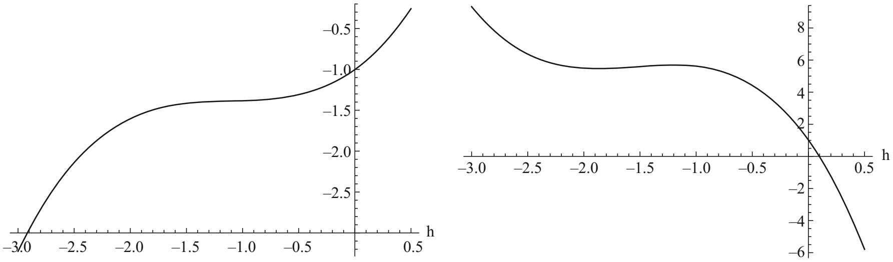


Fig. 7 $\hbar$-curves for $u''(0)$ (left) and $u''''(0)$ (right) for (42)–(43)

4 Numerical Techniques

4.1 The Runge-Kutta Method

The Runge-Kutta methods are a class of iterative methods implicit or explicit, used in temporal discretization for the approximate solutions of ODEs. These methods were developed 100 years ago, around 1900, by the German mathematicians C. Runge and M.W. Kutta. The most widely known member of the Runge-Kutta family is generally referred to as "RK4" or "classical Runge-Kutta method", and it is an explicit method with fourth order accuracy [62]. Runge-Kutta (RK) methods are one-step methods, such as the Euler method, and increase their accuracy at the price of an increase of functional evaluations at each time level. RK methods are widely utilized for the numerical solution of nonlinear ODEs, such as the one discussed in this chapter.

To specify a particular RK method, one needs to provide $q \in \mathbb{N}$, the number of stages, and the coefficients $a_{ij} \in \mathbb{R}$, $i, j = 1, \ldots, q$, $b_i \in \mathbb{R}$, $i = 1, \ldots, q$ and τ_i, $i = 1, \ldots, q$. The matrix $\mathbf{A} = (a_{ij})$ is called the RK matrix, while the b_i and τ_i are known as the weights and the nodes [41]. The RK method that provides the approximate solution for a first order ODE with initial condition, $x(0) = X_0$ is,

$$\begin{cases} k_{n,i} = f(t_{n,i}, x_n + h \displaystyle\sum_{j=1}^{q} a_{ij} k_{n,j}), \quad i - 1, \ldots, q, \\ x_{n+1} = x_n + h \displaystyle\sum_{i=1}^{q} b_i k_{n,i}, \\ x_0 = X_0, \end{cases} \tag{56}$$

where, $k_{n,i} = f(t_{n,j}, x_{n,j})$, $n = 0, 1, \ldots, N-1$ is the domain partition, and h the step size. The coefficients, a_{ij}, b_i and τ_i are usually arranged as a Butcher tableau [4, 13],

$$\begin{array}{cccc|c} a_{11} & a_{12} & \ldots & a_{1q} & \tau_1 \\ a_{21} & a_{22} & \ldots & a_{2q} & \tau_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{q1} & a_{q2} & \ldots & a_{qq} & \tau_q \\ \hline b_1 & b_2 & \ldots & b_q & \end{array}$$

It should be mentioned that an RK method can be extended to systems of ODEs. However, the order of an RK method in the scalar case does not necessarily coincide with that in the vector case [62].

The following theorem concerns the stability of RK method and the uniqueness of the obtained numerical solution.

Theorem 4 ([4]) *For a system of m–ODEs, e.g.*

$$\begin{cases} \bar{x}' = \bar{f}(t, \bar{x}), \quad t \in [a, b], \\ \bar{x}(a) = \bar{X}_0, \end{cases} \tag{57}$$

where, $\bar{x}$ and $\bar{f} \in \mathbb{R}^m$, are smooth functions, and $\bar{f}$ satisfies the Lipschitz condition,

$$\exists\, L \geqslant 0,\ \forall \bar{x}_1,\ \bar{x}_2 \in \mathbb{R}^m : \left\| \bar{f}(t, \bar{x}_1) - \bar{f}(t, \bar{x}_2) \right\|_\infty \leqslant L \left\| \bar{x}_1 - \bar{x}_2 \right\|_\infty ,$$

where $\|\bar{x}\|_\infty = \max_t \max_i |x_i(t)|$, $i = 1, \ldots, m$ *and* $t \in [a, b]$. *We assume that* $\gamma h < 1$, *where* $\gamma = L \max_i \sum_j |a_{ij}|$, *and* $h = \frac{b-a}{N}$, $N \in \mathbb{N}$ *is the partition step. Under these assumptions the discrete system has a unique solution.*

Additionally, if $\bar{y}_n$, $\bar{y}_{n,i} \in \mathbb{R}^m$ *satisfy relation* (56) *with known* $\bar{y}_0 \in \mathbb{R}^m$, *then,*

$$\max_{1\leqslant n\leqslant N} \|\bar{x}_n - \bar{y}_n\|_\infty \leqslant C \|\bar{x}_0 - \bar{y}_0\|_\infty ,$$

where C *does not depend on the partition step,* h*.*

Remark 1 It can be shown that RK methods for systems are stable for every norm, $\|.\|$ of $\mathbb{R}^m$, meaning that there exists C_2 that does not depend on h such that,

$$\max_{1\leqslant n\leqslant N} \|\bar{x}_n - \bar{y}_n\| \leqslant C_2 \|\bar{x}_0 - \bar{y}_0\| .$$

Remark 2 The RK method is of p order if,

$$\max_{0\leq n\leq N-1} \|y_{n+1} - y(t_{n+1})\|_\infty \leq C_3 h^{p+1}$$

where C_3 does not depend on h. The previous inequality gives us the local error of an RK method.

Remark 3 The RK method is consistent if and only if $\sum_{i=1}^{q} b_i = 1$. An extended mathematical analysis about RK methods can be found in [4, 41, 62]

Remark 4 In explicit RK methods, such the RK4, the matrix $\mathbf{A}$ in Butcher tableau is strictly lower triangular, as shown for the RK4,

$$\begin{array}{cccc|c}
0 & 0 & 0 & 0 & 0 \\
1/2 & 0 & 0 & 0 & 1/2 \\
0 & 1/2 & 0 & 0 & 1/2 \\
0 & 0 & 1 & 0 & 1 \\
\hline
1/6 & 1/3 & 1/3 & 1/6 &
\end{array}$$

Similarly to (10), the IVP (1)–(2) can be reduced to the following system

$$\begin{cases} x' = y = g_1(t, x, y), \\ y' = -a(x^2 - 1)y - \beta x - \gamma x + f(t) = g_2(t, x, y), \end{cases} \tag{58}$$

where $t \in [0, T]$ and $x = x(t)$, $y = y(t)$ smooth functions. The RK4 for (58), subject to the initial conditions $x(0) = 0$, $y(0) = 0$ has the form

$$x_{n+1} = x_{n+1} + K \quad and \quad y_{n+1} = y_n + L,$$

where $K = \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$, $L = \frac{1}{6}(l_1 + 2l_2 + 2l_3 + l_4)$ and

$$\begin{aligned}
&k_1 = hg_1(t_n, x_n, y_n), && l_1 = hg_2(t_n, x_n, y_n),\\
&k_2 = hg_1(t_n + 0.5h, x_n + 0.5k_1, y_n + 0.5l_1), && l_2 = hg_2(t_n + 0.5h, x_n + 0.5k_1, y_n + 0.5l_1),\\
&k_3 = hg_1(t_n + 0.5h, x_n + 0.5k_2, y_n + 0.5l_2), && l_3 = hg_2(t_n + 0.5h, x_n + 0.5k_2, y_n + 0.5l_2),\\
&k_4 = hg_1(t_n + h, x_n + k_3, y_n + l_3), && l_4 = hg_2(t_n + h, x_n + k_3, y_n + l_3)
\end{aligned} \tag{59}$$

Then (58) is solved utilizing the function ode45 in Matlab (MathWorks, Natick, MA, USA). The corresponding numerical solution for $f(t) = \cos t$, $a = 0.002$, $\beta = 0.008$ and $\gamma = 0.003$ is shown in Fig. 2.

4.2 The Finite Difference Method

The finite difference method (FDM) is the dominated approach to numerical solutions of ODEs as well as PDEs. This is due to the fact that as a method is very easy to implement in a numerical code and most times provides very reliable results in good agreement with analytical solutions. FDM was mainly developed during the period of 1940–1960 and it can be used for linear and nonlinear differential equations [4, 24]. Newer developments of the FDM is the finite volume method (FVM) usually applied in fluid mechanics applications.

In order to demonstrate FDM we'll apply it to (1)–(2), where $x \in C^2[0, T]$ for $f(t) = \delta \cos(\omega t)$. The main idea of the FDM is to replace (1) with a difference equation by partitioning the domain $[0, T]$, in N equidistant sub-domains, such that $h = T/N$ is the partition step.

The values of the unknown function, x, at the partition points t_i, with $t_i = (i - 1)h$, $i = 1, 2, \ldots, N + 1$, is $x(t_i) = x_i$. For three neighboring points of the domain $[0, T]$, e.g. t_{i-1}, t_i, t_{i+1} we can approximate the first and second order derivatives of the unknown function, x, as in Eq. (60), and both schemes have a second order truncation error.

$$\begin{cases} x_i' = \dfrac{x_{i+1} - x_{i-1}}{2h} + O(h^2), \\[2ex] x_i'' = \dfrac{x_{i+1} - 2x_i + x_{i-1}}{h^2} + O(h^2). \end{cases} \tag{60}$$

Introducing formulae (60) to (1) the difference equation

$$c_{i1}\; x_{i+1} + c_{i2}\; x_i + c_{i3}\; x_{i-1} = c_{i4}, \tag{61}$$

is obtained, where the coefficients, c_{ij}, $j = 1, \ldots, 4$ are functions of x and t, namely

$$c_{i1} = \frac{1}{h^2} + \frac{a}{2h}x_i^2 - \frac{a}{2h}, \quad c_{i2} = -\frac{2}{h^2} + \beta + \gamma x_i^2,$$
$$c_{i3} = \left(\frac{1}{h^2} - \frac{a}{2h}x_i^2 + \frac{a}{2h}\right), \quad c_{i4} = g\cos(\omega t_i).$$

The obtained algebraic system of equations that arise from (61) is nonlinear and of the form $\mathbf{A}\bar{x} = \bar{b}$, where the matrix of the coefficients $\mathbf{A} = (a_{ij})$ contains the unknown variable x and it is an $N \times N$ matrix. To numerically solve this system, nonlinear algorithms such as Newton–like methods, can be used. Newton–like methods are faster with a good initial guess but may suffer from local minima. To solve the nonlinear discretized algebraic system of equations, the Levenberg–Marquardt algorithm (LMA) is used [76]. The LMA interpolates between the Gauss–Newton algorithm and the method of Gradient Descent [62]. The LMA tends to be slightly slower than the Gauss–Newton method, but it is a more robust algorithm compared with the classical Gauss–Newton method.

The matrix $\mathbf{A}$ is either a diagonal (tridiagonal, pentadiagonal etc.) or a block–diagonal matrix. It is also symmetric and positive definite. These properties of $\mathbf{A}$ enable its inversion, which leads to the existence and uniqueness of the obtained numerical solution [24, 62].

The invertibility of $\mathbf{A}$ is connected with its condition number. Namely, at every iteration of the LMA the matrix $\mathbf{A}$ should not be "ill–conditioned". Since $\mathbf{A}$ is hermitian and positive definite, its condition number is given by the expression, $K(\lambda) = \frac{\lambda_{\max}}{\lambda_{\min}}$, where $\lambda_{\max}$ is its largest and $\lambda_{\min}$ its smallest eigenvalue. In order for $\mathbf{A}$ to be invertible, it is important its condition number to remain "small" [24, 62].

In other FDM or FVM approaches, the unknown variables are evaluated using iterative schemes [24, 25]. The iterative solution approach seems to have an advantage compared to the straightforward adoption of the direct matrix methodology used in this section. However, the coupling between the equations should be taken under consideration for achieving faster convergence of the solution. Implementing the direct approach in a Matlab script (MathWorks, Natick, MA, USA) the nonlinear system of equations obtained from (61) is solved. The corresponding numerical solution for $f(t) = \cos t$, $a = 0.002$, $\beta = 0.008$ and $\gamma = 0.003$ is shown in Fig. 2.

Instead of applying a numerical method to the IVP (1)–(2), we may apply a numerical method to (18). Comparing the standard FDM, described in this subsection, with the FDM as applied to (18) for $X_0 = X_1 = 0$, $f(t) = \cos t$, $a = 0$, $\beta = 0.008$ and $\gamma = 0.003$, we observe that the two solutions are in very good agrement, as depicted in Fig. 8. This approach seems to decrease the errors. Actually, the FDM solution of (18) gives a numerical solution that is closer (almost coincides) to the analytical solution, Eq. (4), compared with the standard FDM, as depicted in Fig. 9.

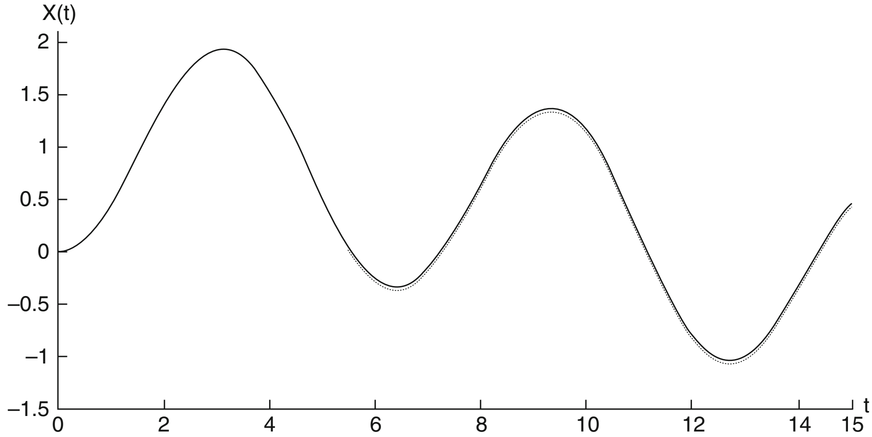


Fig. 8 Graph of the FDM solution of (1)–(2) (continuous line) with the FDM solution of (18) (dotted line) of , for $X_0 = X_1 = 0$, $f(t) = \cos t$, $a = 0$, $\beta = 0.008$ and $\gamma = 0.003$

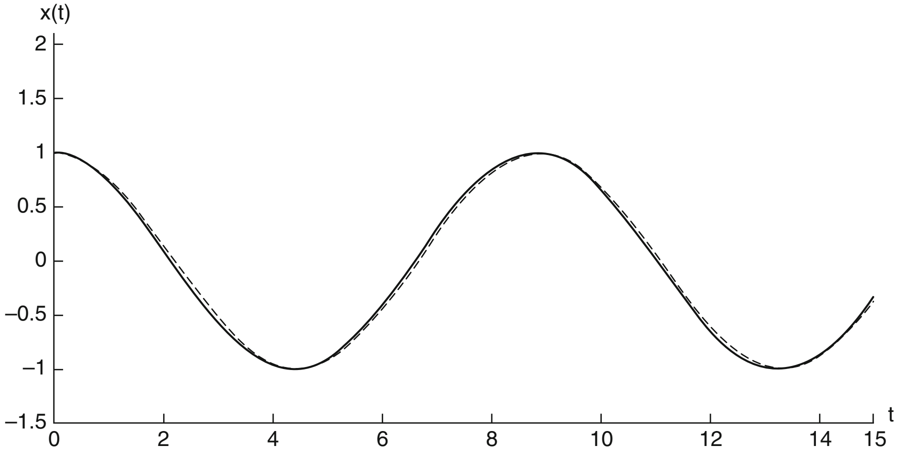


Fig. 9 Graph of the FDM solution of (1)–(2) (continuous line) with the FDM solution of (18) (dotted line) and the analytical solution (4) (dashed line), for $X_0 = 1$, $X_1 = 0$, $f(t) = 0$, $a = 0$, $\beta = 0.5$ and $\gamma = 0$

4.3 The Finite Elements Method

In the previous subsection we introduced the FDM. In this subsection we will discuss the finite element method (FEM) and specifically the Galerkin FEM. With the help of FEM we are able to construct an approximate solution of the initial value problem under consideration. These approximations are defined in the entire domain, e.g. $I = [0, T]$, they are continuous and piecewise polynomials, e.g. polynomials of order r in every subdomain of the partition of I. FEMs were developed from 1970 and beyond, they are more systematic compared to FDM and they take place in more complicated spaces [4, 62]. The Galerkin FEM is usually applied to dynamic PDEs and rarely to initial value ODE problems [5].

In order to demonstrate FEM, we'll apply it to the IVP (1)–(2), where $x \in C^2[0, T]$ in the case where $a = 0$. In this case, (1)–(2) takes the form

$$x'' + \beta x + \gamma x^3 = f(t), \tag{62}$$

$$x(0) = X_0, \quad x'(0) = X_1, \tag{63}$$

which is called the "strong form" [4, 12].

Introducing the piecewise continuously differentiable function $v \in C_0[0, T]$, we write,

$$(x'', v) + (g(x, t), v) = (f, v), \tag{64}$$

where, $g(x, t) = \beta x + \gamma x^3$ and $(\cdot, \cdot)$ is the L^2 dot product defined for piecewise continuous functions, x, v in $[0, T]$ as,

$$(x, v) = \int_0^T x(t)v(t)dt = \sum_{j=0}^{N} \int_{t_j}^{t_{j+1}} x(\tau)v(\tau)d\tau,$$

where, $0 = t_0 < t_1 < t_2 < \ldots < t_{J+1} = T$ is the domain partition. The induced norm from the dot product is expressed as $\|\cdot\|$ and the function $v \in V$, where $V = \{v \in C_0[0, T] : v$ are piecewise continuously differentiable functions$\}$. Applying integration by parts, the term (x'', v) can be written as,

$$\big(x'', v\big) = \sum_{j=0}^{N} \int_{t_j}^{t_{j+1}} x''(\tau)v(\tau)d\tau = \sum_{j=0}^{N} \left[x'(\tau)v(\tau)\Big|_{t_j}^{t_{j+1}} - \int_{t_j}^{t_{j+1}} x'(\tau)v'(\tau)d\tau \right]. \tag{65}$$

Assuming that the function $v \in C_0[0, T]$ is a linear polynomial with $v(0) = v(T) = 0$, (65) can be rewritten as,

$$\big(x'', v\big) = -(x', v'), \tag{66}$$

Finally, we can write that,

$$-(x', v') + (g, v) = (f, v), \tag{67}$$

which is called the "weak form".

We introduce linear polynomials (linear splines) to construct the simplest Galerkin FEM [4]. We impose the domain partition, $0 = t_0 < t_1 < t_2 < \ldots < t_{J+1} = T$ with $\max_j(t_{j+1}, t_j) = h$ and the space $S_h^2 =$

$\{\phi \in C[0, T] : \phi[t_j, t_{j+1}] \in \mathbb{P}_1\}$. The functions ϕ_j, $j = 1, 2, \ldots, J$ are defined in $[0, T]$ as,

$$\phi_j(t) = \begin{cases} \dfrac{t - t_{j-1}}{t_j - t_{j-1}}, & t_{j-1} \leqslant t \leqslant t_j \\ \dfrac{t_{j+1} - t}{t_{j+1} - t_j}, & t_{j-1} \leqslant t \leqslant t_j \\ 0, & \text{elsewhere} \end{cases} \tag{68}$$

and for $i = 0$ is defined as,

$$\phi_0(t) = \begin{cases} \dfrac{t_1 - t}{t_1 - t_0}, & t_0 \leqslant t \leqslant t_1. \\ 0, & \text{elsewhere} \end{cases} \tag{69}$$

The set $\{\phi_j\}_{j=0}^{J}$ is a basis of S_h^2 and is called nodal basis and the discrete points t_j are called nodes [12]. We assume that the approximate solution, $x_h \in S_h^2$, satisfy the "weak form" Eq. (67) in space S_h^2. For the linear case it has been shown that: If the weak form is $a(x, v) = (f, v)$, where $a(\cdot, \cdot)$ is the bilinear form of a linear operator A, and $x \in C^2[0, T]$ and $f \in C^0[0, T]$ satisfy the weak form, then x also satisfies the strong form with the appropriate initial conditions. Moreover, if $f \in L^2[0, T]$, equation $A(x_h, v) = (f_h, v)$, $\forall v \in S_h^2$, has a unique solution. More details on the linear problem can be found in [12].

Writing the approximate solution in the form, $x_h = a_0\phi_0 + a_1\phi_1 + \ldots + a_J\phi_J$ we obtain the algebraic system, $\mathbf{A}\,\bar{a} = \bar{F}$, where $\bar{a} = (a_0, a_1, \ldots, a_J)^T$, $\bar{F} = (F_i)$, $F_i = (f, \phi_i)$, $i = 0, 1, \ldots, J$ and $\mathbf{A} = (A_{ij})$, with $A_{ij} = -(\phi_j', \phi_i') + (\beta\phi_j, \phi_i) + (\gamma\phi_j^3, \phi_i)$, $i, j = 0, 1, \ldots, J$.

The system of equations obtained from the "weak form" (67) is numerically solved implementing FEM in Matlab. The corresponding numerical solution for $X_0 = X_1 = 0$, $f(t) = \frac{1}{2}\cos t$, $a = 0$, $\beta = 0.035$ and $\gamma = 0.0001$ is shown (dashed line) in Fig. 10. In the same figure, the numerical solutions obtained by RK4 (continuous line), FDM (dotted line) are also depicted. The solution obtained by FAT (see next section) is not depicted since for theses values of parameters and initial conditions it coincides graphically with the RK4 solution. However, in Table 4 we compare the three numerical solutions obtained by FEM, FDM and RK4 with the solution obtained by HAM and the solution obtained by FAT that follows.

5 A Functional-Analytic Technique

In this final section, we'll describe a functional analytic technique (FAT) for obtaining the solution of ODEs. Actually, this method is the combination of two

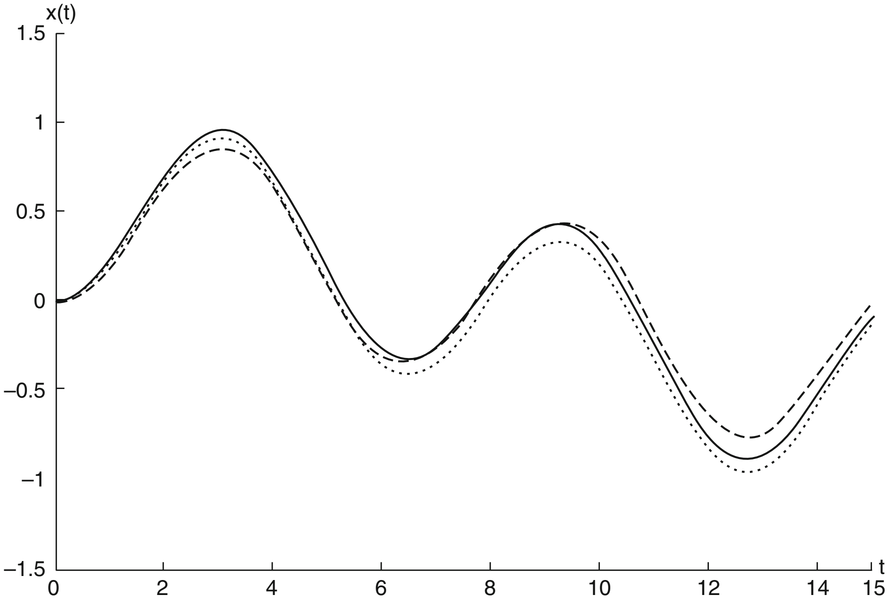


Fig. 10 Graph of the three numerical methods, RK4 (continuous line), FDM (dotted line) and FEM (dashed line) of (1)–(2), for $X_0 = X_1 = 0$, $f(t) = \frac{1}{2}\cos t$, $a = 0$, $\beta = 0.035$ and $\gamma = 0.0001$

Table 4 Comparison of the three numerical methods, RK4, FEM, FDM, with HAM and FAT for various values of time, t and for $X_0 = 1$, $X_1 = 0$, $f(t) = 0$, $a = 0$, $\beta = 1.0$ and $\gamma = 10.0$

Time (s)	HAM	FAT	RK4	FDM	FEM
0	1.0000000000	1.0000000000	1.0000000000	1.0000000000	1.0000000000
2	0.8410511400	0.8356150796	0.8352800715	0.8176114534	0.8199895952
4	0.4318527148	0.4283707319	0.4883298390	0.4239989518	0.4820233720
6	−0.0523297550	−0.0649773903	−0.0639698268	−0.0704801933	−0.0850110357
8	−0.5094371206	−0.5177897201	−0.5166629311	−0.5121772708	−0.5171721184
10	−0.8873520147	−0.8847765340	−0.8925938207	−0.8788896663	−0.8425358431
12	−0.9898718736	−0.9898506062	−0.9847090261	−0.9863830224	−0.9713292225
14	−0.7368877482	−0.7333500846	−0.7209509149	−0.7360511664	−0.7717958122
16	−0.3047155397	−0.3009010000	−0.3675830852	−0.3201920195	−0.3666390525
18	0.1569462253	0.1727534941	0.1612739037	0.1271600464	0.2110826426
20	0.6075260295	0.6265640712	0.6463317609	0.6106077360	0.6218574555

FATs originally developed by Ifantis in [35–38] and [39], for obtaining existence and uniqueness results of solutions of IVPs for ODEs and ordinary difference equations in the Hilbert spaces

$$H_2(\mathbb{D}) = \left\{ f : \mathbb{D} \to \mathbb{C}, \quad f(z) = \sum_{n=1}^{\infty} f_n z^{n-1} \ \text{ with } \ \sum_{n=1}^{\infty} |f_n|^2 < +\infty \right\}, \tag{70}$$

where $\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$ and

$$\ell_2 = \left\{ f_n : \mathbb{N} \to \mathbb{C} \ \text{ with } \ \sum_{n=1}^{\infty} |f_n|^2 < +\infty \right\} \tag{71}$$

or the Banach spaces

$$H_1(\mathbb{D}) = \left\{ f : \mathbb{D} \to \mathbb{C}, \quad f(z) = \sum_{n=1}^{\infty} f_n z^{n-1} \ \text{ with } \ \sum_{n=1}^{\infty} |f_n| < +\infty \right\} \tag{72}$$

and

$$\ell_1 = \left\{ f_n : \mathbb{N} \to \mathbb{C} \ \text{ with } \ \sum_{n=1}^{\infty} |f_n| < +\infty \right\}. \tag{73}$$

Later on, these FATs were also used and extended, not only by Ifantis, but also by his students and collaborators. More recently, in [59], they were combined in order to develop a “discretization” technique for the solution of IVPs of nonlinear ODEs in the real plane, such as the Duffing equation. In [60], this “discretization” technique was extended to BVPs and was applied to the well-known Blasius problem. Moreover, it was extended in order to compute complex solutions of this problem. The same “discretization” technique was also used in [58], in order to study the behavior of the famous logistic equation in the complex plane.

The basic idea of this FAT is to “discretize” an ODE by converting it into an equivalent difference equation, utilizing specific mappings among $H_2(\mathbb{D})$, ℓ_2 and an abstract Hilbert space H, or among $H_1(\mathbb{D})$, ℓ_1 and an abstract Banach space H_1. The obtained solution of the ODE under consideration, by use of this technique, is an analytic solution of the form

$$x(t) = \sum_{n=1}^{\infty} x_n \left(\frac{t}{T} \right)^{n-1}, \quad |t| < T, \quad T > 0, \quad t \in \mathbb{R} \ \text{ or } \ t \in \mathbb{C}, \tag{74}$$

where the coefficients x_n are uniquely determined by a specific difference equation which is called the discrete equivalent of the ODE under consideration.

In order to proceed, we'll start by first defining H, H_1 and the mappings among H, H_1, $H_2(\mathbb{D})$, $H_1(\mathbb{D})$, ℓ_2 and ℓ_1, on which this FAT is based. Let's denote by H an abstract separable Hilbert space over the complex field, with the orthonormal base $\{e_n\}$, $n = 1, 2, 3, \ldots$ and by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ the inner product and the norm in H, respectively. Define also in H the shift operator V:

$$V e_n = e_{n+1}, \quad n = 1, 2, 3, \ldots$$

and its adjoint V^*:

$$V^*e_n = e_{n-1}, \quad n = 2, 3, \ldots, \quad V^*e_1 = 0,$$

as well as the diagonal operator C_0:

$$C_0 e_n = n e_n, \quad n = 1, 2, 3, \ldots .$$

The following two propositions hold.

Proposition 2 ([37]) *The representation*

$$\langle f_z, f \rangle = \sum_{n=1}^{\infty} f_n z^{n-1} = f(z), \quad z \in \mathbb{D}, \tag{75}$$

is a one-by-one mapping from H onto $H_2(\mathbb{D})$ which preserves the norm, where $f_z = \sum_{n=1}^{\infty} z^{n-1} e_n$, $f_0 = e_1$, is the complete system in H of eigenvectors of V^* and $f = \sum_{n=1}^{\infty} \overline{f}_n e_n$ an element of H.

The unique element $f = \sum_{n=1}^{\infty} \overline{f_n} e_n$ appearing in (75) is called the *abstract form* of $f(z)$ in H. In general, if $G(f(z))$ is a function from $H_2(\mathbb{D})$ to $H_2(\mathbb{D})$ and $N(f)$ is the unique element in H for which

$$G(f(z)) = \langle f_z, N(f) \rangle ,$$

then $N(f)$ is called the *abstract form* of $G(f(z))$ in H.

The corresponding by the representation (75), abstract Banach space of the elements $f = \sum_{n=1}^{\infty} \overline{f}_n e_n \in H$ for which $\sum_{n=1}^{\infty} |f_n| < +\infty$, will be denoted by H_1 and the norm in H_1 by $\| \cdot \|_1$. For H_1 it is known [38, pp. 348–349] that it is invariant under the operators V^k, $(V^*)^k$, as well as under every bounded diagonal operator.

Proposition 3 ([39]) *The linear function*

$$\phi : H(H_1) \to \ell_2(\ell_1)$$

defined by

$$\phi(f) = \langle f, e_n \rangle = f_n \tag{76}$$

is an isomorphism from H (H_1) onto ℓ_2 (ℓ_1), i.e. it is a $1-1$ *mapping from H (H_1) onto ℓ_2 (ℓ_1) which preserves the norm.*

The application of this FAT can be summarized in the following steps:

Step 1 Reduce the problem under consideration with unknown function $x(t)$ with $|t| < T, T > 0$, to a new one with independent variable z with $|z| < 1$. This is done by using the simple transformation:

$$z = \frac{t}{T}, \quad x(t) = x(zT) = \tilde{x}(z). \tag{77}$$

Step 2 Use the mapping (75) and rewrite the problem for $\tilde{x}(z)$ in the form of an inner product like

$$\langle f_z, Ef \rangle = 0, \tag{78}$$

where E is an operator defined on H or H_1.

Step 3 Use the completeness of f_z, $|z| < 1$ and find from (78) the equivalent to the problem for $\tilde{x}(z)$, operator equation in H or H_1, i.e.

$$Ef = 0, \tag{79}$$

Step 4 Take the inner product of both parts of (79) with e_n and use (76) to obtain the equivalent to (79) difference equation, with unknown the sequence x_n appearing in (74).

Step 5 Compute x_n and find

$$\tilde{x}(z) = \sum_{n=1}^{\infty} x_n z^{n-1} \Leftrightarrow x(t) = \sum_{n=1}^{\infty} x_n \left(\frac{t}{T}\right)^{n-1}. \tag{80}$$

In this way a "numerical scheme" is found for the ODE under consideration. From this "numerical scheme", the coefficients x_n are found and thus the truncated solution

$$x_{tr}(t) = \sum_{n=1}^{N} x_n \left(\frac{t}{T}\right)^{n-1}, \quad |t| < T, \tag{81}$$

can be obtained, where N finite number.

If we're interested in complex solutions of the ODE under consideration, we should further proceed by writing $x_n = u_n + iv_n$ and $t = re^{i\theta}$, so as to obtain the recurrence relations satisfied by u_n and v_n, as well as the Re$[x(t)]$ and Im$[x(t)]$. There are many reasons for studying complex solutions of ODEs, even for ODEs arising in real physical problems, such as (1). One of them is of course mathematical curiosity. Another one, as already mentioned in Sect. 2.1, is the fact that singularity analysis for the complex solutions of the ODE, can provide evidence for its integrability or not. In addition, the investigation of physical problems for complex values of physical parameters can in many cases provide a further insight of the properties of their solutions. Actually, complex solutions of (1)–(2), for $\beta = 1$, $f(t) = \delta\cos(\omega t)$, $\alpha, \gamma, \delta, \omega \in \mathbb{R}$, were investigated in [10]. Moreover, the chaos control of chaotic unstable limit cycles of real and complex nonlinear van der Pol oscillators was investigated in [51].

In order to demonstrate this FAT, we'll apply it to the IVP (1)–(2), for $|t| < T$, $T > 0$. Thus, Steps 1–6, become the following Steps A1–A6.

Step A1 Using (77), the IVP (1)–(2) becomes

$$\tilde{x}'' - aT\tilde{x}' + aT\tilde{x}^2\tilde{x}' + \beta T^2\tilde{x} + \gamma T^2\tilde{x}^3 = T^2\tilde{f}(z) \tag{82}$$

$$\tilde{x}(0) = X_0, \quad \tilde{x}'(0) = TX_1, \tag{83}$$

where $|z| < 1$, $\tilde{x} = \tilde{x}(z)$ and $\tilde{f}(z) = f(zT) = f(t) = \sum_{n=1}^{\infty} \tilde{f}_n z^{n-1}$.

Step A2 In order to rewrite (82)–(83) in the form of an inner product in H_1, we need the abstract forms of the terms appearing in (82). These were already found in [37] and [38] and we give them in the form of the next proposition.

Proposition 4 *The following relations hold:*

(i) $\dfrac{d^n g(z)}{dz^n} = \left\langle f_z, (C_0V^*)^n g\right\rangle$,

(ii) $\phi(z)g(z) = \langle f_z, \phi(V)g\rangle$,

(iii) $[h(z)]^n = \left\langle f_z, [h_1(V)]^{n-1}h\right\rangle$,

where $n = 1, 2, \ldots$, $g(z) = \displaystyle\sum_{n=1}^{\infty} \overline{g}_n z^{n-1} \in H_2(\mathbb{D})$, $h(z) = \displaystyle\sum_{n=1}^{\infty} \overline{h}_n z^{n-1} \in H_2(\mathbb{D})$, $\phi(z) = \displaystyle\sum_{n=1}^{\infty} c_n z^{n-1}$ *analytic in some neighborhood of* $\overline{\mathbb{D}} = [-1, 1]$, $\phi(V) = \displaystyle\sum_{n=1}^{\infty} \overline{c}_n V^{n-1}$, $h_1(V) = \displaystyle\sum_{n=1}^{\infty} \overline{h}_n V^{n-1}$, $g = \displaystyle\sum_{n=1}^{\infty} \overline{g}_n e_n \in H$ *and* $h = \displaystyle\sum_{n=1}^{\infty} \overline{h}_n e_n \in H_1$.

Combining (i) and (iii) of the previous proposition, we easily find that

$$g^2(z)g'(z) = \frac{d}{dz}\left[\frac{1}{3}g^3(z)\right] = \left\langle f_z, \frac{1}{3}C_0V^*[g_1(V)]^2 g\right\rangle.$$

Thus, (82) is rewritten in the form

$$< f_z, (C_0V^*)^2\tilde{x} - \overline{a}TC_0V^*\tilde{x} + \frac{\overline{a}T}{3}C_0V^*[\tilde{x}_1(V)]^2\tilde{x} + \overline{\beta}T^2\tilde{x} + \overline{\gamma}T^2[\tilde{x}_1(V)]^2\tilde{x} - T^2\tilde{f} >= 0, \tag{84}$$

where $\tilde{f}$ and $\tilde{x}$ are the abstract forms in H_1 of $\tilde{f}(z)$ and $\tilde{x}(z)$, respectively.

Step A3 Using the completeness of f_z, $|z| < 1$ the equivalent to (82), operator equation in H_1 is

$$(C_0V^*)^2\tilde{x} - \overline{a}TC_0V^*\tilde{x} + \overline{\beta}T^2\tilde{x} = T^2\tilde{f} - \frac{\overline{a}T}{3}C_0V^*[\tilde{x}_1(V)]^2\tilde{x} - \overline{\gamma}T^2[\tilde{x}_1(V)]^2\tilde{x}. \tag{85}$$

Step A4 Taking the inner product of both parts of (85) with e_n and using (76) we eventually obtain

$$\begin{gathered} x_{n+2} = \frac{T^2}{n(n+1)}\tilde{f}_n + \frac{aT}{n+1}x_{n+1} - \frac{\beta T^2}{n(n+1)}x_n - \\ \frac{\gamma T^2}{n(n+1)}\sum_{k=1}^{n} x_k \sum_{s=1}^{n-k+1} x_s x_{n-k-s+2} - \frac{aT}{3(n+1)}\sum_{k=1}^{n+1} x_k \sum_{s=1}^{n-k+2} x_s x_{n-k-s+3}. \end{gathered} \tag{86}$$

Moreover, from the initial conditions (83), we find:

$$\tilde{x}(0) = X_0 \Rightarrow \sum_{n=1}^{\infty} x_n z^{n-1}\Big|_{z=0} = X_0 \Rightarrow x_1 = X_0. \tag{87}$$

$$\tilde{x}'(0) = TX_1 \Rightarrow \sum_{n=2}^{\infty}(n-1)x_n z^{n-2}\Big|_{z=0} = TX_1 \Rightarrow x_2 = TX_1. \tag{88}$$

Step A5 Computing x_n from (86)–(88) we find the solution (80) of (1)–(2) for $t \in \mathbb{R}$.

Step A6 Setting $x_n = u_n + iv_n$, $a = a_1 + ia_2$, $\beta = \beta_1 + i\beta_2$, $\gamma = \gamma_1 + i\gamma_2$, $t = re^{i\theta}$, (86)–(88) become:

$$\begin{aligned} u_{n+2} = {} & \frac{T^2\mathrm{Re}[\tilde{f}_n]}{n(n+1)} - \frac{T^2\beta_1 u_n}{n(n+1)} + \frac{T^2\beta_2 v_n}{n(n+1)} + \frac{Ta_1u_{n+1}}{n+1} - \frac{Ta_2v_{n+1}}{n+1} - \\ & \frac{T^2\gamma_1}{n(n+1)}\sum_{k=1}^{n}\sum_{s=1}^{-k+n+1}(-v_ku_sv_{-k+n-s+2} - v_kv_su_{-k+n-s+2} - u_kv_sv_{-k+n-s+2} + u_ku_su_{-k+n-s+2}) - \\ & \frac{T^2\gamma_2}{n(n+1)}\sum_{k=1}^{n}\sum_{s=1}^{-k+n+1}(-v_ku_s - u_{-k+n-s+2} - u_ku_sv_{-k+n-s+2} - u_kv_su_{-k+n-s+2} + v_kv_sv_{-k+n-s+2}) - \\ & \frac{Ta_1}{3(n+1)}\sum_{k=1}^{n+1}\sum_{s=1}^{-k+n+2}(-v_ku_sv_{-k+n-s+3} - v_kv_su_{-k+n-s+3} - u_kv_sv_{-k+n-s+3} + u_ku_su_{-k+n-s+3}) - \\ & \frac{Ta_2}{3(n+1)}\sum_{k=1}^{n+1}\sum_{s=1}^{-k+n+2}(-v_ku_su_{-k+n-s+3} - u_ku_sv_{-k+n-s+3} - u_kv_su_{-k+n-s+3} + v_kv_sv_{-k+n-s+3}) \end{aligned}$$

$$v_{n+2}=\frac{T^2\mathrm{Im}[\tilde{f}_n]}{n(n+1)}-\frac{T^2\beta_2u_n}{n(n+1)}-\frac{T^2\beta_1v_n}{n(n+1)}+\frac{Ta_2u_{n+1}}{n+1}+\frac{Ta_1v_{n+1}}{n+1}-$$

$$\frac{T^2\gamma_2}{n(n+1)}\sum_{k=1}^{n}\sum_{s=1}^{-k+n+1}(-v_ku_sv_{-k+n-s+2}-v_kv_su_{-k+n-s+2}-u_kv_sv_{-k+n-s+2}+u_ku_su_{-k+n-s+2})-$$

$$\frac{T^2\gamma_1}{n(n+1)}\sum_{k=1}^{n}\sum_{s=1}^{-k+n+1}(v_ku_su_{-k+n-s+2}+u_ku_sv_{-k+n-s+2}+u_kv_su_{-k+n-s+2}-v_kv_sv_{-k+n-s+2})-$$

$$\frac{Ta_2}{3(n+1)}\sum_{k=1}^{n+1}\sum_{s=1}^{-k+n+2}(-v_ku_sv_{-k+n-s+3}-v_kv_su_{-k+n-s+3}-u_kv_sv_{-k+n-s+3}+u_ku_su_{-k+n-s+3})-$$

$$\frac{Ta_1}{3(n+1)}\sum_{k=1}^{n+1}\sum_{s=1}^{-k+n+2}(v_ku_su_{-k+n-s+3}+u_ku_sv_{-k+n-s+3}+u_kv_su_{-k+n-s+3}-v_kv_sv_{-k+n-s+3})$$

with

$$u_1=\mathrm{Re}[X_0],\quad u_2=T\mathrm{Re}[X_1],\quad v_1=\mathrm{Im}[X_0],\quad v_2=T\mathrm{Im}[X_1].$$

Thus, the truncated solution (80) becomes

$$x_{tr}(t)=\sum_{n=1}^{N}\frac{(u_n+iv_n)}{T^{n-1}}\left(re^{i\theta}\right)^{n-1}$$

$$=\sum_{n=1}^{N}\frac{r^{n-1}}{T^{n-1}}(u_n+iv_n)\,[\cos(n-1)\theta+i\sin((n-1)\theta)]$$

$$\Rightarrow\begin{cases}\mathrm{Re}[x(t)]=\displaystyle\sum_{n=1}^{N}\frac{r^{n-1}}{T^{n-1}}\,[u_n\cos((n-1)\theta)-v_n\sin((n-1)\theta)]\\ \mathrm{Im}[x(t)]=\displaystyle\sum_{n=1}^{N}\frac{r^{n-1}}{T^{n-1}}\,[u_n\sin((n-1)\theta)+v_n\cos((n-1)\theta)]\end{cases}\tag{89}$$

Having described this FAT, some questions arise such as:

Q1: How do we know that the operator equation (85) has a solution in H? If it has such a solution, is it unique?

Q2: How do we choose N appearing in the truncated solution (81) or (89)?

Q3: How large T can be?

Q4: Regardless of the value of T, it may seem that a limitation of this FAT is the upper bound of T. Is there a way to estimate solutions formed for values of t larger than T?

The answers to Q1 and Q3 are provided by the following theorem

Theorem 5 *Suppose* $\tilde{f}(z)\in H_1(\mathbb{D})$,

$$|a|T+|\beta|T^2<2\tag{90}$$

and

$$|X_0| + T|X_1| + \frac{T^2}{2}\|\tilde{f}(z)\|_{H_1(\mathbb{D})} < \frac{(2-|a|T-|\beta|T^2)^{3/2}}{3\sqrt{|a|T+3|\gamma|T^2}}. \tag{91}$$

Then, the operator equation (85) *has a unique solution in* H_1 *bounded by* $R_0 = \sqrt{\dfrac{2-|a|T-|\beta|T^2}{|a|T+3|\gamma|T^2}}$. *Equivalently, the IVP* (82)–(83) *has a unique solution in* $H_1(\mathbb{D})$.

Proof The operator equation (85) can also be written as

$$C_0(C_0+I)(V^*)^2\tilde{x} - \overline{a}TC_0V^*\tilde{x} + \overline{\beta}T^2\tilde{x} = T^2\tilde{f} - \frac{\overline{a}T}{3}C_0V^*[\tilde{x}_1(V)]^2\tilde{x} - \overline{\gamma}T^2[\tilde{x}_1(V)]^2\tilde{x},$$

since $(C_0V^*)^2 = C_0(C_0+I)(V^*)^2$ (see [37, p. 91]) or

$$(V^*)^2\tilde{x} - \overline{a}TB_1V^*\tilde{x} + \overline{\beta}T^2B\tilde{x} = T^2B\tilde{f} - \frac{\overline{a}T}{3}B_1V^*[\tilde{x}_1(V)]^2\tilde{x} - \overline{\gamma}T^2B[\tilde{x}_1(V)]^2\tilde{x}, \tag{92}$$

where B, B_1 are the diagonal operators:

$$Be_n = \frac{1}{n(n+1)}e_n, \quad B_1e_n = \frac{1}{n+1}e_n, \quad , n = 1, 2, \ldots$$

with norms $\|B\|_1 = \|B_1\|_1 = \frac{1}{2}$.

Taking into consideration that $V^*e_1 = 0$, Eq. (92) becomes:

$$\begin{aligned}\tilde{x} - \overline{a}TV^2B_1V^*\tilde{x} + \overline{\beta}T^2V^2B\tilde{x} = T^2V^2B\tilde{f} - \tfrac{\overline{a}T}{3}V^2B_1V^*[\tilde{x}_1(V)]^2\tilde{x} \\ -\overline{\gamma}T^2V^2B[\tilde{x}_1(V)]^2\tilde{x} + c_1e_1 + c_2e_2,\end{aligned} \tag{93}$$

where c_1, c_2 are arbitrary constants which are determined by taking the inner product of both parts of (93) with e_1 and e_2. Thus, (93) is eventually rewritten as

$$\begin{aligned}\left(I - \overline{a}TV^2B_1V^* + \overline{\beta}T^2V^2B\right)\tilde{x} = \overline{X}_0e_1 + T\overline{X}_1e_2 + T^2V^2B\tilde{f} - \\ \tfrac{\overline{a}T}{3}V^2B_1V^*[\tilde{x}_1(V)]^2\tilde{x} - \overline{\gamma}T^2V^2B[\tilde{x}_1(V)]^2\tilde{x}.\end{aligned} \tag{94}$$

Due to (90), the linear operator $I - K$, where $K = \overline{a}TV^2B_1V^* - \overline{\beta}T^2V^2B$ is invertible, since $\|K\| < 1$, and its inverse is defined on all H and bounded by $\frac{1}{1-\|K\|}$ (see for example [27, pp. 70–71]). Thus, (94) takes the following form

$$\begin{aligned}&\tilde{x} = (I-K)^{-1} \\ &\quad\left[\overline{X}_0e_1 + T\overline{X}_1e_2 + T^2V^2B\tilde{f} - \frac{\overline{a}T}{3}V^2B_1V^*[\tilde{x}_1(V)]^2\tilde{x} - \overline{\gamma}T^2V^2B[\tilde{x}_1(V)]^2\tilde{x}\right] = \phi(\tilde{x}),\end{aligned} \tag{95}$$

which is convenient for the application of fixed point theorems. We usually use the fixed point theorem of Earle and Hamilton [21]:

Theorem 6 *If $p : X \to X$ is holomorphic, i.e. its Fréchet derivative exists, and $p(X)$ lies strictly inside X, then p has a unique fixed point in X, where X is a bounded, connected and open subset of a Banach space Y. (By saying that a subset X' of X lies strictly inside X is meant that there exists an $\epsilon > 0$ such that $\|x' - y\| > \epsilon$ for all $x' \in X'$ and $y \in Y - X$.)*

In order to apply the previous theorem to (95), we begin by assuming that $\|x\|_1 \le R$, R sufficiently large but finite. Then (95) gives

$$\|\phi(\tilde{x})\|_1 \le \frac{2}{2-|a|T-|\beta|T^2}\left(|X_0|+T|X_1|+\frac{T^2}{2}\|\tilde{f}\|_1\right)+\frac{|a|T+3|\gamma|T^2}{3\left(2-|a|T-|\beta|T^2\right)}R^3, \tag{96}$$

since $\|\tilde{x}_1(V)\|_1 = \|\tilde{x}\|_1$ (see [38, p. 349]). Let

$$P(R) = R - \frac{|a|T+3|\gamma|T^2}{3\left(2-|a|T-|\beta|T^2\right)}R^3,$$

which has the maximum $P(R_0) = \dfrac{2}{3}R_0$ at $R_0 = \sqrt{\dfrac{2-|a|T-|\beta|T^2}{|a|T+3|\gamma|T^2}}$. Then, if

$$\frac{2}{2-|a|T-|\beta|T^2}\left(|X_0|+T|X_1|+\frac{T^2}{2}\|\tilde{f}\|_H\right) \le P(R_0)-\epsilon,$$

where $\epsilon > 0$ arbitrary we have

$$\|\phi(\tilde{x})\|_1 \le R_0 - \epsilon < R_0$$

and ϕ is a holomorphic map (since $[\tilde{x}_1(V)]^2$ is Frechét differentiable, see [38, p. 355]) from $S(0, R_0) = \{\tilde{x} \in H_1 : \|\tilde{x}\|_1 < R_0\}$ strictly inside $S(0, R_0)$. Thus if (91) holds, the fixed point theorem 6 can be applied to (95). As a consequence the IVP (82)–(83) has a unique solution in $H_1(\mathbb{D})$, which proves Theorem 5.

Remark 5 Notice that the initial conditions accompanying (82), are incorporated in operator equation (94) or (95).

Remark 6 Since (1) is very similar to the Duffing equation (the extra term appearing in (1) is x^2x'), Theorem 5 is very similar to theorem 3.1 of [59].

Now let's return to questions Q1–Q4. As already mentioned, the existence and uniqueness of (85) is guaranteed by Theorem 5, i.e. T should be chosen so that inequalities (90)–(91) are valid.

In order to determine N, the fact that the coefficients x_n form a sequence belonging in ℓ_1 must be taken into consideration. Thus, due to the definition of ℓ_1, the coefficients x_n tend to zero as $n \to \infty$ and consequently the same holds for u_n (or v_n). Practically, this means that after some $n = m$, the coefficients x_n (or u_n, v_n) computed by the method in Step 5 (or 6) will be very small, practically zero (within the round-off error of the computer). Thus, N can be chosen greater or equal to m. Typical values of m are 30, 20 or even 10, depending on the desired accuracy of the computed solution.

Finally, in order to proceed to the numerical implementation of the method and the computation of the solution, the procedure followed in [58–60] is considered. The first step is to determine T as already mentioned and then, determine the coefficients x_n. One way to determine N for the numerical procedure is to monitor the coefficients x_n during their calculation. It can be set for example, that if five successive coefficients are below 10^{-20} then N is sufficiently large. Once the solution is calculated for the corresponding T, up to $t = t_1$ the quantity $x(t_1)$ is known. Considering now $x(t_1)$ known, it is used as a new initial value. A new IVP is again solved, after determining T and N as already described. In this way, successive IVPs are solved considering as initial values of the next problem the last calculated value of the previous one.

We calculated several solutions of (1)–(2) using this FAT, not only for $t \in \mathbb{R}$, but also for $t \in \mathbb{C}$. Indicatively we mention that in the case where $X_0 = X_1 = 0$, $f(t) = \cos t$, $a = 0.002$, $\beta = 0.008$ and $\gamma = 0.003$, the FAT solution coincides graphically with the RK4 solution depicted in Fig. 2. In the case where $X_0 = 1$, $X_1 = 0$, $f(t) = 0$, $a = 0$, $\beta = 1$ and $\gamma = 10$ some values of $x(t)$ are given in Table 4 and compared not only with the corresponding three numerical solutions obtained by FEM, FDM and RK4, but also with the solution obtained by HAM. It is obvious that the FAT solution is very close to all other solutions.

Now let's turn our attention to the complex solutions of (1)–(2), which can be calculated using (89). We performed several numerical experiments, but here we'll present the results only for the case where

$$X_0 = 1.88 + 0.5i, \quad X_1 = 0.5, \quad f(t) = 0, \quad a = -0.08, \quad \beta = 1, \quad \gamma = 0. \tag{97}$$

The reason is that the complex solutions for real t of the similar IVP

$$x'' + a\,(x\overline{x} - 1)\,x' + x = 0, \quad x(0) = X_0, \quad x'(0) = X_1. \tag{98}$$

where investigated in [51] for various values of the appearing parameters including (97). Notice that the aforementioned ODE is the complex van der Pol equation. It is remarkable, that the results for (1)–(2) are very similar to the corresponding results for (98).

In an effort to locate possible singularities in the complex plane, (1)–(2) is solved using the values (97), for various values of the angle θ all over the complex plane ($\theta \in [0°, 360°]$). It is remarkable, that the use of a complex variable permits the integration from a single point of the complex plane towards different directions

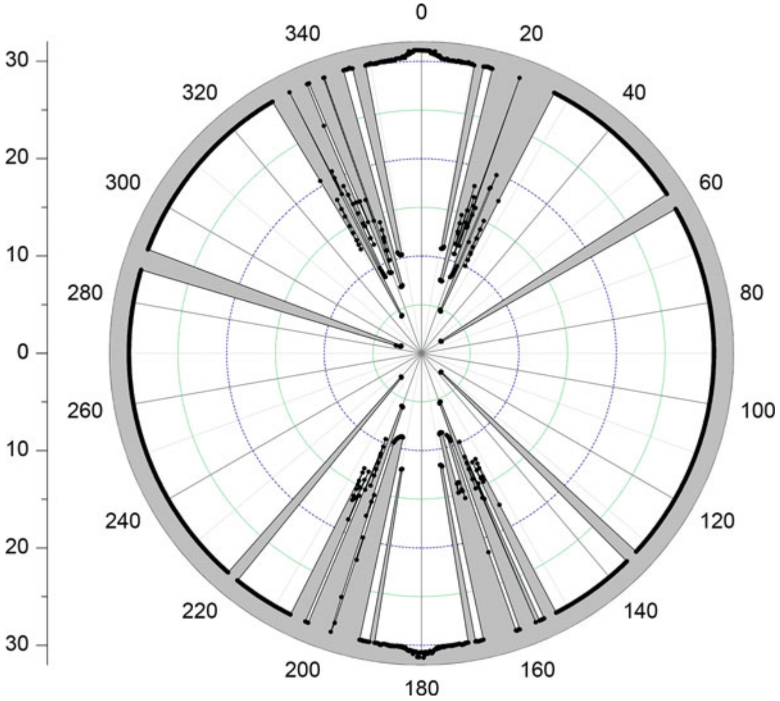


Fig. 11 Location of singularities in the complex plane of the solution of (1)–(2) for values of the parameters as given in (97)

defined by the angle θ. After performing the integration for one angle, a new integration with the same starting point takes place for θ increased by $d\theta$, which in our case was taken to be 0.2°. Obviously, for $\theta = 0°$ and $\theta = 180°$, solutions of (1)–(2) for real t are obtained. Figure 11 pictures these integrations in the polar plane and shows the location of singularities (gray sectors) of $x(t)$. These sectors are defined from the first point, with respect to (r, θ), for which the solution "blows up". These points are singularities and the solution there ceases to be analytic. Unfortunately, only the first singularity can be detected in each sector and there is no ability by using this FAT, to examine the regions for larger r for which the first singularity is detected. This means that this FAT cannot give any information on what happens within the grey regions of Fig. 11. Also, this method cannot give any information regarding the kind of the detected singularity.

Figures 12 and 13 show variations of the real and imaginary part of the solution $x(t)$ of (1)–(2) for the values (97) and for two values of the angle: $\theta = 0°$ and $\theta = 5°$. It is obvious that both $\text{Re}[x(t)]$ and $\text{Im}[x(t)]$ exhibit an oscillatory behavior. For $t > 0$ (case $\theta = 0°$), the $\text{Re}[x(t)]$ resembles a damped oscillator, where the $\text{Im}[x(t)]$ resembles a harmonic oscillator. However, for complex t both $\text{Re}[x(t)]$ and $\text{Im}[x(t)]$ oscillate with continuously increasing amplitude. The graph of $\text{Re}[x(t)]$ for $\theta = 5°$ is closer to the corresponding graph given in [51].

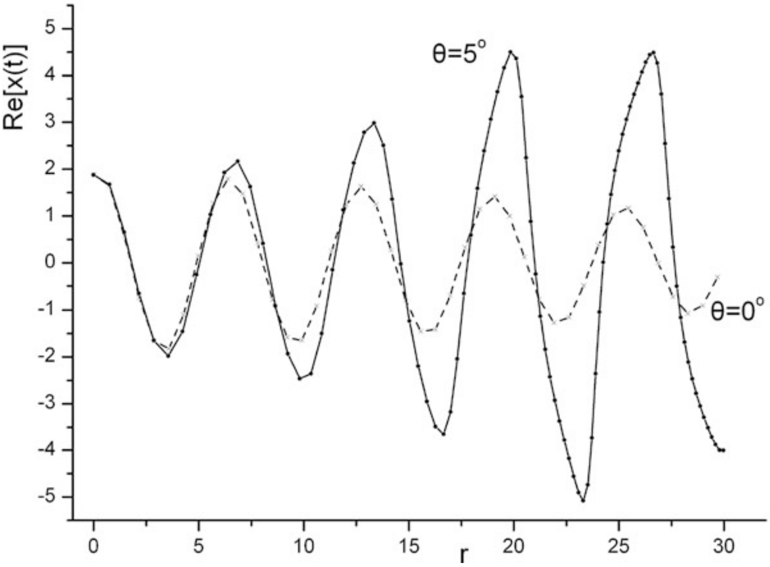


Fig. 12 Variation of the real part of the solution of (1)–(2) for values of the parameters as given in (97) and for values of θ 0° and 5°

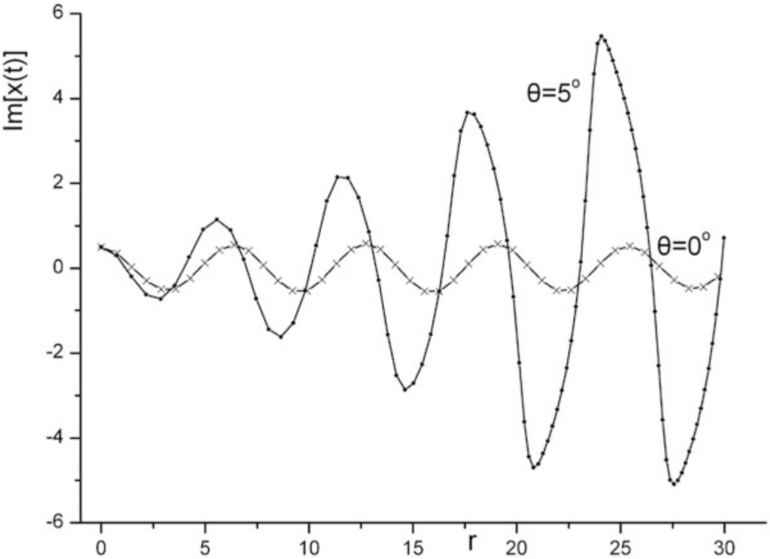


Fig. 13 Variation of the imaginary part of the solution of (1)–(2) for values of the parameters as given in (97) and for values of θ 0° and 5°

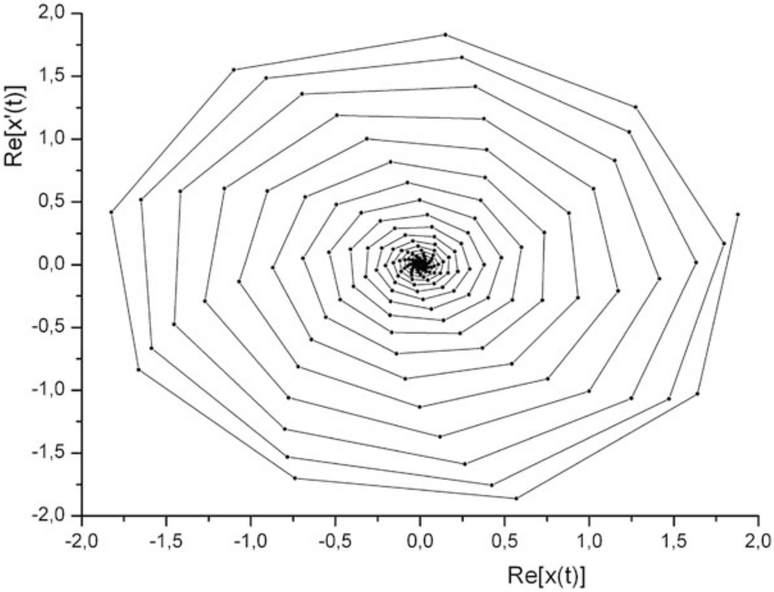


Fig. 14 Phase plane for the real part of the solution of (1)–(2) for values of the parameters as given in (97) and $\theta = 0°$

Finally, Figs. 14 and 15 show the phase plane for Re$[x(t)]$ of the solution $x(t)$ of (1)–(2) for the values (97), again for $\theta = 0°$ and $\theta = 5°$. In the first case, a spiral trajectory is observed which as t elapses approaches $(0, 0)$, whereas in the second case the formed spiral trajectory is drawn away from $(0, 0)$ and close to a limit cycle which seems to be formed. This limit cycle looks very similar to the well-known limit cycle of the real van der Pol equation. Similar graphs were also obtained for the phase plane of the corresponding Im$[x(t)]$ for the same values of θ.

From the application of this FAT to the IVP (1)–(2), some advantages and disadvantages of this technique should be obvious. Let's begin with the disadvantages:

- the continuous and not necessarily analytic solutions of ODEs, that often appear in physical problems, cannot be obtained with this technique.
- The method is not easily applicable for someone not familiar with operator theory.

On the other hand, the corresponding advantages compensate us for the aforementioned difficulties. These are

- The obtained solution always converges to the true solution of the ODE under consideration, due to the used isomorphisms.
- The method is accurate, since the only errors encountered in practice are the round-off errors for the truncated series (81) or (89), after taking into consideration a sufficiently large number of terms.

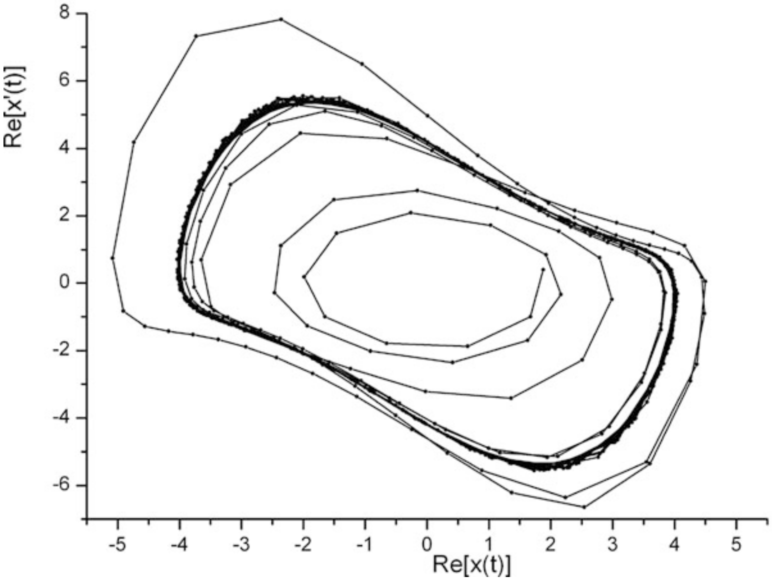


Fig. 15 Phase plane for the real part of the solution of (1)–(2) for values of the parameters as given in (97) and $\theta = 5°$

- It does not depend on the grid used, due to the fact that the computed solution of the ODE under consideration is based on the calculation of the coefficients x_n (or u_n, v_n) and is calculated analytically.
- It is very fast, as already demonstrated in [59].
- It is only slightly modified when complex solutions are calculated instead of real solutions, as Step A6 implies.
- The analysis of complex solutions may reveal the existence of points where the solutions "blow up", i.e. points where these solutions cease to be analytic. The location of these singularities in $\mathbb{C}$ form sectors in polar graphs. However only the first singularity can be detected in each sector and there is no ability by using this method, to examine the regions for larger r for which the first singularity is detected. Also, this method cannot give any information regarding the kind of the detected singularity.

6 Conclusions

The differential equations that describe many realistic problems are nonlinear and most of these cannot be solved explicitly using standard analytic techniques. In this chapter, we present several qualitative, approximate or numerical techniques in order to deal with such kind of equations. This was accomplished by considering the nonlinear ODE (1), subject to the initial conditions (2), which is used in various physical problems where oscillations appear.

We begun by obtaining some qualitative type of results regarding the trajectories of the corresponding autonomous equation (1) in the phase plane. This was achieved by explicitly finding its first integrals for specific choices of the appearing parameters and, by performing a standard phase plane analysis of the associated to (1), auxiliary first order system of ODEs (10). Finding a first integral for a nonlinear ODE is very useful, but unfortunately not always easy to be done. Furthermore, studying the dynamic properties of a nonlinear ODE is of paramount importance for understanding the behavior of its solution, since this includes the investigation of the existence of limit cycles, strange attractors, chaos, etc. The dynamic properties of a nonlinear ODE are investigated via rigorous mathematical analysis and/or with the aid of numerical techniques.

Then, we connected the nonlinear IVP (1)–(2) with the Green function of an auxiliary linear IVP. In this way, (1)–(2) was rewritten in the form of an integro-differential Volterra equation, which in the case where $a = 0$ is reduced to a pure integral equation. This integral equation is more appropriate than (1)–(2), for the application of fixed point theorems, which permits establishing existence and uniqueness results in specific spaces of interest. Instead of applying a numerical method to the IVP, we may apply a numerical method to the corresponding integrodifferential equation. This approach decreases the errors when using an FDM method. Moreover, this integral form of (1)–(2) is connected with the finite elements method.

Next, we calculated the approximate solution of (1)–(2) via the classical perturbation method and the homotopy analysis method. Both methods rely in calculating several consecutive approximations of the true solution of the IVP under consideration. This is done by solving a series of "easier" than (1)–(2) IVPs. Perturbation methods can be applied only when a small ($\ll 1$) positive parameter, the perturbation parameter, appears in the problem and may be successful when at least the first approximation of the solution (leading order term) can be found in closed form. In (1), three parameters appear. If any of them is chosen to be the perturbation parameter, the corresponding IVP for the leading order term, although "easier" than (1)–(2), remains nonlinear and thus, cannot be solved for all values of the remaining parameters. However, choosing all appearing parameters simultaneously as perturbation parameters and performing a three-parameter perturbation, provides a satisfactory approximate solution of (1)–(2).

The homotopy analysis method on the other hand, does not need the existence of a small parameter in the problem under consideration and reduces the solution of (1)–(2), to the solution of successive linear IVPs associated with (1)–(2). HAM gives us great freedom in constructing these linear IVPs, although certain rules should be followed. The obtained approximate solution depends on an auxiliary parameter $\hbar$. Appropriate choices of this parameter give us very satisfactory approximations of the solution of (1)–(2).

After having used two approximate techniques, we numerically solved (1)–(2) using the fourth order Runge-Kutta, the standard finite differences and the finite elements method. These methods are widely utilized for the numerical solution of nonlinear ODEs, while FDM and FEM are also used for the numerical solution of

PDEs. All numerical methods gave solutions which are in very good agreement with the approximate ones. Especially the fourth order RK method provided a solution which is very close to the corresponding solution obtained by HAM. The FDM is the dominated approach to numerical solutions of ODEs and PDEs, due to the fact that as a method is very easy to implement in a numerical code and most times provides very reliable results. Moreover, the FDM is a differential numerical method, whereas the FEM is an integral method. The integral methods are superior compared to the differential methods concerning the required smoothness of the functions needed for the numerical solution. The FEM is more systematic compared to FDM, but it takes place in more complicated spaces.

Finally, we solved (1)–(2) using a non-standard "discretization" based on a functional analytic technique. This technique enables finding a difference equation which is equivalent to (1) and not a discrete analogue of it. This FAT is very accurate and fast and the obtained solution always converges to the true solution of the problem under consideration. However, it requires a specific knowledge of operator theory and cannot by applied if the required abstract forms cannot be founded, in contrast for example with FDM methods which can be applied to almost all forms of ODEs. Maybe the most significant advantage of this FAT, is the fact that it can be only slightly modified in order to compute solutions of ODEs in the complex plane.

Overall, we would like to point out that there is a huge variety of techniques for studying and solving differential equations and here we presented very few of them. The choice of the technique depends not only on the problem under consideration, but also on the personal taste of the researcher.

Acknowledgements The authors would like to thank associate professor Efstratios E. Tzirtzilakis of the Department of Mechanical Engineering, Technological Educational Institute of Western Greece, for many fruitful discussions and comments.

References

1. G. Adomian, A review of the decomposition method in applied mathematics. J. Math. Anal. Appl. **135**(3), 501–544 (1988)
2. G. Adomian, *Solving Frontier Problems of Physics. The Decomposition Method* (Springer, Berlin, 1994)
3. R.P. Agarwal, D. O'Regan, *Ordinary and Partial Differential Equations: With Special Functions, Fourier Series, and Boundary Value Problems* (Springer Science+Business Media, LLC, New York, 2009)
4. G.D. Akrivis, V.A. Dougalis, *Numerical Methods for Ordinary Differential Equations* (Crete University Press, Heraklion, 2006)
5. G. Akrivis, Ch. Makridakis, Galerkin time-stepping method for nonlinear parabolic equations. M2AN Math. Model. Numer. Anal. **38**, 261–289 (2004)
6. V.I. Arnol'd, *Ordinary Differential Equations* (Springer, Berlin, 1992)
7. J. Awreicewicz, On the occurence of chaos in van der Pol-Duffing's oscillator. J. Sound Vib. **109**(3), 519–522 (1986)
8. C.M. Bender, S.A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers. Asymptotic Methods and Perturbation Theory* (Springer, New York, 1999)

9. L.A. Bergman, J.E. Hyatt, Green functions for transversely vibrating uniform Euler-Bernoulli beams subject to constant axial preload. J. Sound Vib. **134**(1), 175–180 (1989)
10. T.C. Bountis, L.B. Drossos, M. Lakshmanan, S. Parthasarathy, On the non-integrability of a family of Duffing-van der Pol oscillators. J. Phys. A Math. Gen. **26**, 6927–6942 (1993)
11. J.P. Boyd, *Chebyshev and Fourier Spectral Methods*, 2nd edn. (Dover, New York, 2000)
12. S.C. Brenner, L.R. Scott, *The Mathematical Theory of Finite Element Method* (Springer, New York, 1994)
13. J.C. Butcher, A stability property of implicit Runge-Kutta methods. BIT Numer. Math. **15**, 358–361 (1975)
14. J.H.E. Cartwright, O. Piro, The dynamics of Runge–Kutta methods. Int. J. Bifurc. Chaos **2**, 427–449 (1992)
15. V.K. Chandrasekar, M. Senthilvelan, M. Lakshmanan, New aspects of integrability of force-free Duffing-van der Pol oscillator and related nonlinear systems. J. Phys. A Math. Gen. **37**, 4527–4534 (2004)
16. A. Chen, J. Jiang, Periodic solution of the Duffing-Van der Pol oscillator by homotopy perturbation method. Int. J. Comput. Math. **87**(12), 2688–2696 (2010)
17. Y.M. Chen, J.K. Liu, Uniformly valid solution of limit cycle of the Duffing-van der Pol equation. Mech. Res. Commun. **36**, 845–850 (2009)
18. A. Chudzik, Synchronisation and periodisation of Duffing oscillators coupled by elastic beam: finite element method approach. J. Theor. Appl. Mech. **48**(2), 517–524 (2010)
19. F. Dal, The method of multiple time scales and finite differences method for the van del Pol oscillator with small fractional damping. Asian Res. J. Math. **2**(2), 1–11 (2017)
20. D.G. Duffy, *Green's Functions with Applications*, 2nd edn. (Chapman and Hall/CRC, Boca Raton, 2017)
21. C.J. Earle, R.S. Hamilton, A fixed point theorem for holomorphic mappings, In *Global Analysis Proceedings Symposium Pure Mathematics, vol. XVI, Berkeley, CA, (1968)* (American Mathematical Society, Providence, 1970), pp. 61–65
22. Z. Feng, Duffing-van der Pol-type oscillator systems. Discret. Contin. Dyn. Syst. Ser. S **7**(6), 1231–1257 (2014)
23. Z. Feng, G. Gao, J. Cui, Duffing-van der Pol-type oscillator system and its first integrals. Commun. Pure Appl. Anal. **10**(5), 1377–1392 (2011)
24. C.A.J. Fletcher, *Computational Techniques for Fluid Dynamics I* (Spinger, Berlin, 1988)
25. C.A.J. Fletcher, *Computational Techniques for Fluid Dynamics II* (Spinger, Berlin, 1988)
26. J. Gao, A.S. Selvarathinam, Y.J. Weitsman, Analysis of adhesively joined composite beams. J. Sandw. Struct. Mater. **1**, 323–339 (1999)
27. I. Gohberg, S. Goldberg, *Basic Operator Theory* (Birkhäuser, Basel, 1980)
28. D.H. Griffel, *Applied Functional Analysis* (Dover, New York, 2002)
29. M. Hatami, D.D. Ganji, M. Sheikholeslami, *Differential Transformation Method for Mechanical Engineering Problems* (Academic, Cambridge, 2016)
30. J.-H. He, Homotopy perturbation technique. Comput. Methods Appl. Mech. Eng. **178**(3–4), 257–262 (1999)
31. J.-H. He, Recent development of the homotopy perturbation method. Topol. Methods Nonlinear Anal. **31**(2), 205–209 (2008)
32. P.J. Hilton, *An Introduction to Homotopy Theory* (Cambridge University Press, Cambridge, 1953)
33. A.J.T. Horvath, Periodic solutions of a combined Van der Pol-Duffing differential equation. Int. J. Mech. Sci. **17**, 677–680 (1975)
34. P. Hou, K. Yuan, B. Chen, Study on the 3D Green's functions of the fluid and piezoelectric bimaterials. Theor. Appl. Mech. Lett. **7**, 105–116 (2017)
35. E.K. Ifantis, Spectral theory of the difference equation $f(n+1)+f(n-1)=[E-\phi(n)]f(n)$. J. Math. Phys. **10**(3), 421–425 (1969)
36. E.K. Ifantis, Solution of the Schrödinger equation in the Hardy–Lebesgue space. J. Math. Phys. **12**, 1961–1965 (1971)

37. E.K. Ifantis, An existence theory for functional-differential equations and functional-differential systems. J. Differ. Equ. **29**, 86–104 (1978)
38. E.K. Ifantis, Analytic solutions for nonlinear differential equations. J. Math. Anal. Appl. **124**(2), 339–380 (1987)
39. E.K. Ifantis, On the convergence of power-series whose coefficients satisfy a Poincaré-type linear and nonlinear difference equation. Complex Var. **9**, 63–80 (1987)
40. G. Iooss, D.D. Joseph, *Elementary Stability and Bifurcation Theory*, 2nd edn. (Springer, New York, 1990)
41. A. Iserles, *A First Course in the Numerical Analysis of Differential Equations* (Cambridge University Press, Cambridge, 1996)
42. Z. Jing, Z. Yang, T. Jiang, Complex dynamics in Duffing-Van der Pol equation. Chaos Solitons Fractals **27**, 722–747 (2006)
43. D.W. Jordan, P. Smith, *Nonlinear Ordinary Differential Equations*, 2nd edn. (Oxford University Press, Oxford, 1987)
44. A.Y.T. Leung, Q.C. Zhang, Complex normal form for strongly non-linear vibration systems exemplified by Duffing-van der Pol equation. J. Sound Vib. **213**(5), 907–914 (1998)
45. R.J. LeVeque, *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems* (Society for Industrial and Applied Mathematics, Philadelphia, 2007)
46. S. Liao, *Beyond Perturbation. Introduction to the Homotopy Analysis Method* (Chapman and Hall/CRC, Boca Raton, 2004)
47. S. Liao, Notes on the homotopy analysis method: some definitions and theorems. Commun. Nonlinear Sci. Numer. Simul. **14**(4), 983–997 (2009)
48. Y. Liu, S.J. Liao, Z. Li, A Maple package of automated derivation of homotopy analysis solution for periodic nonlinear oscillations. J. Syst. Sci. Complex. **25**(3), 594–616 (2012)
49. Y. Liu, S.J. Liao, Z. Li, Symbolic computation of strongly nonlinear periodic oscillations. J. Symb. Comput. **55**, 72–95 (2013)
50. J. Logan, *Applied Mathematics*, 2nd edn. (Wiley, New York, 1997)
51. G.M. Mahmoud, A.A.M. Farghaly, Chaos control of chaotic limit cycles of real and complex van der Pol oscillators. Chaos Solitons Fractals **21**, 915–924 (2004)
52. F.M. Moukam Kakmeni, S. Bowong, C. Tchawoua, E. Kaptouom, Strange attractors and chaos control in a Duffing-Van der Pol oscillator with two external periodic forces. J. Sound Vib. **277**, 783–799 (2004)
53. R.E. Mickens, *Nonstandard Finite Difference Models of Differential Equations* (World Scientific Publishing, River Edge, 1994)
54. R.E. Mickens, Nonstandard finite difference schemes for differential equations. J. Differ. Equ. Appl. **8**(9), 823–847 (2002)
55. A. Okasha El-Nady, M.M.A. Lashin, Approximate solution of nonlinear Duffing oscillator using Taylor expansion. J. Mech. Eng. Autom. **6**(5), 110–116 (2016)
56. A.H. Nayfeh, *Introduction to Perturbation Techniques* (Wiley, New York, 1981)
57. V.V. Nemytskii, V.V. Stepanov, *Qualitative Theory of Differential Equations* (Dover, New York, 1989)
58. E.N. Petropoulou, E.E. Tzirtzilakis, On the logistic equation in the complex plane. Numer. Funct. Anal. Optim. **34**(7), 770–790 (2013)
59. E.N. Petropoulou, P.D. Siafarikas, E.E. Tzirtzilakis, A "discretization" technique for the solution of ODEs. J. Math. Anal. Appl. **331**, 279–296 (2007)
60. E.N. Petropoulou, P.D. Siafarikas, E.E. Tzirtzilakis, A "discretization" technique for the solution of ODEs II. Numer. Funct. Anal. Optim. **30**(5–6), 613–631 (2009)
61. Z.-H. Qin, Y.-S. Chen, Singularity analysis of Duffing-van der Pol system with two bifurcation parameters under multi-frequency excitations. Appl. Math. Mech. (English Ed.) **31**(8), 1019–1026 (2010)
62. A. Quarteroni, R. Sacco, F. Saleri, *Numerical Mathematics* (Springer, New York, 2000)
63. J.A. Rad, S. Kazem, K. Parand, A numerical solution of the nonlinear controlled Duffing oscillator by radial basis functions. Comput. Math. Appl. **64**, 2049–2065 (2012)

64. S.S. Rao, *Mechanical Vibrations*, 5th edn. (Pearson Education, London, 2011)
65. J. Rebenda, Z. Šmarda, A differential transformation approach for solving functional differential equations with multiple delays. Commun. Nonlinear Sci. Numer. Simul. **48**, 246–257 (2017)
66. J. Rebenda, Z. Šmarda, Y. Khan, A new semi-analytical approach for numerical solving of Cauchy problem for differential equations with delay, Filomat **31**(15), 4725–4733 (2017)
67. M. Sathyamoorthy, *Nonlinear Analysis of Structures* (CRC Press, Boca Raton, 1998)
68. A.S. Soomro, G.A. Tularam, M.M. Shaikh, A comparison of numerical methods for solving the unforced van der Pol's equation. Math. Theory Model. **3**(2), 66–77 (2013)
69. W.-H. Steeb, N. Euler, *Nonlinear Evolution Equations and Painlevé Test* (World Scientific Publishing, Singapore, 1988)
70. W. Szemplińska-Stupnicka, J. Rudowski, The coexistence of periodic, almost-periodic and chaotic attractors in the van der Pol-Duffing oscillator. J. Sound Vib. **199**(2), 165–175 (1997)
71. M.E. Taylor, *Partial Differerential Equations I. Basic Theory*, 2nd edn. (Springer Science+Business Media, LLC, New York, 2011)
72. M.E. Taylor, *Partial Differerential Equations II. Qualitative Studies of Linear Equations*, 2nd edn. (Springer Science+Business Media, LLC, New York, 2011)
73. M.E. Taylor, *Partial Differerential Equations III. Nonlinear Equations*, 2nd edn. (Springer Science+Business Media, LLC, New York, 2011)
74. A. Venkatesan, M. Lakshmanan, Bifurcation and chaos in the double-well Duffing-van der Pol oscillator: numerical and analytical studies, Phys. Rev. E **56**(6), 6321–6330 (1997)
75. D. Wu, L. Yang, Y. Gao, Three-dimensional Green's functions of thermoporoelastic axisymmetric cones. Appl. Math. Model. **42**, 315–329 (2017)
76. M.A. Xenos, An Euler-Lagrange approach for studying blood flow in an aneurysmal geometry. Proc. R. Soc. A **473**, 20160774 (2017)
77. L. Xie, C. Zhang, C. Hwu, E. Pan, On novel explicit expressions of Green's function and its derivatives for magnetoelectroelastic materials. Eur. J. Mech. A Solid **60**, 134–144 (2016)
78. R. Yamapi, G. Filatrella, Strange attractors and synchronization dynamics of coupled Van der Pol-Duffing oscillators. Commun. Nonlinear Sci. Numer. Simul. **13**, 1121–1130 (2008)
79. J. Yu, W.-Z. Pan, R.-B. Zhang, Period-doubling cascades and strange attractors in extended Duffing-van der pol oscillator. Commun. Theor. Phys. (Beijing, China) **51**, 865–868 (2009)

On a Hilbert-Type Integral Inequality in the Whole Plane

Michael Th. Rassias and Bicheng Yang

1 Introduction

If $f(x), g(y) \geq 0$,

$$0 < \int_0^\infty f^2(x)dx < \infty \text{ and } 0 < \int_0^\infty g^2(y)dy < \infty ,$$

then we have the following well known Hilbert integral inequality (cf. [1]):

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{x+y} dxdy < \pi \left(\int_0^\infty f^2(x)dx \int_0^\infty g^2(y)dy \right)^{\frac{1}{2}} , \tag{1}$$

where the constant factor π is the best possible.

Recently, by the use of methods of weight functions and by introducing multi-parameters, several extensions of (1) were presented in books of B. Yang (cf. [2, 3]). Some Hilbert-type inequalities with the homogenous kernels of degree 0 and non-homogenous kernels were obtained in [4–9]. Some other kinds of Hilbert-type inequalities were established in [10–15]. Several of these inequalities are built in the quarter plane of the first quadrant.

M. Th. Rassias
Institute of Mathematics, University of Zurich, Zurich, Switzerland
Moscow Institute of Physics and Technology, Dolgoprudny, Russia
Institute for Advanced Study, Program in Interdisciplinary Studies, Princeton, NJ, USA
e-mail: michail.rassias@math.uzh.ch

B. Yang (✉)
Department of Mathematics, Guangdong University of Education, Guangzhou, Guangdong, People's Republic of China
e-mail: bcyang@gdei.edu.cn

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_23

By the use of methods of weight functions, in 2007, Yang [16] proved the following Hilbert-type integral inequality in the whole plane as follows:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{f(x)g(y)}{(1+e^{x+y})^{\lambda}}dxdy$$
$$< B\left(\frac{\lambda}{2},\frac{\lambda}{2}\right)\left(\int_{-\infty}^{\infty}e^{-\lambda x}f^{2}(x)dx\int_{-\infty}^{\infty}e^{-\lambda y}g^{2}(y)dy\right)^{\frac{1}{2}}, \tag{2}$$

where the constant factor $B(\frac{\lambda}{2},\frac{\lambda}{2})(\lambda > 0)$ is the best possible, and $B(u, v)$ is the beta function (cf. [17]). He et al. [18–31] proved some new Hilbert-type integral inequalities in the whole plane with the best possible constant factors. The methods in these papers are interesting and technically challenging.

In the present paper, still using methods of real analysis and weight functions, a new Hilbert-type integral inequality in the whole plane with multi-parameters and a best possible constant factor is formulated and proved. In the form of applications, the equivalent forms, some particular cases and the operator expressions are also considered.

2 Some Lemmas

In the following, we let $0 < \alpha_1 \leq \alpha_2 < \pi$, $\mu, \sigma > 0$, $\mu + \sigma = \lambda$, $\delta \in \{-1, 1\}$,

$$\gamma \in \{a; a = \frac{1}{2k+1}, 2k-1 \ \ (k \in \mathbf{N} = \{1, 2, \cdots\})\}.$$

Definition 1 We define the following weight functions:

$$\omega(\sigma, y) := \int_{-\infty}^{\infty}\max_{i\in\{1,2\}}\frac{|y|^{\sigma}|x|^{\delta\sigma-1}}{[\max\{|x^{\delta}y|^{\gamma}+(x^{\delta}y)^{\gamma}\cos\alpha_i, 1\}]^{\lambda/\gamma}}dx \ \ (y \in \mathbf{R}), \tag{3}$$

$$\varpi(\sigma, x) := \int_{-\infty}^{\infty}\max_{i\in\{1,2\}}\frac{|x|^{\delta\sigma}|y|^{\sigma-1}}{[\max\{|x^{\delta}y|^{\gamma}+(x^{\delta}y)^{\gamma}\cos\alpha_i, 1\}]^{\lambda/\gamma}}dy \ \ (x \in \mathbf{R}). \tag{4}$$

Lemma 1 *The following expressions hold true:*

$$\omega(\sigma, y) = \varpi(\sigma, x) = K(\sigma) \ \ (y, x \in \mathbf{R}\backslash\{0\}), \tag{5}$$

where

$$K(\sigma) := \frac{1}{2^{\frac{\sigma}{\gamma}}}\left[\left(\sec\frac{\alpha_2}{2}\right)^{\frac{2\sigma}{\gamma}}+\left(\csc\frac{\alpha_1}{2}\right)^{\frac{2\sigma}{\gamma}}\right]\frac{\lambda}{\mu\sigma} \in \mathbf{R}_{+} = (0, \infty). \tag{6}$$

Proof For $y \in \mathbf{R}\backslash\{0\}$, setting $u = x^{\delta} y$ in (3), we derive that

$$x = y^{\frac{-1}{\delta}} u^{\frac{1}{\delta}},\ dx = \frac{1}{\delta} y^{\frac{-1}{\delta}} u^{\frac{1}{\delta}-1} du$$

and

$$\begin{aligned}\omega(\sigma, y) &= |\frac{1}{\delta}| \int_{-\infty}^{\infty} \max_{i\in\{1,2\}} \frac{1}{(\max\{|u|^{\gamma} + u^{\gamma}\cos\alpha_i, 1\})^{\lambda/\gamma}} |u|^{\sigma-1} du \\ &= K_1(\sigma) + K_2(\sigma),\end{aligned} \tag{7}$$

where

$$\begin{aligned}K_1(\sigma) &:= \int_{-\infty}^{0} \max_{i\in\{1,2\}} \frac{1}{[\max\{(-u)^{\gamma}(1-\cos\alpha_i), 1\}]^{\lambda/\gamma}} (-u)^{\sigma-1} du, \\ K_2(\sigma) &:= \int_{0}^{\infty} \max_{i\in\{1,2\}} \frac{1}{[\max\{u^{\gamma}(1+\cos\alpha_i), 1\}]^{\lambda/\gamma}} u^{\sigma-1} du.\end{aligned}$$

Setting $v = u^{\gamma}(1 + \cos\alpha_i)$ in the integral $K_2(\sigma)$, we get

$$u = \frac{1}{(1+\cos\alpha_i)^{1/\gamma}} v^{\frac{1}{\gamma}},\ \ du = \frac{1}{\gamma(1+\cos\alpha_i)^{1/\gamma}} v^{\frac{1}{\gamma}-1} dv$$

and

$$\begin{aligned}K_2(\sigma) &= \int_0^{\infty} \max_{i\in\{1,2\}} \frac{1}{\gamma(1+\cos\alpha_i)^{\sigma/\gamma}} \frac{1}{(\max\{v, 1\})^{\lambda/\gamma}} v^{\frac{\sigma}{\gamma}-1} dv \\ &= \frac{1}{\gamma(1+\cos\alpha_2)^{\sigma/\gamma}} \int_0^{\infty} \frac{1}{(\max\{v, 1\})^{\lambda/\gamma}} v^{\frac{\sigma}{\gamma}-1} dv \\ &= \frac{1}{\gamma(1+\cos\alpha_2)^{\sigma/\gamma}} \left(\int_0^1 v^{\frac{\sigma}{\gamma}-1} dv + \int_1^{\infty} \frac{1}{v^{\lambda/\gamma}} v^{\frac{\sigma}{\gamma}-1} dv \right) \\ &= \frac{1}{(1+\cos\alpha_2)^{\sigma/\gamma}} \frac{\lambda}{\mu\sigma} = \frac{1}{2^{\sigma/\gamma}} (\sec\frac{\alpha_2}{2})^{\frac{2\sigma}{\gamma}} \frac{\lambda}{\mu\sigma} \in \mathbf{R}_+.\end{aligned}$$

Setting $v = -u$ in the integral of $K_1(\sigma)$, we similarly obtain that

$$\begin{aligned}K_1(\sigma) &= \int_0^{\infty} \max_{i\in\{1,2\}} \frac{1}{\{\max\{v^{\gamma}[1+\cos(\pi-\alpha_i)], 1\}\}^{\lambda/\gamma}} v^{\sigma-1} dv \\ &= \frac{1}{[1+\cos(\pi-\alpha_1)]^{\sigma/\gamma}} \frac{\lambda}{\mu\sigma} = \frac{1}{2^{\sigma/\gamma}} (\csc\frac{\alpha_1}{2})^{\frac{2\sigma}{\gamma}} \frac{\lambda}{\mu\sigma} \in \mathbf{R}_+,\end{aligned}$$

namely, we have $\omega(\sigma, y) = K(\sigma) \in \mathbf{R}_+$.

For $x \in \mathbf{R}\backslash\{0\}$, setting $u = x^{\delta} y$ in (4), we find $y = x^{-\delta}u$, $dy = x^{-\delta}du$ and

$$\varpi(\sigma, x) = \int_{-\infty}^{\infty} \max_{i\in\{1,2\}} \frac{1}{(\max\{|u|^{\gamma} + u^{\gamma}\cos\alpha_i, 1\})^{\lambda/\gamma}} |u|^{\sigma-1} du = K(\sigma).$$

Hence (5) follows and thus the lemma is proved.

Remark 1 If we replace $\max_{i\in\{1,2\}}$ with $\min_{i\in\{1,2\}}$ in (3) and (4), then (6) is valid by exchanging α_1 and α_2.

Lemma 2 *Let us assume that* $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$, $K(\sigma)$ *is as defined in (6), and* $f(x)$ *is a non-negative measurable function in* $\mathbf{R}$. *The following inequality holds true:*

$$J := \int_{-\infty}^{\infty} |y|^{p\sigma-1} \left\{ \int_{-\infty}^{\infty} \max_{i\in\{1,2\}} \frac{f(x)}{[\max\{|x^{\delta}y|^{\gamma} + (x^{\delta}y)^{\gamma}\cos\alpha_i, 1\}]^{\frac{\lambda}{\gamma}}} dx \right\}^{p} dy$$
$$\leq K^{p}(\sigma) \int_{-\infty}^{\infty} |x|^{p(1-\delta\sigma)-1} f^{p}(x) dx. \tag{8}$$

Proof In the sequel, we set

$$H^{(\delta)}(x, y) := \max_{i\in\{1,2\}} \frac{1}{[\max\{|x^{\delta}y|^{\gamma} + (x^{\delta}y)^{\gamma}\cos\alpha_i, 1\}]^{\lambda/\gamma}} \quad (x, y \in \mathbf{R}). \tag{9}$$

By Hölder's inequality (cf. [32]), we obtain that

$$\left(\int_{-\infty}^{\infty} H^{(\delta)}(x, y) f(x) dx \right)^{p}$$
$$= \left\{ \int_{-\infty}^{\infty} H^{(\delta)}(x, y) [\frac{|x|^{(1-\delta\sigma)/q}}{|y|^{(1-\sigma)/p}} f(x)][\frac{|y|^{(1-\sigma)/p}}{|x|^{(1-\delta\sigma)/q}}] dx \right\}^{p}$$
$$\leq \int_{-\infty}^{\infty} H^{(\delta)}(x, y) \frac{|x|^{(1-\delta\sigma)(p-1)}}{|y|^{1-\sigma}} f^{p}(x) dx$$
$$\times \left[\int_{-\infty}^{\infty} H^{(\delta)}(x, y) \frac{|y|^{(1-\sigma)(q-1)}}{|x|^{1-\delta\sigma}} dx \right]^{p-1} \tag{10}$$
$$= \frac{(\omega(\sigma, y))^{p-1}}{|y|^{p\sigma-1}} \int_{-\infty}^{\infty} H^{(\delta)}(x, y) \frac{|x|^{(1-\delta\sigma)(p-1)}}{|y|^{1-\sigma}} f^{p}(x) dx.$$

Thus by (5) and Fubini's theorem (cf. [33]), we have

$$J \leq K^{p-1}(\sigma) \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} H^{(\delta)}(x, y) \frac{|x|^{(1-\delta\sigma)(p-1)}}{|y|^{1-\sigma}} f^p(x) dx \right] dy$$
$$= K^{p-1}(\sigma) \int_{-\infty}^{\infty} \varpi(\sigma, x) |x|^{p(1-\delta\sigma)-1} f^p(x) dx.$$

By (5) we also obtain the inequality (8). This completes the proof of the lemma.

3 Main Results and Corollaries

Theorem 1 *If* $p > 1, \frac{1}{p} + \frac{1}{q} = 1,\ K(\sigma)$ *is as defined by (6),* $f(x), g(y) \geq 0,$

$$0 < \int_{-\infty}^{\infty} |x|^{p(1-\delta\sigma)-1} f^p(x) dx < \infty \ \ and \ \ 0 < \int_{-\infty}^{\infty} |y|^{q(1-\sigma)-1} g^q(y) dy < \infty\,.$$

then we have the following equivalent inequalities:

$$I := \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \max_{i \in \{1,2\}} \frac{f(x) g(y)}{(\max\{|x^\delta y|^\gamma + (x^\delta y)^\gamma \cos\alpha_i, 1\})^{\frac{\lambda}{\gamma}}} dx dy$$
$$< K(\sigma) \left[\int_{-\infty}^{\infty} |x|^{p(1-\delta\sigma)-1} f^p(x) dx \right]^{\frac{1}{p}} \left[\int_{-\infty}^{\infty} |y|^{q(1-\sigma)-1} g^q(y) dy \right]^{\frac{1}{q}}, \tag{11}$$

$$J := \int_{-\infty}^{\infty} |y|^{p\sigma-1} \left\{ \int_{-\infty}^{\infty} \max_{i \in \{1,2\}} \frac{f(x)}{[\max\{|x^\delta y|^\gamma + (x^\delta y)^\gamma \cos\alpha_i, 1\}]^{\frac{\lambda}{\gamma}}} dx \right\}^p dy$$
$$< K^p(\sigma) \int_{-\infty}^{\infty} |x|^{p(1-\delta\sigma)-1} f^p(x) dx, \tag{12}$$

where, the constant factors $K(\sigma)$ *and* $K^p(\sigma)$ *are the best possible.*

In particular, for $\alpha_1 = \alpha_2 = \alpha \in (0, \pi),\ \gamma = 1$ *in (11) and (12), setting*

$$k(\sigma) := \frac{1}{2^\sigma} \left[\left(\sec\frac{\alpha}{2} \right)^{2\sigma} + \left(\csc\frac{\alpha}{2} \right)^{2\sigma} \right] \frac{\lambda}{\mu\sigma}, \tag{13}$$

we deduce the following equivalent inequalities:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{(\max\{|x^{\delta}y|+x^{\delta}y\cos\alpha,1\})^{\lambda}}f(x)g(y)dxdy$$
$$< k(\sigma)\left[\int_{-\infty}^{\infty}|x|^{p(1-\delta\sigma)-1}f^{p}(x)dx\right]^{\frac{1}{p}}\left[\int_{-\infty}^{\infty}|y|^{q(1-\sigma)-1}g^{q}(y)dy\right]^{\frac{1}{q}}, \quad (14)$$

$$\int_{-\infty}^{\infty}|y|^{p\sigma-1}\left[\int_{-\infty}^{\infty}\frac{1}{(\max\{|x^{\delta}y|+x^{\delta}y\cos\alpha,1\})^{\lambda}}f(x)dx\right]^{p}dy$$
$$< k^{p}(\sigma)\int_{-\infty}^{\infty}|x|^{p(1-\sigma)-1}f^{p}(x)dx. \quad (15)$$

Proof If (10) takes the form of equality for a $y \neq 0$, then there exists constants A and B, satisfying $A^2+B^2>0$, and

$$A\frac{|x|^{(1-\delta\sigma)(p-1)}}{|y|^{1-\sigma}}f^{p}(x)=B\frac{|y|^{(1-\sigma)(q-1)}}{|x|^{1-\delta\sigma}} \quad \text{a. e. in } \mathbf{R}$$

(cf. [32]). We have $A \neq 0$ (otherwise, $B=A=0$), from which it follows that

$$|x|^{p(1-\delta\sigma)-1}f^{p}(x)=|y|^{q(1-\sigma)}\frac{B}{A|x|} \quad \text{a. e. in } \mathbf{R}.$$

However, this contradicts the fact that

$$0<\int_{-\infty}^{\infty}|x|^{p(1-\delta\sigma)-1}f^{p}(x)dx<\infty.$$

Hence, (10) takes the form of strict inequality. So does (8), namely, (12) follows.

In view of Hölder's inequality (cf. [32]), we also obtain that

$$I=\int_{-\infty}^{\infty}\left(|y|^{\sigma-\frac{1}{p}}\int_{-\infty}^{\infty}H^{(\delta)}(x,y)f(x)dx\right)(|y|^{\frac{1}{p}-\sigma}g(y))dy$$
$$\leq J^{\frac{1}{p}}\left[\int_{-\infty}^{\infty}|y|^{q(1-\sigma)-1}g^{q}(y)dy\right]^{\frac{1}{q}}. \quad (16)$$

Then by (12) we derive (11). On the other hand, assuming that (11) holds true, we set

$$g(y):=|y|^{p\sigma-1}\left(\int_{-\infty}^{\infty}H^{(\delta)}(x,y)f(x)dx\right)^{p-1} \quad (y\in\mathbf{R}).$$

Then it follows that

$$J=\int_{-\infty}^{\infty}|y|^{q(1-\sigma)-1}g^q(y)dy\,.$$

In view of (9), we have $J<\infty$. If $J=0$, then (12) is trivially valid. If $0<J<\infty$, then in view of (11), we get

$$\begin{aligned}&\int_{-\infty}^{\infty}|y|^{q(1-\sigma)-1}g^q(y)dy=J=I\\&<K(\sigma)\left[\int_{-\infty}^{\infty}|x|^{p(1-\delta\sigma)-1}f^p(x)dx\right]^{\frac{1}{p}}\left[\int_{-\infty}^{\infty}|y|^{q(1-\sigma)-1}g^q(y)dy\right]^{\frac{1}{q}},\end{aligned}\tag{17}$$

$$J^{\frac{1}{p}}=\left[\int_{-\infty}^{\infty}|y|^{q(1-\sigma)-1}g^q(y)dy\right]^{\frac{1}{p}}<K(\sigma)\left[\int_{-\infty}^{\infty}|x|^{p(1-\delta\sigma)-1}f^p(x)dx\right]^{\frac{1}{p}},\tag{18}$$

namely, (12) follows, which is equivalent to (11).

We set $E_\delta:=\{x\in\mathbf{R};\,|x|^\delta\geq 1\}$, and

$$E_\delta^+:=E_\delta\cap\mathbf{R}_+=\{x\in\mathbf{R}_+;\,x^\delta\geq 1\}.$$

For any $\varepsilon>0$, we define $\tilde{f}(x)$, $\tilde{g}(y)$ as follows:

$$\tilde{f}(x):=\begin{cases}|x|^{\delta(\sigma-\frac{2\varepsilon}{p})-1}, & x\in E_\delta\\ 0, & x\in\mathbf{R}\backslash E_\delta\end{cases},$$

$$\tilde{g}(y):=\begin{cases}0, & y\in(-\infty,-1)\cup(1,\infty)\\ |y|^{\sigma+\frac{2\varepsilon}{q}-1}, & y\in[-1,1]\end{cases}.$$

Then we get

$$\begin{aligned}\tilde{L}\;:&=\left[\int_{-\infty}^{\infty}|x|^{p(1-\delta\sigma)-1}\tilde{f}^p(x)dx\right]^{\frac{1}{p}}\left[\int_{-\infty}^{\infty}|y|^{q(1-\sigma)-1}\tilde{g}^q(y)dy\right]^{\frac{1}{q}}\\&=2(\int_{E_\delta^+}x^{-2\delta\varepsilon-1}dx)^{\frac{1}{p}}(\int_0^1 y^{2\varepsilon-1}dy)^{\frac{1}{q}}=\frac{1}{\varepsilon}.\end{aligned}$$

We have

$$\begin{aligned}h(x):&=\int_{-1}^{1}\max_{i\in\{1,2\}}\frac{|y|^{\sigma+\frac{2\varepsilon}{q}-1}}{[\max\{|x^\delta y|^\gamma+(x^\delta y)^\gamma\cos\alpha_i,1\}]^{\lambda/\gamma}}dy\\&=\int_{-1}^{1}\max_{i\in\{1,2\}}\frac{|Y|^{\sigma+\frac{2\varepsilon}{q}-1}}{[\max\{|-x^\delta Y|^\gamma+(-x^\delta Y)^\gamma\cos\alpha_i,1\}]^{\lambda/\gamma}}dY\\&=h(-x),\end{aligned}$$

namely, $h(x)$ is an even function. Since

$$\begin{aligned}\tilde{I} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H^{(\delta)}(x,y)\tilde{f}(x)\tilde{g}(y)dxdy \\ &= \int_{E_\delta} |x|^{\delta(\sigma-\frac{2\varepsilon}{p})-1} h(x)dx = 2\int_{E_\delta^+} x^{\delta(\sigma-\frac{2\varepsilon}{p})-1} h(x)dx \\ &= 2\int_{E_\delta^+} x^{-2\delta\varepsilon-1}\left[\int_{-x^\delta}^{x^\delta} \max_{i\in\{1,2\}} \frac{|u|^{\sigma+\frac{2\varepsilon}{q}-1}}{(\max\{|u|^\gamma + u^\gamma\cos\alpha_i, 1\})^{\lambda/\gamma}}du\right]dx \;(u = x^\delta y),\end{aligned}$$

setting $v = x^\delta$ in the above integral, by Fubini's theorem (cf. [33]), we obtain

$$\begin{aligned}\tilde{I} &= 2\int_1^\infty v^{-2\varepsilon-1}\left[\int_{-v}^{v} \max_{i\in\{1,2\}} \frac{|u|^{\sigma+\frac{2\varepsilon}{q}-1}}{(\max\{|u|^\gamma + u^\gamma\cos\alpha_i, 1\})^{\lambda/\gamma}}du\right]dv \\ &= 2\int_1^\infty v^{-2\varepsilon-1}\left\{\int_0^v [\max_{i\in\{1,2\}} \frac{1}{(\max\{u^\gamma(1+\cos\alpha_i), 1\})^{\lambda/\gamma}}\right. \\ &\quad \left. + \max_{i\in\{1,2\}} \frac{1}{(\max\{u^\gamma(1-\cos\alpha_i), 1\})^{\lambda/\gamma}}]u^{\sigma+\frac{2\varepsilon}{q}-1}du\right\}dv \\ &= 2\int_1^\infty v^{-2\varepsilon-1}\left\{\int_0^v [\frac{1}{(\max\{u^\gamma(1+\cos\alpha_2), 1\})^{\lambda/\gamma}}\right. \\ &\quad \left. + \frac{1}{(\max\{u^\gamma(1-\cos\alpha_1), 1\})^{\lambda/\gamma}}]u^{\sigma+\frac{2\varepsilon}{q}-1}du\right\}dv \\ &= 2\int_1^\infty v^{-2\varepsilon-1}\left\{\int_0^1 [\frac{u^{\sigma+\frac{2\varepsilon}{q}-1}}{(\max\{u^\gamma(1+\cos\alpha_2), 1\})^{\lambda/\gamma}}\right. \\ &\quad \left. + \frac{u^{\sigma+\frac{2\varepsilon}{q}-1}}{(\max\{u^\gamma(1-\cos\alpha_1), 1\})^{\lambda/\gamma}}]du\right\}dv \\ &\quad + 2\int_1^\infty v^{-2\varepsilon-1}\left\{\int_1^v [\frac{u^{\sigma+\frac{2\varepsilon}{q}-1}}{(\max\{u^\gamma(1+\cos\alpha_2), 1\})^{\lambda/\gamma}}\right. \\ &\quad \left. + \frac{u^{\sigma+\frac{2\varepsilon}{q}-1}}{(\max\{u^\gamma(1-\cos\alpha_1), 1\})^{\lambda/\gamma}}]du\right\}dv\end{aligned}$$

$$= \frac{1}{\varepsilon}\int_0^1 \left\{ \frac{u^{\sigma+\frac{2\varepsilon}{q}-1}}{[\max\{u^\gamma(1+\cos\alpha_2), 1\}]^{\frac{\lambda}{\gamma}}} + \frac{u^{\sigma+\frac{2\varepsilon}{q}-1}}{[\max\{u^\gamma(1-\cos\alpha_1), 1\}]^{\frac{\lambda}{\gamma}}} \right\} du + 2\int_1^\infty \left(\int_u^\infty v^{-2\varepsilon-1}dv\right)$$

$$\times \left\{ \frac{u^{\sigma+\frac{2\varepsilon}{q}-1}}{[\max\{u^\gamma(1+\cos\alpha_2), 1\}]^{\lambda/\gamma}} + \frac{u^{\sigma+\frac{2\varepsilon}{q}-1}}{[\max\{u^\gamma(1-\cos\alpha_1), 1\}]^{\lambda/\gamma}} \right\} du$$

$$= \frac{1}{\varepsilon}\left\{ \int_0^1 [\frac{u^{\sigma+\frac{2\varepsilon}{q}-1}}{(\max\{u(1+\cos\alpha_2)^{1/\gamma}, 1\})^\lambda} + \frac{u^{\sigma+\frac{2\varepsilon}{q}-1}}{(\max\{u(1-\cos\alpha_1)^{1/\gamma}, 1\})^\lambda}]du \right.$$

$$\left. + \int_1^\infty [\frac{u^{\sigma-\frac{2\varepsilon}{p}-1}}{(\max\{u(1+\cos\alpha_2)^{1/\gamma}, 1\})^\lambda} + \frac{u^{\sigma-\frac{2\varepsilon}{p}-1}}{(\max\{u(1-\cos\alpha_1)^{1/\gamma}, 1\})^\lambda}]du \right\}.$$

If the constant factor $K(\sigma)$ in (11) is not the best possible, then there exists a positive constant $k \le K(\sigma)$, such that (11) is valid when replacing $K(\sigma)$ by k. In particular, we have $\varepsilon\tilde{I} < \varepsilon k\tilde{L}$, and

$$\int_0^1 \left\{ \frac{u^{\sigma+\frac{2\varepsilon}{q}-1}}{[\max\{u(1+\cos\alpha_2)^{1/\gamma}, 1\}]^\lambda} + \frac{u^{\sigma+\frac{2\varepsilon}{q}-1}}{[\max\{u(1-\cos\alpha_1)^{1/\gamma}, 1\}]^\lambda} \right\} du$$

$$+ \int_1^\infty \left\{ \frac{u^{\sigma-\frac{2\varepsilon}{p}-1}}{[\max\{u(1+\cos\alpha_2)^{1/\gamma}, 1\}]^\lambda} + \frac{u^{\sigma-\frac{2\varepsilon}{p}-1}}{[\max\{u(1-\cos\alpha_1)^{1/\gamma}, 1\}]^\lambda} \right\} du$$

$$= \varepsilon\tilde{I} < \varepsilon k\tilde{L} = k. \tag{19}$$

By (7) and Levi's theorem (cf. [33]), we get

$$K(\sigma) = K_2(\sigma) + K_1(\sigma) = \int_0^\infty \max_{i\in\{1,2\}} \frac{1}{[\max\{u^\gamma(1+\cos\alpha_i), 1\}]^{\lambda/\gamma}} u^{\sigma-1}du$$

$$+ \int_0^\infty \max_{i\in\{1,2\}} \frac{1}{[\max\{u^\gamma(1-\cos\alpha_i), 1\}]^{\lambda/\gamma}} u^{\sigma-1}du$$

$$= \int_0^\infty \frac{u^{\sigma-1}du}{[\max\{u(1+\cos\alpha_2)^{\frac{1}{\gamma}}, 1\}]^\lambda} + \int_0^\infty \frac{u^{\sigma-1}du}{[\max\{u(1-\cos\alpha_1)^{\frac{1}{\gamma}}, 1\}]^\lambda}$$

$$= \int_0^1 \lim_{\varepsilon\to 0^+}\left\{\frac{u^{\sigma+\frac{2\varepsilon}{q}-1}}{[\max\{u(1+\cos\alpha_2)^{1/\gamma},1\}]^\lambda}+\frac{u^{\sigma+\frac{2\varepsilon}{q}-1}}{[\max\{u(1-\cos\alpha_1)^{1/\gamma},1\}]^\lambda}\right\}du$$

$$+\int_1^\infty \lim_{\varepsilon\to 0^+}\left\{\frac{u^{\sigma-\frac{2\varepsilon}{p}-1}}{[\max\{u(1+\cos\alpha_2)^{1/\gamma},1\}]^\lambda}+\frac{u^{\sigma-\frac{2\varepsilon}{p}-1}}{[\max\{u(1-\cos\alpha_1)^{1/\gamma},1\}]^\lambda}\right\}du$$

$$= \lim_{\varepsilon\to 0^+}\left\{\int_0^1 [\frac{u^{\sigma+\frac{2\varepsilon}{q}-1}}{(\max\{u(1+\cos\alpha_2)^{1/\gamma},1\})^\lambda}+\frac{u^{\sigma+\frac{2\varepsilon}{q}-1}}{(\max\{u(1-\cos\alpha_1)^{1/\gamma},1\})^\lambda}]du\right.$$

$$\left.+\int_1^\infty [\frac{u^{\sigma-\frac{2\varepsilon}{p}-1}}{(\max\{u(1+\cos\alpha_2)^{1/\gamma},1\})^\lambda}+\frac{u^{\sigma-\frac{2\varepsilon}{p}-1}}{(\max\{u(1-\cos\alpha_1)^{1/\gamma},1\})^\lambda}]du\right\}$$

$$\leq k.$$

Hence, the constant factor $k = K(\sigma)$ in (11) is the best possible.

The constant factor in (12) is still the best possible. Otherwise, we would reach a contradiction by (16), that the constant factor in (11) is not the best possible. This completes the proof of the theorem.

Corollary 1 *For $\delta = 1$ in (11) and (12), we obtain the following equivalent inequalities with the non-homogeneous kernel:*

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\max_{i\in\{1,2\}}\frac{f(x)g(y)}{(\max\{|xy|^\gamma+(xy)^\gamma\cos\alpha_i,1\})^{\frac{\lambda}{\gamma}}}dxdy$$
$$< K(\sigma)\left[\int_{-\infty}^{\infty}|x|^{p(1-\sigma)-1}f^p(x)dx\right]^{\frac{1}{p}}\left[\int_{-\infty}^{\infty}|y|^{q(1-\sigma)-1}g^q(y)dy\right]^{\frac{1}{q}}, \quad (20)$$

$$\int_{-\infty}^{\infty}|y|^{p\sigma-1}\left\{\int_{-\infty}^{\infty}\max_{i\in\{1,2\}}\frac{f(x)}{[\max\{|xy|^\gamma+(xy)^\gamma\cos\alpha_i,1\}]^{\frac{\lambda}{\gamma}}}dx\right\}^p dy$$
$$< K^p(\sigma)\int_{-\infty}^{\infty}|x|^{p(1-\sigma)-1}f^p(x)dx, \quad (21)$$

where, the constant factors $K(\sigma)$ and $K^p(\sigma)$ are the best possible. In particular, for $\alpha_1 = \alpha_2 = \alpha \in (0,\pi)$, $\gamma = 1$ in (20) and (21), we have the following equivalent inequalities:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{f(x)g(y)}{(\max\{|xy|+xy\cos\alpha,1\})^{\lambda}}dxdy$$
$$< k(\sigma)\left[\int_{-\infty}^{\infty}|x|^{p(1-\sigma)-1}f^{p}(x)dx\right]^{\frac{1}{p}}\left[\int_{-\infty}^{\infty}|y|^{q(1-\sigma)-1}g^{q}(y)dy\right]^{\frac{1}{q}}, \quad (22)$$

$$\int_{-\infty}^{\infty}|y|^{p\sigma-1}\left[\int_{-\infty}^{\infty}\frac{f(x)}{(\max\{|xy|+xy\cos\alpha,1\})^{\lambda}}dx\right]^{p}dy$$
$$< k^{p}(\sigma)\int_{-\infty}^{\infty}|x|^{p(1-\sigma)-1}f^{p}(x)dx, \quad (23)$$

where $k(\sigma)$ is indicated by (13).

Corollary 2 *For $\delta = -1$ in (11) and (12), replacing $|x|^{\lambda}f(x)$ by $f(x)$, we obtain*

$$0 < \int_{-\infty}^{\infty}|x|^{p(1-\mu)-1}f^{p}(x)dx < \infty,$$

as well as the following equivalent inequalities with the homogeneous kernel:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\max_{i\in\{1,2\}}\frac{f(x)g(y)}{(\max\{|y|^{\gamma}+sgn(x)y^{\gamma}\cos\alpha_i,|x|^{\gamma}\})^{\lambda/\gamma}}dxdy$$
$$< K(\sigma)\left[\int_{-\infty}^{\infty}|x|^{p(1-\mu)-1}f^{p}(x)dx\right]^{\frac{1}{p}}\left[\int_{-\infty}^{\infty}|y|^{q(1-\sigma)-1}g^{q}(y)dy\right]^{\frac{1}{q}}, \quad (24)$$

$$\int_{-\infty}^{\infty}|y|^{p\sigma-1}\left[\int_{-\infty}^{\infty}\max_{i\in\{1,2\}}\frac{f(x)}{(\max\{|y|^{\gamma}+sgn(x)y^{\gamma}\cos\alpha_i,|x|^{\gamma}\})^{\lambda/\gamma}}dx\right]^{p}dy$$
$$< K^{p}(\sigma)\int_{-\infty}^{\infty}|x|^{p(1-\mu)-1}f^{p}(x)dx, \quad (25)$$

where the constant factors $K(\sigma)$ and $K^{p}(\sigma)$ are the best possible. In particular, for $\alpha_1 = \alpha_2 = \alpha \in (0,\pi)$, $\gamma = 1$ in (24) and (25), we obtain the following equivalent inequalities:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{(\max\{|y|+sgn(x)y\cos\alpha,|x|\})^{\lambda}}f(x)g(y)dxdy$$
$$< k(\sigma)\left[\int_{-\infty}^{\infty}|x|^{p(1-\mu)-1}f^{p}(x)dx\right]^{\frac{1}{p}}\left[\int_{-\infty}^{\infty}|y|^{q(1-\sigma)-1}g^{q}(y)dy\right]^{\frac{1}{q}}, \quad (26)$$

$$\int_{-\infty}^{\infty} |y|^{p\sigma-1}\left[\int_{-\infty}^{\infty} \frac{1}{(\max\{|y| + sgn(x)y\cos\alpha, |x|\})^{\lambda}} f(x)dx\right]^{p} dy$$
$$< k^{p}(\sigma)\int_{-\infty}^{\infty} |x|^{p(1-\mu)-1} f^{p}(x)dx, \tag{27}$$

where $k(\sigma)$ is indicated by (13).

4 Operator Expressions

Suppose that $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$. We define the following functions:

$$\varphi(x) := |x|^{p(1-\delta\sigma)-1},\ \psi(y) := |y|^{q(1-\sigma)-1},\ \phi(x) := |x|^{p(1-\mu)-1}(x, y \in \mathbf{R}),$$

wherefrom $\psi^{1-p}(y) = |y|^{p\sigma-1}$. Define the following real normed linear space:

$$L_{p,\varphi}(\mathbf{R}) \ : \ = \left\{f : ||f||_{p,\varphi} := \left(\int_{-\infty}^{\infty} \varphi(x)|f(x)|^{p}dx\right)^{\frac{1}{p}} < \infty\right\},$$

$$L_{p,\psi^{1-p}}(\mathbf{R}) = \left\{h : ||h||_{p,\psi^{1-p}} = \left(\int_{-\infty}^{\infty} \psi^{1-p}(y)|h(y)|^{p}dy\right)^{\frac{1}{p}} < \infty\right\},$$

$$L_{p,\phi}(\mathbf{R}) = \left\{g : ||g||_{p,\phi} = \left(\int_{-\infty}^{\infty} \phi(x)|g(x)|^{p}dx\right)^{\frac{1}{p}} < \infty\right\}.$$

(a) In view of Theorem 1, for $f \in L_{p,\varphi}(\mathbf{R})$, setting

$$H_1(y) := \int_{-\infty}^{\infty} \max_{i\in\{1,2\}} \frac{|f(x)|}{[\max\{|x^{\delta}y|^{\gamma} + (x^{\delta}y)^{\gamma}\cos\alpha_i, 1\}]^{\frac{\lambda}{\gamma}}}dx\ \ (y \in \mathbf{R}),$$

by (12), we have

$$||H_1||_{p,\psi^{1-p}} := \left(\int_{-\infty}^{\infty} \psi^{1-p}(y)H_1^{p}(y)dy\right)^{\frac{1}{p}} < K(\sigma)||f||_{p,\varphi} < \infty. \tag{28}$$

Definition 2 Let us define the Hilbert-type integral operator with the non-homogeneous kernel in the whole plane $T_1 : L_{p,\varphi}(\mathbf{R}) \to L_{p,\psi^{1-p}}(\mathbf{R})$ as follows: For any $f \in L_{p,\varphi}(\mathbf{R})$, there exists a unique representation $T_1 f = H_1 \in L_{p,\psi^{1-p}}(\mathbf{R})$, satisfying $T_1 f(y) = H_1(y)$ for any $y \in \mathbf{R}$.

In view of (28), it follows that

$$||T_1 f||_{p,\psi^{1-p}} = ||H_1||_{p,\psi^{1-p}} \leq K(\sigma)||f||_{p,\varphi},$$

and thus the operator T_1 is bounded satisfying

$$||T_1|| = \sup_{f(\neq\theta)\in L_{p,\varphi}(\mathbf{R})} \frac{||T_1 f||_{p,\psi^{1-p}}}{||f||_{p,\varphi}} \leq K(\sigma).$$

Since the constant factor $K(\sigma)$ in (28) is the best possible, we have $||T_1|| = K(\sigma)$.

If we define the formal inner product of $T_1 f$ and g as follows:

$$\begin{aligned}(T_1 f, g) \ : &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} H^{(\delta)}(x, y) f(x) dx\right) g(y) dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H^{(\delta)}(x, y) f(x) g(y) dx dy,\end{aligned}$$

then we can rewrite (11) and (12) as :

$$(T_1 f, g) < ||T_1|| \cdot ||f||_{p,\varphi} ||g||_{q,\psi}, \quad ||T_1 f||_{p,\psi^{1-p}} < ||T_1|| \cdot ||f||_{p,\varphi}.$$

(b) In view of Corollary 2, for $f \in L_{p,\phi}(\mathbf{R})$, setting

$$H_2(y) := \int_{-\infty}^{\infty} \max_{i\in\{1,2\}} \frac{|f(x)|}{(\max\{|y|^{\gamma} + sgn(x) y^{\gamma} \cos\alpha_i, |x|^{\gamma}\})^{\lambda/\gamma}} dx \ \ (y \in \mathbf{R}),$$

by (25), we have

$$||H_2||_{p,\psi^{1-p}} := \left(\int_{-\infty}^{\infty} \psi^{1-p}(y) H_2^p(y) dy\right)^{\frac{1}{p}} < K(\sigma)||f||_{p,\phi} < \infty. \tag{29}$$

Definition 3 Let us define the Hilbert-type integral operator with the homogeneous kernel in the whole plane $T_2 : L_{p,\phi}(\mathbf{R}) \to L_{p,\psi^{1-p}}(\mathbf{R})$ as follows:

For any $f \in L_{p,\phi}(\mathbf{R})$, there exists a unique representation $T_2 f = H_2 \in L_{p,\psi^{1-p}}(\mathbf{R})$, satisfying $T_2 f(y) = H_2(y)$ for any $y \in \mathbf{R}$.

In view of (29), it follows that

$$||T_2 f||_{p,\psi^{1-p}} = ||H_2||_{p,\psi^{1-p}} \leq K(\sigma)||f||_{p,\phi},$$

and then the operator T_2 is bounded satisfying

$$||T_2|| = \sup_{f(\neq\theta)\in L_{p,\phi}(\mathbf{R})} \frac{||T_2 f||_{p,\psi^{1-p}}}{||f||_{p,\phi}} \leq K(\sigma).$$

Since the constant factor $K(\sigma)$ in (29) is the best possible, we have $||T_2|| = K(\sigma)$.

If we define the formal inner product of $T_2 f$ and g as

$$(T_2 f, g) := \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \max_{i\in\{1,2\}} \frac{f(x)g(y)}{(\max\{|y|^\gamma + sgn(x)y^\gamma \cos\alpha_i, |x|^\gamma\})^{\frac{\lambda}{\gamma}}} dxdy,$$

then we can rewrite (24) and (25) as follows:

$$(T_2 f, g) < ||T_2|| \cdot ||f||_{p,\phi}||g||_{q,\psi}, \quad ||T_2 f||_{p,\psi^{1-p}} < ||T_2|| \cdot ||f||_{p,\phi}.$$

Acknowledgements This work is supported by the National Natural Science Foundation (No. 61772140), and Appropriative Researching Fund for Professors and Doctors, Guangdong University of Education (No. 2015ARF25). We are grateful for their help.

References

1. G.H. Hardy, J.E. Littlewood, G. Pólya, *Inequalities* (Cambridge University Press, Cambridge, 1934)
2. B.C. Yang, *The Norm of Operator and Hilbert-Type Inequalities* (Science Press, Beijing, 2009)
3. B.C. Yang, *Hilbert-Type Integral Inequalities* (Bentham Science Publishers Ltd., Sharjah, 2009)
4. B.C. Yang, On the norm of an integral operator and applications. J. Math. Anal. Appl. **321**, 182–192 (2006)
5. J.S. Xu, Hardy-Hilbert's inequalities with two parameters. Adv. Math. **36**(2), 63–76 (2007)
6. B.C. Yang, On the norm of a Hilbert's type linear operator and applications. J. Math. Anal. Appl. **325**, 529–541 (2007)
7. D.M. Xin, A Hilbert-type integral inequality with the homogeneous kernel of zero degree. Math. Theory Appl. **30**(2), 70–74 (2010)
8. B.C. Yang, A Hilbert-type integral inequality with the homogenous kernel of degree 0. J. Shandong Univ. **45**(2), 103–106 (2010)
9. L. Debnath, B.C. Yang, Recent developments of Hilbert-type discrete and integral inequalities with applications. Int. J. Math. Math. Sci. **2012**, 871845, 29 (2012)
10. M.Th. Rassias, B.C. Yang, On a half-discrete Hilbert's inequality. Appl. Math. Comput. **220**, 75–93 (2013)
11. B.C. Yang, M. Krnić, A half-discrete Hilbert-type inequality with a general homogeneous kernel of degree 0. J. Math. Inequal. **6**(3), 401–417 (2012)
12. Th.M. Rassias, B.C. Yang, A multidimensional half - discrete Hilbert - type inequality and the Riemann zeta function. Appl. Math. Comput. **225**, 263–277 (2013)
13. M.Th. Rassias, B.C. Yang, On a multidimensional half - discrete Hilbert - type inequality related to the hyperbolic cotangent function. Appl. Math. Comput. **242**, 800–813 (2013)
14. M.Th. Rassias, B.C. Yang, A multidimensional Hilbert - type integral inequality related to the Riemann zeta function, in *Applications of Mathematics and Informatics in Science and Engineering*, ed. by N.J. Daras (Springer, New York, 2014), pp. 417–433

15. Q. Chen, B.C. Yang, A survey on the study of Hilbert-type inequalities. J. Inequal. Appl. **2015**, 302 (2015)
16. B.C. Yang, A new Hilbert-type integral inequality. Soochow J. Math. **33**(4), 849–859 (2007)
17. Z.Q. Wang, D.R. Guo, *Introduction to Special Functions* (Science Press, Beijing, 1979)
18. B. He, B.C. Yang, On a Hilbert-type integral inequality with the homogeneous kernel of 0-degree and the hypergeometrc function. Math. Pract. Theory **40**(18), 105–211 (2010)
19. B.C. Yang, A new Hilbert-type integral inequality with some parameters. J. Jilin Univ. **46**(6), 1085–1090 (2008)
20. B.C. Yang, A Hilbert-type integral inequality with a non-homogeneous kernel. J. Xiamen Univ. **48**(2), 165–169 (2008)
21. Z. Zeng, Z.T. Xie, On a new Hilbert-type integral inequality with the homogeneous kernel of degree 0 and the integral in whole plane. J. Inequal. Appl. **2010**, 256796, 9 (2010)
22. B.C. Yang, A reverse Hilbert-type integral inequality with some parameters. J. Xinxiang Univ. **27**(6), 1–4 (2010)
23. A.Z. Wang, B.C. Yang, A new Hilbert-type integral inequality in whole plane with the non-homogeneous kernel. J. Inequal. Appl. **2011**, 123 (2011)
24. D.M. Xin, B.C. Yang, A Hilbert-type integral inequality in whole plane with the homogeneous kernel of degree -2. J. Inequal. Appl. **2011**, 401428, 11 (2011)
25. B. He, B.C. Yang, On an inequality concerning a non-homogeneous kernel and the hypergeometric function. Tamsul Oxford J. Inf. Math. Sci. **27**(1), 75–88 (2011)
26. Z.T. Xie, Z. Zeng, Y.F. Sun, A new Hilbert-type inequality with the homogeneous kernel of degree -2. Adv. Appl. Math. Sci. **12**(7), 391–401 (2013)
27. Q.L. Huang, S.H. Wu, B.C. Yang, Parameterized Hilbert-type integral inequalities in the whole plane. Sci. World J. **2014**, 169061, 8 (2014)
28. Z. Zhen, K. Raja Rama Gandhi, Z.T. Xie, A new Hilbert-type inequality with the homogeneous kernel of degree -2 and with the integral. Bull. Math. Sci. Appl. **3**(1), 11–20 (2014)
29. M.Th. Rassias, B.C. Yang, A Hilbert - type integral inequality in the whole plane related to the hyper geometric function and the beta function. J. Math. Anal. Appl. **428**(2), 1286–1308 (2015)
30. X.Y. Huang, J.F.Cao, B. He, B.C. Yang, Hilbert-type and Hardy-type integral inequalities with operator expressions and the best constants in the whole plane. J. Inequal. Appl. **2015**, 129 (2015)
31. Z.H. Gu, B.C. Yang, A Hilbert-type integral inequality in the whole plane with a non-homogeneous kernel and a few parameters. J. Inequal. Appl. **2015**, 314 (2015)
32. J.C. Kuang, *Applied Inequalities* (Shangdong Science and Technology Press, Jinan, 2004)
33. J.C. Kuang, *Real Analysis and Functional Analysis (Continuation) (Second Volume)* (Higher Education Press, Beijing, 2015)

Four Conjectures in Nonlinear Analysis

Biagio Ricceri

In this chapter, I intend to formulate four challenging conjectures in Nonlinear Analysis which have their roots in certain results that I have obtained in the past years.

1 A Conjecture on the Monge-Ampère Equation

Conjecture 1.1 Let $\Omega \subset \mathbf{R}^n$ $(n \geq 2)$ be a non-empty open bounded set and let $h : \Omega \to \mathbf{R}$ be a non-negative continuous function.

Then, each $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$ satisfying in Ω the Monge-Ampère equation

$$\det(D^2 u) = h$$

has the following property:

$$\nabla(\Omega) \subseteq \operatorname{conv}(\nabla(\partial\Omega)) \; .$$

This conjecture is motivated by Ricceri [26] where I proved that it is true for $n = 2$. I am going to produce such a proof here.

In what follows, Ω is a non-empty relatively compact and open set in a topological space E, with $\partial\Omega \neq \emptyset$, and Y is a real locally convex Hausdorff topological vector space. $\overline{\Omega}$ and $\partial\Omega$ denote the closure and the boundary of Ω, respectively. Since $\overline{\Omega}$ is compact, $\partial\Omega$, being closed, is compact too. Let us first recall some well-known definitions.

B. Ricceri (✉)
Department of Mathematics, University of Catania, Catania, Italy
e-mail: ricceri@dmi.unict.it

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization
and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_24

Let S be a subset of Y and let $y_0 \in S$. As usual, we say that S is supported at y_0 if there exists $\varphi \in Y^* \setminus \{0\}$ such that $\varphi(y_0) \le \varphi(y)$ for all $y \in S$. If this happens, of course $y_0 \in \partial S$. Further, extending a maximum principle definition for real-valued functions, a continuous function $f : \overline{\Omega} \to Y$ is said to satisfy the convex hull property in $\overline{\Omega}$ (see [7, 13] and references therein) if

$$f(\Omega) \subseteq \overline{\mathrm{conv}}(f(\partial\Omega)) ,$$

$\overline{\mathrm{conv}}(f(\partial\Omega))$ being the closed convex hull of $f(\partial\Omega)$. When $\dim(Y) < \infty$, since $f(\partial\Omega)$ is compact, $\mathrm{conv}(f(\partial\Omega))$ is compact too and so $\overline{\mathrm{conv}}(f(\partial\Omega)) = \mathrm{conv}(f(\partial\Omega))$.

A function $\psi : Y \to \mathbf{R}$ is said to be quasi-convex if, for each $r \in \mathbf{R}$, the set $\psi^{-1}(]-\infty, r])$ is convex.

Notice the following proposition:

Proposition 1.1 *For each pair A, B of non-empty subsets of Y, the following assertions are equivalent:*

(a_1) $A \subseteq \overline{\mathrm{conv}}(B)$.

(a_2) *For every continuous and quasi-convex function $\psi : Y \to \mathbf{R}$, one has*

$$\sup_A \psi \le \sup_B \psi .$$

Proof Let (a_1) hold. Fix any continuous and quasi-convex function $\psi : Y \to \mathbf{R}$. Fix $\tilde{y} \in A$. Then, there is a net $\{y_\alpha\}$ in $\mathrm{conv}(B)$ converging to $\tilde{y}$. So, for each α, we have $y_\alpha = \sum_{i=1}^k \lambda_i z_i$, where $z_i \in B$, $\lambda_i \in [0, 1]$ and $\sum_{i=1}^k \lambda_i = 1$. By quasi-convexity, we have

$$\psi(y_\alpha) = \psi\left(\sum_{i=1}^k \lambda_i z_i\right) \le \max_{1\le i\le k} \psi(z_i) \le \sup_B \psi$$

and so, by continuity,

$$\psi(\tilde{y}) = \lim_\alpha \psi(y_\alpha) \le \sup_B \psi$$

which yields (a_2).

Now, let (a_2) hold. Let $x_0 \in A$. If $x_0 \notin \overline{\mathrm{conv}}(B)$, by the standard separation theorem, there would be $\psi \in Y^* \setminus \{0\}$ such that $\sup_{\overline{\mathrm{conv}}(B)} \psi < \psi(x_0)$, against (a_2). So, (a_1) holds. △

Clearly, applying Proposition 1.1, we obtain the following one:

Proposition 1.2 *For any continuous function $f : \overline{\Omega} \to Y$, the following assertions are equivalent:*

(b_1) *f satisfies the convex hull property in $\overline{\Omega}$.*

(b_2) *For every continuous and quasi-convex function* $\psi : Y \to \mathbf{R}$, *one has*

$$\sup_{x\in\Omega} \psi(f(x)) = \sup_{x\in\partial\Omega} \psi(f(x)) .$$

In view of Proposition 1.2, we now introduce the notion of convex hull-like property for functions defined in Ω only.

Definition 1.1 A continuous function $f : \Omega \to Y$ is said to satisfy the convex hull-like property in Ω if, for every continuous and quasi-convex function $\psi : Y \to \mathbf{R}$, there exists $x^* \in \partial\Omega$ such that

$$\limsup_{x\to x^*} \psi(f(x)) = \sup_{x\in\Omega} \psi(f(x)) .$$

We have

Proposition 1.3 *Let* $g : \overline{\Omega} \to Y$ *be a continuous function and let* $f = g_{|\Omega}$.

Then, the following assertions are equivalent:

(c_1) f *satisfies the convex hull-like property in* Ω .
(c_2) g *satisfies the convex hull property in* $\overline{\Omega}$.

Proof Let (c_1) hold. Let $\psi : Y \to \mathbf{R}$ be any continuous and quasi-convex function. Then, by Definition 1.1, there exists $x^* \in \partial\Omega$ such that

$$\limsup_{x\to x^*} \psi(f(x)) = \sup_{x\in\Omega} \psi(f(x)) .$$

But

$$\limsup_{x\to x^*} \psi(f(x)) = \psi(g(x^*))$$

and hence

$$\sup_{x\in\partial\Omega} \psi(g(x)) = \sup_{x\in\Omega} \psi(g(x)) .$$

So, by Proposition 1.2, (c_2) holds.

Now, let (c_2) hold. Let $\psi : Y \to \mathbf{R}$ be any continuous and quasi-convex function. Then, by Proposition 1.2, one has

$$\sup_{x\in\partial\Omega} \psi(g(x)) = \sup_{x\in\Omega} \psi(g(x)) .$$

Since $\partial\Omega$ is compact and $\psi \circ g$ is continuous, there exists $x^* \in \partial\Omega$ such that

$$\psi(g(x^*)) = \sup_{x\in\partial\Omega} \psi(g(x)) .$$

But

$$\psi(g(x^*)) = \lim_{x \to x^*} \psi(f(x))$$

and, by continuity again,

$$\sup_{x \in \Omega} \psi(g(x)) = \sup_{x \in \overline{\Omega}} \psi(g(x))$$

and so

$$\lim_{x \to x^*} \psi(f(x)) = \sup_{x \in \Omega} \psi(f(x))$$

which yields (c_1). △

The central result is as follows:

Theorem 1.1 *For any continuous function $f : \Omega \to Y$, at least one of the following assertions holds:*

- (i) *f satisfies the convex hull-like property in Ω .*
- (ii) *There exists a non-empty open set $X \subseteq \Omega$, with $\overline{X} \subseteq \Omega$, satisfying the following property: for every continuous function $g : \Omega \to Y$, there exists $\tilde{\lambda} \geq 0$ such that, for each $\lambda > \tilde{\lambda}$, the set $(g + \lambda f)(X)$ is supported at one of its points.*

Proof Assume that (i) does not hold. So, we are assuming that there exists a continuous and quasi-convex function $\psi : Y \to \mathbf{R}$ such that

$$\limsup_{x \to z} \psi(f(x)) < \sup_{x \in \Omega} \psi(f(x)) \tag{1.1}$$

for all $z \in \partial\Omega$.

In view of (1.1), for each $z \in \partial\Omega$, there exists an open neighbourhood U_z of z such that

$$\sup_{x \in U_z \cap \Omega} \psi(f(x)) < \sup_{x \in \Omega} \psi(f(x)) .$$

Since $\partial\Omega$ is compact, there are finitely many $z_1, \ldots, z_k \in \partial\Omega$ such that

$$\partial\Omega \subseteq \bigcup_{i=1}^{k} U_{z_i} . \tag{1.2}$$

Put

$$U = \bigcup_{i=1}^{k} U_{z_i} .$$

Hence

$$\sup_{x\in U\cap\Omega}\psi(f(x))=\max_{1\leq i\leq k}\sup_{x\in U_{z_i}\cap\Omega}\psi(f(x))<\sup_{x\in\Omega}\psi(f(x))\ .$$

Now, fix a number r so that

$$\sup_{x\in U\cap\Omega}\psi(f(x))<r<\sup_{x\in\Omega}\psi(f(x)) \tag{1.3}$$

and set

$$K=\{x\in\Omega:\psi(f(x))\geq r\}\ .$$

Since f, ψ are continuous, K is closed in Ω. But, since $K\cap U=\emptyset$ and U is open, in view of (1.2), K is closed in E. Hence, K is compact since $\overline{\Omega}$ is so. By (1.3), we can fix $\bar{x}\in\Omega$ such that $\psi(f(\bar{x}))>r$. Notice that the set $\psi^{-1}(]-\infty,r])$ is closed and convex. So, thanks to the standard separation theorem, there exists a non-zero continuous linear functional $\varphi:Y\to\mathbf{R}$ such that

$$\varphi(f(\bar{x}))<\inf_{y\in\psi^{-1}(]-\infty,r])}\varphi(y)\ . \tag{1.4}$$

Then, from (1.4), it follows

$$\varphi(f(\bar{x}))<\inf_{x\in\Omega\setminus K}\varphi(f(x))\ .$$

Now, choose ρ so that

$$\varphi(f(\bar{x}))<\rho<\inf_{x\in\Omega\setminus K}\varphi(f(x))$$

and set

$$X=\{x\in\Omega:\varphi(f(x))<\rho\}\ .$$

Clearly, X is a non-empty open set contained in K. Now, let $g:\Omega\to Y$ be any continuous function. Set

$$\tilde{\lambda}=\inf_{x\in X}\frac{\varphi(g(x))-\inf_{z\in K}\varphi(g(z))}{\rho-\varphi(f(x))}\ .$$

Fix $\lambda>\tilde{\lambda}$. So, there is $x_0\in X$ such that

$$\frac{\varphi(g(x_0))-\inf_{z\in K}\varphi(g(z))}{\rho-\varphi(f(x_0))}<\lambda\ .$$

From this, we get

$$\varphi(g(x_0)) + \lambda\varphi(f(x_0)) < \lambda\rho + \inf_{z\in K} \varphi(g(z)) \ . \tag{1.5}$$

By continuity and compactness, there exists $\hat{x} \in K$ such that

$$\varphi(g(\hat{x}) + \lambda f(\hat{x})) \leq \varphi(g(x)) + \lambda f(x)) \tag{1.6}$$

for all $x \in K$. Let us prove that $\hat{x} \in X$. Arguing by contradiction, assume that $\varphi(f(\hat{x})) \geq \rho$. Then, taking (1.5) into account, we would have

$$\varphi(g(x_0)) + \lambda\varphi(f(x_0)) < \lambda\varphi(f(\hat{x})) + \varphi(g(\hat{x}))$$

contradicting (1.6). So, it is true that $\hat{x} \in X$, and, by (1.6), the set $(g + \lambda f)(X)$ is supported at its point $g(\hat{x}) + \lambda f(\hat{x})$. △

An application of Theorem 1.1 shows a strongly bifurcating behaviour of certain equations in $\mathbf{R}^n$.

Theorem 1.2 *Let Ω be a non-empty bounded open subset of $\mathbf{R}^n$ and let $f : \Omega \to \mathbf{R}^n$ a continuous function.*

Then, at least one of the following assertions holds:

(d_1) *f satisfies the convex hull-like property in Ω .*

(d_2) *There exists a non-empty open set $X \subseteq \Omega$, with $\overline{X} \subseteq \Omega$, satisfying the following property: for every continuous function $g : \Omega \to \mathbf{R}^n$, there exists $\tilde{\lambda} \geq 0$ such that, for each $\lambda > \tilde{\lambda}$, there exist $\hat{x} \in X$ and two sequences $\{y_k\}$, $\{z_k\}$ in $\mathbf{R}^n$, with*

$$\lim_{k\to\infty} y_k = \lim_{k\to\infty} z_k = g(\hat{x}) + \lambda f(\hat{x}) \ ,$$

such that, for each $k \in \mathbf{N}$, one has

(j) *the equation*

$$g(x) + \lambda f(x) = y_k$$

has no solution in X ;

(jj) *the equation*

$$g(x) + \lambda f(x) = z_k$$

has two distinct solutions u_k, v_k in X such that

$$\lim_{k\to\infty} u_k = \lim_{k\to\infty} v_k = \hat{x} \ .$$

Proof Apply Theorem 1.1 with $E = Y = \mathbf{R}^n$. Assume that (d_1) does not hold. Let $X \subseteq \Omega$ be an open set as in (ii) of Theorem 1.1. Fix any continuous function $g : \Omega \to \mathbf{R}^n$. Then, there is some $\tilde{\lambda} \geq 0$ such that, for each $\lambda > \tilde{\lambda}$, there exists $\hat{x} \in X$ such that the set $(g + \lambda f)(X)$ is supported at $g(\hat{x}) + \lambda f(\hat{x})$. As we observed at the beginning, this implies that $g(\hat{x}) + \lambda f(\hat{x})$ lies in the boundary of $(g + \lambda f)(X)$. Therefore, we can find a sequence $\{y_k\}$ in $\mathbf{R}^n \setminus (g + \lambda f)(X)$ converging to $g(\hat{x}) + \lambda f(\hat{x})$. So, such a sequence satisfies (j). For each $k \in \mathbf{N}$, denote by B_k the open ball of radius $\frac{1}{k}$ centered at $\hat{x}$. Let k be such that $B_k \subseteq X$. The set $(g + \lambda f)(B_k)$ is not open since its boundary contains the point $g(\hat{x}) + \lambda f(\hat{x})$. Consequently, by the invariance of domain theorem [30, p. 705], the function $g + \lambda f$ is not injective in B_k. So, there are $u_k, v_k \in B_k$, with $u_k \neq v_k$, such that

$$g(u_k) + \lambda f(u_k) = g(v_k) + \lambda f(v_k) .$$

Hence, if we take

$$z_k = g(u_k) + \lambda f(u_k) ,$$

the sequences $\{u_k\}$, $\{v_k\}$, $\{z_k\}$ satisfy (jj) and the proof is complete. △

Remark 1.1 Notice that, in general, Theorem 1.2 is no longer true when $f : \Omega \to \mathbf{R}^m$ with $m > n$. In this connection, consider the case $n = 1$, $m = 2$, $\Omega =]0, \pi[$ and $f(\theta) = (\cos\theta, \sin\theta)$ for $\theta \in [0, \pi]$. So, for each $\lambda > 0$, on the one hand, the function λf is injective, while, on the other hand, $f(]0, \pi[)$ is not contained in $\mathrm{conv}(\{f(0), f(\pi)\})$.

If $S \subseteq \mathbf{R}^n$ is a non-empty open set, $x \in S$ and $h : S \to \mathbf{R}^n$ is a C^1 function, we denote by $\det(J_h(x))$ the Jacobian determinant of h at x.

A very recent and important result by Saint Raymond [27] states what follows (for anything concerning the topological dimension we refer to [8]):

Theorem 1.A ([27, Theorem 10]) *Let $A \subseteq \mathbf{R}^n$ be a non-empty open set and $\varphi : A \to \mathbf{R}^n$ a C^1 function such that the topological dimension of the set*

$$\{x \in A : \det(J_\varphi(x)) = 0\}$$

is not positive.

Then, the function φ is open.

A joint application of Theorems 1.1 and 1.A gives

Theorem 1.3 *Let Ω be a non-empty bounded open subset of $\mathbf{R}^n$ and let $f : \Omega \to \mathbf{R}^n$ be a C^1 function.*

Then, at least one of the following assertions holds:

(a_1) *f satisfies the convex hull-like property in Ω .*

(a_2) *There exists a non-empty open set $X \subseteq \Omega$, with $\overline{X} \subseteq \Omega$, satisfying the following property: for every continuous function $g : \Omega \to \mathbf{R}^n$ which is C^1 in X, there exists $\tilde{\lambda} \geq 0$ such that, for each $\lambda > \tilde{\lambda}$, the topological dimension of* *the set*

$$\{x \in X : \det(J_{g+\lambda f}(x)) = 0\}$$

is greater than or equal 1.

Proof Assume that (a_1) does not hold. Let X be an open set as in (ii) of Theorem 1.1. Let $g : \Omega \to \mathbf{R}^n$ be a continuous function which is C^1 in X. Then, there is some $\tilde{\lambda} \geq 0$ such that, for each $\lambda > \tilde{\lambda}$, there exists $\hat{x} \in X$ such that the set $(g + \lambda f)(X)$ is supported at $g(\hat{x}) + \lambda f(\hat{x})$. As already remarked, this implies that $g(\hat{x}) + \lambda f(\hat{x}) \in \partial(g + \lambda f)(X)$ and so $(g + \lambda f)(X)$ is not open. Now, (a_2) is a direct consequence of Theorem 1.A. △

In turn, here is a consequence of Theorem 1.3 when $n = 2$.

Theorem 1.4 *Let Ω be a non-empty bounded open set of $\mathbf{R}^2$, let $h : \Omega \to \mathbf{R}$ be a continuous function and let $\alpha, \beta : \Omega \to \mathbf{R}$ be two C^1 functions such that $|\alpha_x \beta_y - \alpha_y \beta_x| + |h| > 0$ and $(\alpha_x \beta_y - \alpha_y \beta_x)h \geq 0$ in Ω.*

Then, any C^1 solution (u, v) in Ω of the system

$$\begin{cases} u_x v_y - u_y v_x = h \\ \\ \beta_y u_x - \beta_x u_y - \alpha_y v_x + \alpha_x v_y = 0 \end{cases} \tag{1.7}$$

satisfies the convex hull-like property in Ω.

Proof Arguing by contradiction, assume that (u, v) does not satisfy the convex hull-like property in Ω. Then, by Theorem 1.3, applied taking $f = (u, v)$ and $g = (\alpha, \beta)$, there exist $\lambda > 0$ and $(\hat{x}, \hat{y}) \in \Omega$ such that

$$\det(J_{g+\lambda f}(\hat{x}, \hat{y})) = 0 \, .$$

On the other hand, for each $(x, y) \in \Omega$, we have

$$\det(J_{g+\lambda f}(x, y)) = (u_x v_y - u_y v_x)(x, y)\lambda^2 + (\beta_y u_x - \beta_x u_y - \alpha_y v_x + \alpha_x v_y)(x, y)\lambda \\ + (\alpha_x \beta_y - \alpha_y \beta_x)(x, y)$$

and hence

$$h(\hat{x}, \hat{y})\lambda^2 + (\alpha_x \beta_y - \alpha_y \beta_x)(\hat{x}, \hat{y}) = 0$$

which is impossible in view of our assumptions. △

Finally, taking Proposition 1.3 in mind, here is the proof of Conjecture 1.1 when $n = 2$:

Theorem 1.5 *Let Ω be a non-empty bounded open subset of* $\mathbf{R}^2$*, let $h : \Omega \to \mathbf{R}$ be a continuous non-negative function and let $w \in C^2(\Omega)$ be a function satisfying in Ω the Monge-Ampère equation*

$$w_{xx}w_{yy} - w_{xy}^2 = h \ .$$

Then, the gradient of w satisfies the convex hull-like property in Ω.

Proof It is enough to observe that (w_x, w_y) is a C^1 solution in Ω of the system (1.7) with $\alpha(x, y) = -y$ and $\beta(x, y) = x$ and that such α, β satisfy the assumptions of Theorem 1.4. △

2 A Conjecture on an Eigenvalue Problem

Conjecture 2.1 Let $n \geq 2$ and let $\Omega = \{x \in \mathbf{R}^n : a < |x| < b\}$, with $0 < a < b$.

Then, there exists $\lambda > 0$ such that the problem

$$\begin{cases} \Delta u = \lambda \sin u & \textit{in } \Omega \\ \\ u = 0 & \textit{on } \partial\Omega \end{cases}$$

has at least one non-zero classical solution.

The above conjecture has its roots in Pohozaev identity [19]. Let me recall it.

So, let $\Omega \subset \mathbf{R}^n$ be a smooth bounded domain, and let $f : \mathbf{R} \to \mathbf{R}$ be a continuous function. Put

$$F(\xi) = \int_0^\xi f(t)dt$$

for all $\xi \in \mathbf{R}$. For $\lambda > 0$, consider the problem

$$\begin{cases} -\Delta u = \lambda f(u) & \textit{in } \Omega \\ \\ u = 0 & \textit{on } \partial\Omega \ . \end{cases} \qquad (P_{\lambda f})$$

In the sequel, a classical solution of problem $(P_{\lambda f})$ is any $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$, zero on $\partial\Omega$, satisfying the equation pointwise in Ω. Set

$$\Lambda_f = \{\lambda > 0 : (P_{\lambda f}) \text{ has a non-zero classical solution}\} \ .$$

When $n \geq 2$, the Pohozaev identity tells us that, if u is a classical solution of $(P_{\lambda f})$, then one has

$$\frac{2-n}{2}\int_{\Omega}|\nabla u(x)|^2dx+n\lambda\int_{\Omega}F(u(x))dx=\frac{1}{2}\int_{\partial\Omega}|\nabla u(x)|^2x\cdot\nu(x)ds \qquad (2.1)$$

where ν denotes the unit outward normal to $\partial\Omega$.

From (2.1), in particular, it follows that, if Ω is star-shaped with respect to 0 (so $x\cdot\nu(x)\geq 0$ on $\partial\Omega$), then the set Λ_f is empty in the two following cases:

(a) $f(\xi)=|\xi|^{p-2}\xi$ with $n\geq 3$ and $p\geq\frac{2n}{n-2}$;
(b) $\sup_{\xi\in\mathbf{R}}F(\xi)=0$.

A natural question arises: what about problem $(P_{\lambda f})$ in cases (a) and (b) when Ω is not star-shaped?

It is very surprising to realize that, while a great amount of research has been produced on case (a) (see, for instance, [1–5, 12, 14, 17, 18]), apparently the only papers dealing with case (b) are [9–11, 23].

In [11], the following result has been pointed out:

Theorem 2.1 *Let $n\geq 2$ and $\Omega=\{x\in\mathbf{R}^n : a<|x|<b\}$ with $0<a<b$.*

Then, for every continuous function $f:\mathbf{R}\to\mathbf{R}$, with $\sup_{\xi\in\mathbf{R}}F(\xi)=0$, and every $\lambda>0$, problem $(P_{\lambda f})$ has no radially symmetric non-zero classical solutions.

Proof Let $f:\mathbf{R}\to\mathbf{R}$ be a continuous function, with $\sup_{\xi\in\mathbf{R}}F(\xi)=0$, let $\lambda>0$, and let u be a radially symmetric classical solution of $(P_{\lambda f})$. Then

$$\begin{cases}-(r^{n-1}u'(r))'=\lambda r^{n-1}f(u(r)) & in\ ]a,b[\\ \\ u(a)=u(b)=0\ ,\end{cases}$$

that is

$$\begin{cases}u''(r)+\frac{n-1}{r}u'(r)+\lambda f(u(r))=0 & in\ ]a,b[\\ \\ u(a)=u(b)=0\ .\end{cases} \qquad (2.2)$$

Multiplying both sides of the equation in (2.2) by u', we have

$$u''(r)u'(r)+\frac{n-1}{r}(u'(r))^2+\lambda f(u(r))u'(r)=0 \qquad (2.3)$$

for all $r\in(a,b)$. Let $r_1\in(a,b)$ be such that $u'(r_1)=0$. Define

$$I_{r_1}(r)=\frac{1}{2}\left|u'(r)\right|^2+(n-1)\int_{r_1}^{r}\frac{(u'(t))^2}{t}dt+\lambda F(u(r))$$

for all $r \in [a, b]$. Then (2.3) shows that $I'_{r_1}(r) = 0$ for all $r \in]a, b[$ and so, for some $c \in \mathbf{R}$, one has

$$I_{r_1}(r) = c$$

for all $r \in [a, b]$. Since

$$I_{r_1}(r_1) = 0 + 0 + \lambda F(u(r_1)) \leq 0 \, ,$$

we have $c \leq 0$. On the other hand, since

$$I_{r_1}(b) = \frac{1}{2}|u'(b)|^2 + (n-1)\int_{r_1}^{b} \frac{(u'(t))^2}{t} dt + 0 \geq 0 \, ,$$

have $c \geq 0$, and so $c = 0$. In particular $I_{r_1}(b) = 0$, which implies $u'(b) = 0$, and consequently $u(r) = 0$ for all $r \in [a, b]$, as claimed. △

Remark 2.1 It is important to note the drastic difference between cases (a) and (b) enlighted by Theorem 2.1 when Ω is an annulus. Actually, in this case, it was remarked in [14] that the problem

$$\begin{cases} -\Delta u = \lambda |u|^{p-2}u & in \ \Omega \\ \\ u = 0 & on \ \partial\Omega \end{cases}$$

has radially symmetric non-zero classical solutions for $p \geq \frac{2n}{n-2}$ $(n \geq 3)$, and $\lambda > 0$.

Now, I recall a very general result proved in [23].

For any real Hilbert space X, denote by $\mathscr{A}_X$ the set of all C^1 functionals $I : X \to \mathbf{R}$ such that 0 is a global maximum of I and I' is Lipschitzian with Lipschitz constant less than 1. Set

$$\gamma_X = \inf_{I \in \mathscr{A}_X} \inf\{\lambda > 0 : x = \lambda I'(x) \ \text{ for some } \ x \neq 0\} \, .$$

We have:

Theorem 2.2 *For any real Hilbert space $(X, \langle \cdot, \cdot \rangle)$, with $X \neq \{0\}$, one has*

$$\gamma_X = 3 \, .$$

We first prove

Proposition 2.1 *One has*

$$\gamma_{\mathbf{R}} = 3 \, .$$

Proof Let $I_0 \in \mathscr{A}_{\mathbf{R}}$ and let $L < 1$ be the Lipschitz constant of I_0'. Set

$$I = I_0 - I_0(0)\ .$$

Fix $\lambda \in]0, 3]$. Let us prove that 0 is the only solution of the equation

$$x = \lambda I'(x)\ .$$

Arguing by contradiction, assume that

$$x_0 = \lambda I'(x_0)$$

for some $x_0 \neq 0$. It is not restrictive to assume that $x_0 > 0$ (otherwise, we would work with $I'(-x)$). Consider now the function $g : \mathbf{R} \to \mathbf{R}$ defined by

$$g(x) = \begin{cases} -\frac{x^2}{2} & if\ x < \frac{x_0}{3} \\ \frac{x^2}{2} - \frac{2x_0x}{3} + \frac{x_0^2}{9} & if\ \frac{x_0}{3} \leq x \leq x_0 \\ -\frac{x^2}{2} + \frac{4x_0x}{3} - \frac{8x_0^2}{9} & if\ x_0 > x\ . \end{cases}$$

Clearly, $g \in C^1(\mathbf{R})$. Let $x > 0$ with $x \neq x_0$. Let us prove that

$$g'(x) < I'(x)\ .$$

We distinguish two cases. If $0 < x \leq \frac{x_0}{3}$, We have

$$g'(x) = -x < -Lx \leq I'(x)\ .$$

If $x > \frac{x_0}{3}$, We have

$$\begin{aligned} g'(x) &= \frac{x_0}{3} - |x - x_0| < \frac{x_0}{3} - L|x - x_0| = \frac{\lambda I'(x_0)}{3} - L|x - x_0| \\ &\leq I'(x_0) - L|x - x_0| \leq I'(x)\ . \end{aligned}$$

So, in particular, we get

$$I\left(\frac{4x_0}{3}\right) = \int_0^{\frac{4x_0}{3}} I'(x)dx > \int_0^{\frac{4x_0}{3}} g'(x)dx = g\left(\frac{4x_0}{3}\right) = 0$$

which contradicts the fact that the function I is non-positive, since 0 is a global maximum of I_0. From what we have just proven, it clearly follows that

$$3 \leq \gamma_{\mathbf{R}} .$$

Now, fix any $\mu > 1$. Continue to consider the function g defined above (for a fixed $x_0 > 0$). Clearly, the function $\frac{1}{\mu}g$ belongs to $\mathscr{A}_{\mathbf{R}}$ and

$$x_0 = 3\mu \frac{g(x_0)}{\mu} .$$

Of course, from this we infer that

$$\gamma_{\mathbf{R}} \leq 3\mu$$

and the conclusion clearly follows. △

Proof of Theorem 2.2 First, let us prove that

$$\gamma_X \leq 3 . \tag{2.4}$$

To this end, fix any $\varphi \in \mathscr{A}_{\mathbf{R}}$ and any $\lambda > 0$ such that

$$t = \lambda\varphi'(t)$$

for some $t \neq 0$. Fix also $u \in X$, with $\|u\| = 1$, and consider the functional I defined by putting

$$I(x) = \varphi(\langle u, x \rangle)$$

for all $x \in X$. It is readily seen that $I \in \mathscr{A}_X$. In particular, note that

$$I'(x) = \varphi'(\langle u, x \rangle)u .$$

Finally, set

$$\hat{x} = \lambda\varphi(t)u .$$

Of course, $\hat{x} \neq 0$. Since

$$\langle u, \hat{x} \rangle = \lambda\varphi'(t)$$

we also have

$$\langle u, \hat{x} \rangle = t$$

and so

$$\hat{x} = \lambda I'(\hat{x}) .$$

From this, it clearly follows that

$$\gamma_X \leq \gamma_{\mathbf{R}}$$

and so (2.4) follows now from Proposition 2.1.

Now, let us prove that

$$3 \leq \gamma_X \ . \tag{2.5}$$

To this end, fix $I \in \mathscr{A}_X$, $\lambda > 0$ and $x \in X \setminus \{0\}$ such that

$$x = \lambda I'(x) \ . \tag{2.6}$$

Then, consider the function $\varphi : \mathbf{R} \to \mathbf{R}$ defined by

$$\varphi(t) = I\left(\frac{tx}{\|x\|}\right)$$

for all $t \in \mathbf{R}$. Clearly, 0 is a global maximum for φ. Moreover, $\varphi \in C^1(\mathbf{R})$ and one has

$$\varphi'(t) = \left\langle I'\left(\frac{tx}{\|x\|}\right), \frac{x}{\|x\|}\right\rangle \ .$$

Therefore, if L is the Lipschitz constant of I', for each $t, s \in \mathbf{R}$, we have

$$|\varphi'(t) - \varphi'(s)| = \left|\left\langle I'\left(\frac{tx}{\|x\|}\right) - I'\left(\frac{sx}{\|x\|}\right), \frac{x}{\|x\|}\right\rangle\right|$$

$$\leq \left\| I'\left(\frac{tx}{\|x\|}\right) - I'\left(\frac{sx}{\|x\|}\right)\right\| \leq L|t-s| \ .$$

This shows that φ' is a contraction, and so $\varphi \in \mathscr{A}_{\mathbf{R}}$. Now, from (2.6), we get

$$\|x\| = \lambda \left\langle I'(x), \frac{x}{\|x\|}\right\rangle$$

that is

$$\|x\| = \lambda\varphi'(\|x\|) \ .$$

From this, we infer that

$$\gamma_{\mathbf{R}} \leq \gamma_X \ .$$

So (2.5) follows from Proposition 2.1, and the proof is complete. △

Now, for each $L > 0$, denote by $\mathscr{C}_L$ the class of all Lipschitzian functions $f : \mathbf{R} \to \mathbf{R}$, with Lipschitz constant L, such that $f(0) = 0$ and $\sup_{\xi \in \mathbf{R}} F(\xi) = 0$. Also denote by λ_1 the first eigenvalue of the problem

$$\begin{cases} -\Delta u = \lambda u & in\ \Omega \\ \\ u = 0 & on\ \partial\Omega\ . \end{cases}$$

From Theorem 2.2, it directly follows that

$$\inf_{f \in \mathscr{C}_L} \inf \Lambda_f \geq \frac{3\lambda_1}{L}\ .$$

In [10], Fan obtained the finer inequality

$$\inf_{f \in \mathscr{C}_L} \inf \Lambda_f > \frac{3\lambda_1}{L}\ .$$

Conjecture 2.1 says that $\Lambda_f \neq \emptyset$ for $f(\xi) = -\sin\xi$, Ω being an annulus. Due to what precedes, if Conjecture 2.1 is true, then λ must necessarily be larger than $3\lambda_1$.

3 A Conjecture on a Non-local Problem

Conjecture 3.1 Let $a \geq 0$, $b > 0$ and let $\Omega \subset \mathbf{R}^n$ be a smooth bounded domain, with $n > 4$.

Then, for each $\lambda > 0$ large enough and for each convex set $C \subseteq L^2(\Omega)$ whose closure in $L^2(\Omega)$ contains $H_0^1(\Omega)$, there exists $v^* \in C$ such that the problem

$$\begin{cases} -\left(a + b\int_\Omega |\nabla u(x)|^2 dx\right) \Delta u = |u|^{\frac{4}{n-2}} u + \lambda(u - v^*(x)) & in\ \Omega \\ \\ u = 0 & on\ \partial\Omega \end{cases}$$

has at least three weak solutions, two of which are global minima in $H_0^1(\Omega)$ of the functional

$$u \to \frac{a}{2}\int_\Omega |\nabla u(x)|^2 dx + \frac{b}{4}\left(\int_\Omega |\nabla u(x)|^2 dx\right)^2 - \frac{n-2}{2n}\int_\Omega |u(x)|^{\frac{2n}{n-2}} dx - \frac{\lambda}{2}\int_\Omega |u(x) - v^*(x)|^2 dx\ .$$

Conjecture 3.1 comes from the results I have obtained in [25]. I am going to reproduce them here.

Let a, b, Ω be as in Conjecture 3.1.

On the Sobolev space $H_0^1(\Omega)$, we consider the scalar product

$$\langle u, v\rangle = \int_\Omega \nabla u(x)\nabla v(x)dx$$

and the induced norm

$$\|u\| = \left(\int_\Omega |\nabla u(x)|^2 dx\right)^{\frac{1}{2}} .$$

Denote by $\mathscr{A}$ the class of all Carathéodory functions $f : \Omega \times \mathbf{R} \to \mathbf{R}$ such that

$$\sup_{(x,\xi)\in\Omega\times\mathbf{R}} \frac{|f(x,\xi)|}{1+|\xi|^p} < +\infty \tag{3.1}$$

for some $p \in \left]0, \frac{n+2}{n-2}\right[$.

Moreover, denote by $\tilde{\mathscr{A}}$ the class of all Carathéodory functions $g : \Omega \times \mathbf{R} \to \mathbf{R}$ such that

$$\sup_{(x,\xi)\in\Omega\times\mathbf{R}} \frac{|g(x,\xi)|}{1+|\xi|^q} < +\infty \tag{3.2}$$

for some $q \in \left]0, \frac{2}{n-2}\right[$. Furthermore, denote by $\hat{\mathscr{A}}$ the class of all functions $h : \Omega \times \mathbf{R} \to \mathbf{R}$ of the type

$$h(x,\xi) = f(x,\xi) + \alpha(x)g(x,\xi)$$

with $f \in \mathscr{A}$, $g \in \tilde{\mathscr{A}}$ and $\alpha \in L^2(\Omega)$. For each $h \in \hat{\mathscr{A}}$, define the functional $I_h : H_0^1(\Omega) \to \mathbf{R}$, by putting

$$I_h(u) = \int_\Omega H(x, u(x))dx$$

for all $u \in H_0^1(\Omega)$, where

$$H(x,\xi) = \int_0^\xi h(x,t)dt$$

for all $(x,\xi) \in \Omega \times \mathbf{R}$.

By classical results (involving the Sobolev embedding theorem), the functional I_h turns out to be sequentially weakly continuous, of class C^1, with compact derivative given by

$$I_h'(u)(w) = \int_\Omega h(x, u(x))w(x)dx$$

for all $u, w \in H_0^1(\Omega)$.

Now, recall that, given $h \in \hat{\mathscr{A}}$, a weak solution of the problem

$$\begin{cases} -\left(a + b\int_\Omega |\nabla u(x)|^2 dx\right)\Delta u = h(x, u) & in\ \Omega \\ u = 0 & on\ \partial\Omega \end{cases}$$

is any $u \in H_0^1(\Omega)$ such that

$$\left(a + b\int_\Omega |\nabla u(x)|^2 dx\right)\int_\Omega \nabla u(x)\nabla w(x)dx = \int_\Omega h(x, u(x))w(x)$$

for all $w \in H_0^1(\Omega)$. Let $\Phi : H_0^1(\Omega) \to \mathbf{R}$ be the functional defined by

$$\Phi(u) = \frac{a}{2}\|u\|^2 + \frac{b}{4}\|u\|^4$$

for all $u \in H_0^1(\Omega)$.

Hence, the weak solutions of the problem are precisely the critical points in $H_0^1(\Omega)$ of the functional $\Phi - I_h$ which is said to be the energy functional of the problem.

The central result is as follows:

Theorem 3.1 *Let $n \geq 4$, let $f \in \mathscr{A}$ and let $g \in \tilde{\mathscr{A}}$ be such that the set*

$$\left\{x \in \Omega : \sup_{\xi \in \mathbf{R}} |g(x, \xi)| > 0\right\}$$

has a positive measure.

Then, there exists $\lambda^* \geq 0$ such that, for each $\lambda > \lambda^*$ and each convex set $C \subseteq L^2(\Omega)$ whose closure in $L^2(\Omega)$ contains the set $\{G(\cdot, u(\cdot)) : u \in H_0^1(\Omega)\}$, there exists $v^* \in C$ such that the problem

$$\begin{cases} -\left(a + b\int_\Omega |\nabla u(x)|^2 dx\right)\Delta u = f(x, u) + \lambda(G(x, u) - v^*(x))g(x, u) & in\ \Omega \\ u = 0 & on\ \partial\Omega \end{cases}$$

has at least three weak solutions, two of which are global minima in $H_0^1(\Omega)$ of the functional

$$u \to \frac{a}{2}\int_{\Omega}|\nabla u(x)|^2dx + \frac{b}{4}\left(\int_{\Omega}|\nabla u(x)|^2dx\right)^2 - \int_{\Omega}F(x,u(x))dx$$
$$-\frac{\lambda}{2}\int_{\Omega}|G(x,u(x)) - v^*(x)|^2dx\ .$$

If, in addition, the functional

$$u \to \frac{a}{2}\int_{\Omega}|\nabla u(x)|^2dx + \frac{b}{4}\left(\int_{\Omega}|\nabla u(x)|^2dx\right)^2 - \int_{\Omega}F(x,u(x))dx$$

has at least two global minima in $H_0^1(\Omega)$ and the function $G(x,\cdot)$ is strictly monotone for all $x \in \Omega$, then $\lambda^* = 0$.

The main tool we use to prove Theorem 3.1 is Theorem 3.C below which, in turn, is a direct consequence of two other results recently established in [24].

To state Theorem 3.C in a compact form, we now introduce some notations.

Here and in what follows, X is a non-empty set, V, Y are two topological spaces, y_0 is a point in Y.

We denote by $\mathcal{G}$ the family of all lower semicontinuous functions $\varphi : Y \to [0, +\infty[$, with $\varphi^{-1}(0) = \{y_0\}$, such that, for each neighbourhood U of y_0, one has

$$\inf_{Y\setminus U}\varphi > 0\ .$$

Moreover, denote by $\mathcal{H}$ the family of all functions $\Psi : X \times V \to Y$ such that, for each $x \in X$, $\Psi(x,\cdot)$ is continuous, injective, open, takes the value y_0 at a point v_x and the function $x \to v_x$ is not constant. Furthermore, denote by $\mathcal{M}$ the family of all functions $J : X \to \mathbf{R}$ whose set of all global minima (noted by M_J) is non-empty.

Finally, for each $\varphi \in \mathcal{G}$, $\Psi \in \mathcal{H}$ and $J \in \mathcal{M}$, put

$$\theta(\varphi,\Psi,J) = \inf\left\{\frac{J(x) - J(u)}{\varphi(\Psi(x,v_u))} : (u,x) \in M_J \times X \text{ with } v_x \neq v_u\right\}\ .$$

When X is a topological space, a function $\psi : X \to \mathbf{R}$ is said to be inf-compact if $\psi^{-1}(]-\infty, r])$ is compact for all $r \in \mathbf{R}$.

Theorem 3.A ([24, Theorem 3.1]) *Let $\varphi \in \mathcal{G}$, $\Psi \in \mathcal{H}$ and $J \in \mathcal{M}$.*

Then, for each $\lambda > \theta(\varphi,\Psi,J)$, one has

$$\sup_{v\in V}\inf_{x\in X}(J(x) - \lambda\varphi(\Psi(x,v))) < \inf_{x\in X}\sup_{z\in X}(J(x) - \lambda\varphi(\Psi(x,v_z)))\ .$$

Theorem 3.B ([24, Theorem 3.2]) *Let X be a topological space, E a real Hausdorff topological vector space, $C \subseteq E$ a convex set, $f : X \times C \to \mathbf{R}$ a function which is lower semicontinuous and inf-compact in X, and upper semicontinuous and concave in C. Assume also that*

$$\sup_{v\in C}\inf_{x\in X} f(x,v) < \inf_{x\in X}\sup_{v\in C} f(x,v)\ .$$

Then, there exists $v^* \in C$ such that the function $f(\cdot, v^*)$ has at least two global minima.

Theorem 3.C *Let $\varphi \in \mathscr{G}$, $\Psi \in \mathscr{H}$ and $J \in \mathscr{M}$. Moreover, assume that X is a topological space, that V is a real Hausdorff topological vector space and that $\varphi(\Psi(x,\cdot))$ is convex and continuous for each $x \in X$. Finally, let $\lambda > \theta(\varphi, \Psi, J)$ and let $C \subseteq V$ be a convex set, with $\{v_x : x \in X\} \subseteq \overline{C}$, such that the function $x \to J(x) - \lambda\varphi(\Psi(x,v))$ is lower semicontinuous and inf-compact in X for all $v \in C$.*

Under such hypotheses, there exists $v^* \in C$ such that the function $x \to J(x) - \lambda\varphi(\Psi(x,v^*))$ has at least two global minima in X.

Proof Set

$$D = \{v_x : x \in X\}$$

and, for each $(x,v) \in X \times V$, put

$$f(x,v) = J(x) - \lambda\varphi(\Psi(x,v))\ .$$

Theorem 3.A ensures that

$$\sup_{v\in V}\inf_{x\in X} f(x,v) < \inf_{x\in X}\sup_{v\in D} f(x,v)\ . \tag{3.3}$$

But, since $f(x,\cdot)$ is continuous and $D \subseteq \overline{C}$, we have

$$\sup_{v\in D} f(x,v) = \sup_{v\in\overline{D}} f(x,v) \le \sup_{v\in\overline{C}} f(x,v) = \sup_{v\in C} f(x,v)$$

for all $x \in X$, and hence, from (3.3), it follows that

$$\sup_{v\in C}\inf_{x\in X} f(x,v) < \inf_{x\in X}\sup_{v\in D} f(x,v) \le \inf_{x\in X}\sup_{v\in C} f(x,v)\ .$$

At this point, the conclusion follows applying Theorem 3.B to the restriction of the function f to $X \times C$. △

Proof of Theorem 3.1 For each $\lambda \ge 0$, $v \in L^2(\Omega)$, consider the function $h_{\lambda,v} : \Omega \times \mathbf{R} \to \mathbf{R}$ defined by

$$h_{\lambda,v}(x,\xi) = f(x,\xi) + \lambda(G(x,\xi) - v(x))g(x,\xi)$$

for all $(x, \xi) \in \Omega \times \mathbf{R}$. Clearly, the function $h_{\lambda,v}$ lies in $\hat{\mathscr{A}}$ and

$$H_{\lambda,v}(x, \xi) = F(x, \xi) + \frac{\lambda}{2}\left(|G(x, \xi) - v(x)|^2 - |v(x)|^2\right) .$$

So, the weak solutions of the problem are precisely the critical points in $H_0^1(\Omega)$ of the functional $\Phi - I_{h_{\lambda,v}}$. Moreover, if $p \in \left]0, \frac{n+2}{n-2}\right[$ and $q \in \left]0, \frac{2}{n-2}\right[$ are such that (3.1) and (3.2) hold, for some constant $c_{\lambda,v}$, we have

$$\int_\Omega |H_{\lambda,v}(x, u(x))|dx \le c_{\lambda,v}\left(\int_\Omega |u(x)|^{p+1} + \int_\Omega |u(x)|^{2(q+1)}dx + 1\right)$$

for all $u \in H_0^1(\Omega)$. Therefore, by the Sobolev embedding theorem, for a constant $\tilde{c}_{\lambda,v}$, we have

$$\Phi(u) - I_{h_{\lambda,v}}(u) \ge \frac{b}{4}\|u\|^4 - \tilde{c}_{\lambda,v}(\|u\|^{p+1} + \|u\|^{2(q+1)} + 1) \tag{3.4}$$

for all $u \in H_0^1(\Omega)$. On the other hand, since $n \ge 4$, one has

$$\max\{p + 1, 2(q + 1)\} < \frac{2n}{n - 2} \le 4 .$$

Consequently, from (3.4), we infer that

$$\lim_{\|u\|\to+\infty} (\Phi(u) - I_{h_{\lambda,v}}(u)) = +\infty . \tag{3.5}$$

Since the functional $\Phi - I_{h_{\lambda,v}}$ is sequentially weakly lower semicontinuous, by the Eberlein-Smulyan theorem and by (3.5), it follows that it is inf-weakly compact.

Now, we are going to apply Theorem 3.C taking $X = H_0^1(\Omega)$ with the weak topology and $V = Y = L^2(\Omega)$ with the strong topology, and $y_0 = 0$. Also, we take

$$\varphi(w) = \frac{1}{2}\int_\Omega |w(x)|^2 dx$$

for all $w \in L^2(\Omega)$. Clearly, $\varphi \in \mathscr{G}$. Furthermore, we take

$$\Psi(u, v)(x) = G(x, u(x)) - v(x)$$

for all $u \in H_0^1(\Omega)$, $v \in L^2(\Omega)$, $x \in \Omega$. Clearly, $\Psi(u, v) \in L^2(\Omega)$, $\Psi(u, \cdot)$ is a homeomorphism, and we have

$$v_u(x) = G(x, u(x)) .$$

We show that the map $u \to v_u$ is not constant in $H^1_0(\Omega)$. For each $x \in \Omega$, set

$$\alpha(x) = \inf_{\xi \in \mathbf{R}} G(x, \xi)$$

and

$$\beta(x) = \sup_{\xi \in \mathbf{R}} G(x, \xi) \ .$$

Since G is a Carathéodory is continuous, we have

$$\alpha(x) = \inf_{\xi \in \mathbf{Q}} G(x, \xi)$$

and

$$\beta(x) = \sup_{\xi \in \mathbf{Q}} G(x, \xi) \ ,$$

and so the functions α, β are measurable. Set

$$A = \{x \in \Omega : \alpha(x) < \beta(x)\} \ .$$

Clearly, we have

$$A = \left\{ x \in \Omega : \sup_{\xi \in \mathbf{R}} |g(x, \xi)| > 0 \right\} \ .$$

Hence, by assumption, $\text{meas}(A) > 0$. Then, by the classical Scorza-Dragoni theorem [6, Theorem 2.5.19], there exists a compact set $K \subset A$, of positive measure, such that the restriction of G to $K \times \mathbf{R}$ is continuous. Fix a point $\tilde{x} \in K$ such that the intersection of K and any ball centered at $\tilde{x}$ has a positive measure. Next, fix $\xi_1, \xi_2 \in \mathbf{R}$ such that

$$G(\tilde{x}, \xi_1) < G(\tilde{x}, \xi_2) \ .$$

By continuity, there is a closed ball $B(\tilde{x}, r)$ such that

$$G(x, \xi_1) < G(x, \xi_2)$$

for all $x \in K \cap B(\tilde{x}, r)$. Finally, consider two functions $u_1, u_2 \in H^1_0(\Omega)$ which are constant in $K \cap B(\tilde{x}, r)$. So, we have

$$G(x, u_1(x)) < G(x, u_2(x))$$

for all $x \in K \cap B(\tilde{x}, r)$. Hence, as $\operatorname{meas}(K \cap B(\tilde{x}, r)) > 0$, we infer that $v_{u_1} \neq v_{u_2}$, as claimed. As a consequence, $\Psi \in \mathscr{H}$. Of course, $\varphi(\Psi(u, \cdot))$ is continuous and convex for all $u \in X$. Finally, take

$$J = \Phi - I_f \ .$$

Clearly, $J \in \mathscr{M}$. So, for what seen above, all the assumptions of Theorem 3.C are satisfied. Consequently, if we take

$$\lambda^* = \theta(\varphi, \Psi, J) \tag{3.6}$$

and fix $\lambda > \lambda^*$ and a convex set $C \subseteq L^2(\Omega)$ whose closure in $L^2(\Omega)$ contains the set $\{G(\cdot, u(\cdot)) : u \in H_0^1(\Omega)\}$, there exists $v^* \in C$ such that the functional $\Phi - I_{h_{\lambda,v^*}}$ has at least two global minima in $H_0^1(\Omega)$ which are, therefore, weak solutions of the problem. To guarantee the existence of a third solution, denote by k the inverse of the restriction of the function $at + bt^3$ to $[0, +\infty[$. Let $T : X \to X$ be the operator defined by

$$T(w) = \begin{cases} \frac{k(\|w\|)}{\|w\|} w & if\, w \neq 0 \\ \\ 0 & if\, w = 0\,, \end{cases}$$

Since k is continuous and $k(0) = 0$, the operator T is continuous in X. For each $u \in X \setminus \{0\}$, we have

$$T(\Phi'(u)) = T((a + b\|u\|^2)u) = \frac{k((a + b\|u\|^2)\|u\|)}{(a + b\|u\|^2)\|u\|}(a + b\|u\|^2)u$$
$$= \frac{\|u\|}{(a + b\|u\|^2)\|u\|}(a + b\|u\|^2)u = u\ .$$

In other words, T is a continuous inverse of Φ'. Then, since $I_{h_{\lambda,v^{*\prime}}}$ is compact, the functional $\Phi - I_{h_{\lambda,v^*}}$ satisfies the Palais-Smale condition [29, Example 38.25] and hence the existence of a third critical point of the same functional is assured by Corollary 1 of [20].

Finally, assume that the functional $\Phi - I_f$ has at least two global minima, say $\hat{u}_1, \hat{u}_2$. Then, the set $D := \{x \in \Omega : \hat{u}_1(x) \neq \hat{u}_2(x)\}$ has a positive measure. By assumption, we have

$$G(x, \hat{u}_1(x)) \neq G(x, u_2(x))$$

for all $x \in D$, and so $v_{\hat{u}_1} \neq v_{\hat{u}_2}$. Then, by definition, we have

$$0 \leq \theta(\varphi, \Psi, J) \leq \frac{J(\hat{u}_1) - J(\hat{u}_2)}{\varphi(\Psi(\hat{u}_1, v_{\hat{u}_2}))} = 0$$

and so $\lambda^* = 0$ in view of (3.6). △

Notice the following corollary of Theorem 3.1:

Corollary 3.1 *Let $n \geq 4$, let $\nu \in \mathbf{R}$ and let $p \in \left]0, \frac{n+2}{n-2}\right[$.*

Then, for each $\lambda > 0$ large enough and for each convex set $C \subseteq L^2(\Omega)$ whose closure in $L^2(\Omega)$ contains $H_0^1(\Omega)$, there exists $v^* \in C$ such that the problem

$$\begin{cases} -\left(a + b\int_\Omega |\nabla u(x)|^2 dx\right) \Delta u = \nu |u|^{p-1}u + \lambda(u - v^*(x)) & \text{in } \Omega \\ \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

has at least three solutions, two of which are global minima in $H_0^1(\Omega)$ of the functional

$$u \to \frac{a}{2}\int_\Omega |\nabla u(x)|^2 dx + \frac{b}{4}\left(\int_\Omega |\nabla u(x)|^2 dx\right)^2 - \frac{\nu}{p+1}\int_\Omega |u(x)|^{p+1} dx$$
$$-\frac{\lambda}{2}\int_\Omega |u(x) - v^*(x)|^2 dx \ .$$

Proof Apply Theorem 3.1 taking $f(x, \xi) = |\xi|^{p-1}\xi$ and $g(x, \xi) = 1$. △

Remark 3.1 In Theorem 3.1, the assumption made on g (besides $g \in \tilde{\mathscr{A}}$) is essential. Indeed, if $g = 0$, for $f = 0$ (which is an allowed choice), the problem would have the zero solution only.

Remark 3.2 The assumption $n \geq 4$ is likewise essential. Indeed, Corollary 3.1 does not hold if $n = 3$. To see this, take $p = 4$ (which, when $n = 3$, is compatible with the condition $p < \frac{n+2}{n-2}$) and observe that the corresponding energy functional is unbounded below.

Besides Corollary 3.1, among the consequences of Theorem 3.1, we highlight the following

Theorem 3.2 *Let $n \geq 4$, let $f \in \mathscr{A}$ and let $g \in \tilde{\mathscr{A}}$ be such the set*

$$\left\{x \in \Omega : \sup_{\xi \in \mathbf{R}} F(x, \xi) > 0\right\}$$

has a positive measure. Moreover, assume that, for each $x \in \Omega$, $f(x, \cdot)$ is odd, $g(x, \cdot)$ is even and $G(x, \cdot)$ is strictly monotone.

Then, for each $\lambda > 0$, there exists $\mu^* > 0$ such that, for each $\mu > \mu^*$ and for each convex set $C \subseteq L^2(\Omega)$ whose closure in $L^2(\Omega)$ contains the set $\{G(\cdot, u(\cdot)) : u \in H_0^1(\Omega)\}$, there exists $v^* \in C$ such that the problem

$$\begin{cases} -\left(a+b\int_\Omega |\nabla u(x)|^2dx\right)\Delta u = \mu f(x,u) - \lambda v^*(x)g(x,u) & in\ \Omega \\ \\ u=0 & on\ \partial\Omega \end{cases}$$

has at least three weak solutions, two of which are global minima in $H^1_0(\Omega)$ *of the functional*

$$u \to \frac{a}{2}\int_\Omega |\nabla u(x)|^2dx + \frac{b}{4}\left(\int_\Omega |\nabla u(x)|^2dx\right)^2 - \mu\int_\Omega F(x,u(x))dx + \lambda\int_\Omega v^*(x)G(x,u(x))dx\ .$$

Proof Set

$$D = \left\{x \in \Omega : \sup_{\xi\in\mathbf{R}} F(x,\xi) > 0\right\}\ .$$

By assumption, meas$(D) > 0$. Then, by the Scorza-Dragoni theorem, there exists a compact set $K \subset D$, of positive measure, such that the restriction of F to $K \times \mathbf{R}$ is continuous. Fix a point $\hat{x} \in K$ such that the intersection of K and any ball centered at $\hat{x}$ has a positive measure. Choose $\hat{\xi} \in \mathbf{R}$ so that $F(\hat{x},\hat{\xi}) > 0$. By continuity, there is $r > 0$ such that

$$F(x,\hat{\xi}) > 0$$

for all $x \in K \cap B(\hat{x},r)$. Set

$$M = \sup_{(x,\xi)\in\Omega\times[-|\hat{\xi}|,|\hat{\xi}|]} |F(x,\xi)|\ .$$

Since $f \in \mathscr{A}$, we have $M < +\infty$. Next, choose an open set $\tilde{\Omega}$ such that

$$K \cap B(\hat{x},r) \subset \tilde{\Omega} \subset \Omega$$

and

$$\text{meas}(\tilde{\Omega} \setminus (K \cap B(\hat{x},r)) < \frac{\int_{K\cap B(\hat{x},r)} F(x,\hat{\xi})dx}{M}\ .$$

Finally, choose a function $\tilde{u} \in H^1_0(\Omega)$ such that

$$\tilde{u}(x) = \hat{\xi}$$

for all $x \in K \cap B(x,r)$,

$$\tilde{u}(x) = 0$$

for all $x \in \Omega \setminus \tilde{\Omega}$ and

$$|\tilde{u}(x)| \leq |\hat{\xi}|$$

for all $x \in \Omega$. Thus, we have

$$\int_{\Omega} F(x, \tilde{u}(x))dx = \int_{K\cap B(\hat{x},r)} F(x, \hat{\xi})dx + \int_{\tilde{\Omega}\setminus(K\cap B(\hat{x},r)} F(x, \tilde{u}(x))dx$$

$$> \int_{K\cap B(\hat{x},r)} F(x, \hat{\xi})dx - M\text{meas}(\tilde{\Omega} \setminus (K \cap B(\hat{x}, r)) > 0\,.$$

Now, fix any $\lambda > 0$ and set

$$\mu^* = \frac{\Phi(\tilde{u}) + \frac{\lambda}{2} I_{Gg}(\tilde{u})}{I_f(\tilde{u})}\,.$$

Fix $\mu > \mu^*$. Hence

$$\Phi(\tilde{u}) - \mu I_f(\tilde{u}) + \frac{\lambda}{2} I_{Gg}(\tilde{u}) < 0\,.$$

From this, we infer that the functional $\Phi - \mu I_f + \frac{\lambda}{2} I_{Gg}$ possesses at least to global minima since it is even. At this point, we can apply Theorem 3.1 to the functions g and $\mu f - \lambda Gg$. Our current conclusion follows from the one of Theorem 3.1 since we have $\lambda^* = 0$ and hence we can take the same fixed $\lambda > 0$. △

4 A Conjecture on Disconnectedness Versus Infinitely Many Solutions

Conjecture 4.1 Let $\Omega \subset \mathbf{R}^n$ be a smooth bounded domain, with $n \geq 3$. Let τ be the strongest vector topology on $H^1_0(\Omega)$.

Then, there exists a continuous function $f : \mathbf{R} \to \mathbf{R}$, with

$$\sup_{\xi\in\mathbf{R}} \frac{|f(\xi)|}{1 + |\xi|^{\frac{n+2}{n-2}}} < +\infty\,,$$

such that the set

$$\left\{(u, v) \in H^1_0(\Omega) \times H^1_0(\Omega) : \int_{\Omega} \nabla u(x)\nabla v(x)dx - \int_{\Omega} f(u(x))v(x)dx = 1\right\}$$

is disconnected in $(H^1_0(\Omega), \tau) \times (H^1_0(\Omega), \tau)$.

The importance of Conjecture 4.1 is shown by Proposition 4.3 below. But, first the relevant theory should be fixed.

The central abstract result, obtained in [22], is as follows (see also [16]):

Theorem 4.1 *Let X be a connected topological space, let E be a real topological vector space, with topological dual E^*, and let $A : X \to E^*$ be an operator such that the set*

$$\{y \in E : x \to \langle A(x), y\rangle \ \ is \ \ continuous\}$$

is dense in E and the set

$$\{(x, y) \in X \times E : \langle A(x), y\rangle = 1\}$$

is disconnected.

Then, A does vanish at some point of X.

Proof Denote by p_X the projection from $X \times E$ onto X. Moreover, for any $C \subseteq X \times E$, $x \in X$, put

$$C_x = \{y \in E : (x, y) \in C\}.$$

Arguing by contradiction, assume that $A(x) \neq 0$ for all $x \in X$. Denote by Γ the set

$$\{(x, y) \in X \times E : \langle A(x), y\rangle = 1\}.$$

Since Γ is disconnected, there are two open sets $\Omega_1, \Omega_2 \subseteq X \times E$ such that

$$\Omega_1 \cap \Gamma \neq \emptyset, \ \ \Omega_2 \cap \Gamma \neq \emptyset, \ \ \Omega_1 \cap \Omega_2 \cap \Gamma = \emptyset, \ \ \Gamma \subseteq \Omega_1 \cup \Omega_2.$$

We now prove that $p_X(\Omega_1 \cap \Gamma)$ is open in X. So, let $(x_0, y_0) \in \Omega_1 \cap \Gamma$. Since E is locally connected [28, p.35], there are a neighbourhood U_0 of x_0 in X and an open connected neighbourhood V_0 of y_0 in E such that $U_0 \times V_0 \subseteq \Omega_1$. Since $\langle A(x_0), \cdot\rangle$ is a non-null continuous linear functional, it has no local extrema. Consequently, since $\langle A(x_0), y_0\rangle = 1$, the sets

$$\{u \in V_0 : \langle A(x_0), u\rangle < 1\},$$

$$\{u \in V_0 : \langle A(x_0), u\rangle > 1\}$$

are both non-empty and open. Then, thanks to our density assumption, there are $u_1, u_2 \in V_0$ such that the set

$$\{x \in U_0 : \langle A(x), u_1\rangle < 1 < \langle A(x), u_2\rangle\}$$

is a neighbourhood of x_0. Then, if x is in this set, due to the connectedness of V_0, there is some $y \in V_0$ such that $\langle A(x), y\rangle = 1$, and so, x actually lies in $p_X(\Omega_1 \cap \Gamma)$, as desired. Likewise, it is seen that $p_X(\Omega_2 \cap \Gamma)$ is open. Now, observe that, for any $x \in X$, the set $\{x\} \times \Gamma_x$ is non-empty and connected, and so it is contained either in Ω_1 or in Ω_2. Summarizing, we then have that the sets $p_X(\Omega_1 \cap \Gamma)$ and $p_X(\Omega_2 \cap \Gamma)$ are non-empty, open, disjoint and cover X. Hence, X would be disconnected, a contradiction. △

Once Theorem 4.1 has been obtained, we can state the following formally more complete result:

Theorem 4.2 *Let X be a topological space, let E be a real topological vector space, and let $A : X \to E^*$ be such that the set*

$$\{y \in E : x \to \langle A(x), y\rangle \ \ is \ \ continuous\}$$

is dense in E.

Then, the following assertions are equivalent:

(i) *The set*

$$\{(x, y) \in X \times E : \langle A(x), y\rangle = 1\}$$

is disconnected.

(ii) *The set $X \setminus A^{-1}(0)$ is disconnected.*

Proof Let (i) hold. Since

$$\{(x, y) \in X \times E : \langle A(x), y\rangle = 1\} = \{(x, y) \in (X \setminus A^{-1}(0)) \times E : \langle A(x), y\rangle = 1\},$$

if $X \setminus A^{-1}(0)$ were connected, we could apply Theorem 4.1 to $A_{|(X \setminus A^{-1}(0))}$, and so A would vanish at some point of $X \setminus A^{-1}(0)$, which is absurd.

Conversely, if (ii) holds, then (i) follows at once observing that, with the notations of the proof of Theorem 4.1, one has $X \setminus A^{-1}(0) = p_X(\Gamma)$. △

Remark 4.1 When X is a connected topological space, E is an infinite-dimensional real vector space (with algebraic dual E'), and $A : X \to E'$ is a $\sigma(E', E)$-continuous operator, one could try to apply Theorem 4.1 endowing E with the strongest vector topology [15, p.53].

Remark 4.2 In Theorem 4.1, the role of the constant 1 can actually be assumed by any continuous real function on X. Precisely, we have the following

Proposition 4.1 *Let X be a topological space, let E be a real topological vector space, and let $A : X \to E'$. Assume that, for some continuous function $\alpha : X \to \mathbf{R}$, the set*

$$\Lambda := \{(x, y) \in X \times E : \langle A(x), y\rangle = \alpha(x)\}$$

is disconnected.

Then, either $A(x) = 0$ for some $x \in X$, or the set

$$\Gamma := \{(x, y) \in X \times E : \langle A(x), y\rangle = 1\}$$

is disconnected.

Proof Assume that $A^{-1}(0) = \emptyset$. So, $p_X(\Gamma) = X$. Consider the function $f : X \times E \to X \times E$ defined by putting $f(x, y) = (x, \alpha(x)y)$ for all $(x, y) \in X \times E$. Of course, f is continuous. Arguing by contradiction, assume that Γ is connected. Then, $f(\Gamma)$ is connected too. Now, observe that

$$\Lambda = \bigcup_{x\in\alpha^{-1}(0)} (f(\Gamma) \cup (\{x\} \times \Lambda_x)).$$

Furthermore, note that, if $x \in \alpha^{-1}(0)$, then $(x, 0) \in f(\Gamma) \cap (\{x\} \times \Lambda_x)$, and so $f(\Gamma) \cup (\{x\} \times \Lambda_x)$ is connected. In turn, the sets $f(\Gamma) \cup (\{x\} \times \Lambda_x)$ $(x \in \alpha^{-1}(0))$ are clearly pairwise non-disjoint, and hence Λ is connected, a contradiction. △

In [21], the following proposition was pointed out:

Proposition 4.2 ([21, Proposition 3]) *Let E be an infinite-dimensional Hausdorff topological vector space and K a relatively compact subset of E.*
Then, the set $E \setminus K$ is connected.

Finally, as said, the following proposition shows the importance of Conjecture 4.1:

Proposition 4.3 *Let f be a function satisfying Conjecture 4.1.*
Then, the problem

$$\begin{cases} -\Delta u = f(u) & in\ \Omega \\ \\ u = 0 & on\ \partial\Omega \end{cases}$$

has infinitely many weak solutions.

Proof Let $X = W_0^{1,2}(\Omega)$, with the usual norm $\|u\| = (\int_\Omega |\nabla u(x)|^2 dx)^{\frac{1}{2}}$. Put

$$J(u) = \frac{1}{2}\int_\Omega |\nabla u(x)|^2 dx - \int_\Omega \left(\int_0^{u(x)} f(\xi)d\xi\right) dx$$

for all $u \in X$.

So, the functional J is of class C^1 on X and one has

$$J'(u)(v) = \int_\Omega \nabla u(x)\nabla v(x)dx - \int_\Omega f(x, u(x))v(x)dx$$

for all $u, v \in X$. Hence, the critical points of J in X are exactly the weak solutions of the problem. Since J is of class C^1, clearly the operator $J' : X \to X^*$ is τ-weakly-star continuous. Hence, by Theorem 4.2, the set $X \setminus (J')^{-1}(0)$ is τ-disconnected. Then, due to Proposition 4.2, the set $(J')^{-1}(0)$ is not τ-relatively compact, and hence is infinite. △

References

1. A. Bahri, J.M. Coron, Sur une équation elliptique non linéaire avec l'exposant critique de Sobolev. C. R. Acad. Sci. Paris Sér. I Math. **301**, 345–348 (1985)
2. A. Bahri, J.M. Coron, On a nonlinear elliptic equation involving the critical Sobolev exponent: the effect of the topology of the domain. Commun. Pure Appl. Math. **41**, 253–294 (1988)
3. M. Clapp, F. Pacella, Multiple solutions to the pure critical exponent in domains with a hole of arbitrary size. Math. Z. **259**, 575–589 (2008)
4. J.M. Coron, Topologie et cas limite des injections de Sobolev. C. R. Acad. Sci. Paris Sér. I Math. **299**, 209–212 (1984)
5. E.N. Dancer, A note on an equation with critical exponent. Bull. London Math. Soc. **20**, 600–602 (1988)
6. Z. Denkowski, S. Migórski, N.S. Papageorgiou, *An Introduction to Nonlinear Analysis: Theory* (Kluwer Academic, Boston, 2003)
7. L. Diening, C. Kreuzer, S. Schwarzacher, Convex hull property and maximum principle for finite element minimisers of general convex functionals. Numer. Math. **124**, 685–700 (2013)
8. R. Engelking, *Theory of Dimensions, Finite and Infinite* (Heldermann, Lemgo, 1995)
9. X.L. Fan, A remark on Ricceri's conjecture for a class of nonlinear eigenvalue problems. J. Math. Anal. Appl. **349**, 436–442 (2009)
10. X.L. Fan, On Ricceri's conjecture for a class of nonlinear eigenvalue problems. Appl. Math. Lett. **22**, 1386–1389 (2009)
11. X.L. Fan, B. Ricceri, On the Dirichlet problem involving nonlinearities with non-positive primitive: a problem and a remark. Appl. Anal. **89**, 189–192 (2010)
12. N. Hirano, Existence of nontrivial solutions for a semilinear elliptic problem with supercritical exponent. Nonlinear Anal. **55**, 543–556 (2003)
13. N.I. Katzourakis, Maximum principles for vectorial approximate minimizers of nonconvex functionals. Calc. Var. Partial Differ. Equ. **46**, 505–522 (2013)
14. J.L. Kazdan, F.W. Warner, Remarks on some quasilinear elliptic equations. Commun. Pure Appl. Math. **28**, 567–597 (1975)
15. J.L. Kelley, I. Namioka, *Linear Topological Spaces* (Van Nostrand, Princeton, 1963)
16. A.J.B. Lopes-Pinto, On a new result on the existence of zeros due to Ricceri. J. Convex Anal. **5**, 57–62 (1998)
17. D. Passaseo, Multiplicity of positive solutions of nonlinear elliptic equations with critical Sobolev exponent in some contractible domains. Manuscripta Math. **65**, 147–175 (1989)
18. D. Passaseo, Nontrivial solutions of elliptic equations with supercritical exponent in contractible domains. Duke Math. J. **92**, 429–457 (1998)
19. S.I. Pohozaev, Eigenfunctions of the equation $\Delta u + \lambda f(u) = 0$. Soviet Math. Dokl., **6**, 1408–1411 (1965)

20. P. Pucci, J. Serrin, A mountain pass theorem. J. Differ. Equ. **60**, 142–149 (1985)
21. B. Ricceri, Applications of a theorem concerning sets with connected sections. Topol. Methods Nonlinear Anal. **5**, 237–248 (1995)
22. B. Ricceri, Existence of zeros via disconnectedness. J. Convex Anal. **2**, 287–290 (1995)
23. B. Ricceri, A remark on a class of nonlinear eigenvalue problems. Nonlinear Anal. **69**, 2964–2968 (2008)
24. B. Ricceri, A strict minimax inequality criterion and some of its consequences. Positivity **16**, 455–470 (2012)
25. B. Ricceri, Energy functionals of Kirchhoff-type problems having multiple global minima. Nonlinear Anal. **115**, 130–136 (2015)
26. B. Ricceri, The convex hull-like property and supported images of open sets. Ann. Funct. Anal. **7**, 150–157 (2016)
27. J. Saint Raymond, Open differentiable mappings. Le Matematiche **71**, 197–208 (2016)
28. H.H. Schaefer, *Topological Vector Spaces* (Springer, New York, 1971)
29. E. Zeidler, *Nonlinear Functional Analysis and Its Applications*, vol. III (Springer, New York, 1985)
30. E. Zeidler, *Nonlinear Functional Analysis and Its Applications*, vol. I (Springer, New York, 1986)

Corelations Are More Powerful Tools than Relations

Árpád Száz

To the Memory of my younger brother Géza Száz

1 Introduction

In our former papers [63, 65], a subset R of a product set $X \times Y$ was called a *relation* on X to Y. In particular, the relation R was called a *function* if the set $R\,(x) = \{ y \in Y : \ (x, \ y) \in R \}$ is either empty or a singleton for all $x \in X$. That is, if $x \in X$ such that $R\,(x) \neq \emptyset$, then there exists $y \in Y$ such that $R\,(x) = \{y\}$.

Note that a singleton $\{x\}$, with $x \in X$, can usually be identified with the element x. Thus, the set X may be considered as a subset of its *power set* $\mathscr{P}(X)$. Recall that, for any $A, \ B \subseteq X$, we have $A \in \mathscr{P}(B)$ if and only if $A \subseteq B$. Therefore, $\mathscr{P}$ is just another notation for the inclusion relation $\subseteq$.

In addition to the fundamental notions of unary and binary operations, in our recent paper [74], a function U of one power set $\mathscr{P}(X)$ to another $\mathscr{P}(Y)$ has been briefly called a *corelation* on X to Y. This definition differs from that of Pöschel and Rössiger [46] who have established a more difficult Galois connection.

Now, a relation R on X to Y can be naturally identified with the function φ_R defined by $\varphi_R(x) = R\,(x)$ for all $x \in X$. However, the corelation Φ_R, defined by $\Phi_R(A) = R\,[\,A\,] = \bigcup_{x \in A} \ R(x)$ for all $A \subseteq X$, will turn out to be a more powerful tool than the relation R and the function φ_R.

More concretely, we shall show that just the union-preserving corelations will correspond to relations. Note that, if the ground sets X and Y are not fixed, then a relation is a more general object than a corelation. Namely, functions have been defined as some very particular relations.

The above definition of a corelation has been mainly motivated by the observations of Höhle and Kubiak [25]. However, to feel the importance of corelations,

Á. Száz (✉)
Department of Mathematics, University of Debrecen, Debrecen, Hungary
e-mail: szaz@science.unideb.hu

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization
and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_25

the reader must only note that, for instance, the *complement* and *closure (interior) operations* on a set X are corelations on X.

In the sequel, in addition to the plausible notions of *increasingness* and *union-preservingness*, a corelation U on X to Y will be called *quasi-increasing* if $U\big(\{x\}\big) \subseteq U(A)$ for all $x \in A \subseteq X$. The importance of this property lies mainly in the Galois connection established in our former paper [74].

Having in mind the ideas of several former mathematicians such as Weil [87], Curtis and Mathews [11] and Nakano and Nakano [39], for instance, families $\mathscr{R}$ and $\mathscr{U}$ of relations and corelations on X to Y will now be called *relators* and *corelators* on X to Y, respectively.

Note that if d is a certain *distance function* on X, then the family $\mathscr{R}_d$ of all *surroundings* $B_r^d = \{(x, y) \in X^2 : \ d\,(x, y < r\}$ is an important relator on X. Namely, several notions can be more easily defined in terms of $\mathscr{R}_d$ than that of d. For instance, we can at once write $\operatorname{cl}_d(A) = \bigcap_{r>0} \big(B_r^d\big)^{-1}[\,A\,]$ for all $A \subseteq X$.

In [53, 69], by using the Davis–Pervin and Hunsaker–Lindgren relations [13, 26, 45], we have shown that relators on X are more powerful tools than generalized proximities, closures, topologies and filters on X. By Efremovič and Švarc [17] and our papers [49, 50], it is clear that the same must be true for convergences.

Therefore, it seems to be a big mistake that, following Tietze [86], just the concept of *open sets* has been chosen to be the starting point by Bourbaki [4], Kelley [28] and Engelking [18]. While, for instance, Sierpinski [47], Kowalsky [29], Isbell [27] and Čech [8] applied less attractive, but more powerful tools.

Common generalizations of topological, proximity and uniform spaces were formerly also given by several authors. See, for instance, Császár [9], Doitčinov [15] and Herrlich [24]. However, following the treatments of Murdeshwar and Naimpally [36] and Fletcher and Lingren [19], the basic topological structures can be more conveniently generalized in the framework of generalized uniformities.

In the present work, we shall show that corelators on X to Y are more powerful tools than relators on X to Y. Therefore, corelators have to be studied before relators. At present, I am considering them to be the most convenient starting point for general algebraic and analytical considerations [81].

If $\mathscr{U}$ is a corelation on X to Y, then instead of the standard notations of $\delta_{\mathscr{U}}$ and $\Subset_{\mathscr{U}}$ of Efremovič [16] and Smirnov [48], for any $A \subseteq X$ and $B \subseteq Y$, we write

1. $A \in \operatorname{Cl}_{\mathscr{U}}(B)$ if $U(A) \cap B \neq \emptyset$ for all $U \in \mathscr{U}$;
2. $A \in \operatorname{Int}_{\mathscr{U}}(B)$ if $U(A) \subseteq B$ for some $U \in \mathscr{U}$.

Thus, we can easily see that $\operatorname{Cl}_{\mathscr{U}}(B) = \mathscr{P}(X) \setminus \operatorname{Int}_{\mathscr{U}}(Y \setminus B)$ for all $B \subseteq Y$. Therefore, $\operatorname{Cl}_{\mathscr{U}}$ and $\operatorname{Int}_{\mathscr{U}}$ are also equivalent tools. Moreover, for any $x \in X$, we may also naturally write $x \in \operatorname{cl}_{\mathscr{U}}(B)$ if $\{x\} \in \operatorname{Cl}_{\mathscr{U}}(B)$, and $x \in \operatorname{int}_{\mathscr{U}}(B)$ if $\{x\} \in \operatorname{Int}_{\mathscr{U}}(B)$. Thus, we also have $\operatorname{cl}_{\mathscr{U}}(B) = X \setminus \operatorname{int}_{\mathscr{U}}(Y \setminus B)$ for all $B \subseteq Y$.

However, it is now more important to note that $\operatorname{Int}_{\mathscr{U}}$ is a relation on $\mathscr{P}(Y)$ to $\mathscr{P}(X)$ such that $\operatorname{Int}_{\mathscr{U}} = \bigcup_{U \in \mathscr{U}} \operatorname{Int}_U$ with $\operatorname{Int}_U = \operatorname{Int}_{\{U\}}$. Therefore, the properties of the relation $\operatorname{Int}_{\mathscr{U}}$ can be immediately derived from those of the relations Int_U. This shows that corelations have to be studied before corelators.

Following the ideas of Höhle and Kubiak [25] and the notations of Davey and Priestly [12, p. 155], for any relation R and corelation U on X to Y, we define a corelation $R^{\triangleright}$ and a relation $U^{\triangleleft}$ on X to Y such that, for all $A \subseteq X$ and $x \in X$,

$$R^{\triangleright}(A) = R\,[\,A\,] \qquad\qquad \text{and} \qquad\qquad U^{\triangleleft}(x) = U\big(\{x\}\big).$$

Here, for any two corelations U and V on X to Y, we write $U \leq V$ if $U(A) \subseteq V(A)$ for all $A \subseteq X$. Thus, the maps $\triangleright$ and $\triangleleft$ establish a Galois connection in the sense that, for any relation R and quasi-increasing corelation U on X to Y we have $R^{\triangleright} \leq U \iff R \subseteq U^{\triangleleft}$.

This important Galois connection has the particular property that $R^{\triangleright\triangleleft} = R$ for all relation R on X to Y. Moreover, if U is a corelation on X to Y, then under the notation $U^{\circ} = U^{\triangleleft\triangleright}$, we have $U^{\circ} = U$ $\big($resp. $U^{\circ} \leq U\big)$ if and only if U is union-preserving (resp. quasi-increasing).

In this respect, it is also worth mentioning that $\circ$ is always a *projection* (increasing and idempotent) operation on the family of all corelations U on X to Y such that $U^{\circ}(A) = \bigcup_{x \in A} U\big(\{x\}\big)$ for all $A \subseteq X$. Moreover, the operation $\circ$ is also compatible with certain set and relation theoretic operations.

By using the maps $\triangleright$ and $\triangleleft$, for any two corelations U on X to Y and V on Y to Z, we may also naturally define $U^{\mathbf{-1}} = U^{\triangleright -1 \triangleright}$ and $V \bullet U = \big(V^{\triangleleft} \circ U^{\triangleleft}\big)^{\triangleright}$. Moreover, for instance, for any relator $\mathcal{R}$ on X to Y, we may also naturally define $\operatorname{Int}_{\mathcal{R}} = \operatorname{Int}_{\mathcal{R}^{\triangleright}}$ and $\operatorname{int}_{\mathcal{R}} = \operatorname{int}_{\mathcal{R}^{\triangleright}}$, where $\mathcal{R}^{\triangleright} = \big\{R^{\triangleright} : \; R \in \mathcal{R}\big\}$.

Thus, $\operatorname{Int}_{\mathcal{U}}$ is a more powerful tool than $\operatorname{Int}_{\mathcal{R}}$. However, for instance, we already have $\operatorname{Int}_{\mathcal{U}^{\circ}} = \operatorname{Int}_{\mathcal{U}^{\triangleleft}}$ and $\operatorname{int}_{\mathcal{U}} = \operatorname{int}_{\mathcal{U}^{\circ}} = \operatorname{int}_{\mathcal{U}^{\triangleleft}}$. Thus, our former results on the relation $\operatorname{int}_{\mathcal{R}}$ and the families $\mathcal{E}_{\mathcal{R}} = \{B \subseteq Y : \; \operatorname{int}_{\mathcal{R}}(B) \neq \emptyset\}$ and $\mathcal{T}_{\mathcal{R}} = \{A \subseteq X : A \subseteq \operatorname{int}_{\mathcal{R}}(A)\}$, whenever $X = Y$, will not be generalized.

2 Some Basic Definitions on Relations

A subset F of a product set $X \times Y$ is called a *relation* on X to Y. In particular, a relation on X to itself is called a relation on X. And, $\Delta_X = \{(x, x) : \; x \in X\}$ is called the *identity relation* on X.

If F is a relation on X to Y, then by the above definitions we can also state that F is a relation on $X \cup Y$. However, for several purposes, the latter view of the relation F would be quite unnatural.

If F is a relation on X to Y, then for any $x \in X$ and $A \subseteq X$ the sets $F(x) = \{y \in Y : \; (x, y) \in F\}$ and $F\,[\,A\,] = \bigcup_{a \in A} F(a)$ are called the *images* of x and A under F. If $(x, y) \in F$, then we may also write $x\,F\,y$.

Moreover, the sets $D_F = \{x \in X : \; F(x) \neq \emptyset\}$ and $R_F = F\,[\,X\,]$ are called the *domain* and *range* of F, respectively. If in particular $D_F = X$, then we say that F is a relation of X to Y, or that F is a *total relation* on X to Y.

In particular, a relation f on X to Y is called a *function* if for each $x \in D_f$ there exists $y \in Y$ such that $f(x) = \{y\}$. In this case, by identifying singletons with their elements, we may simply write $f(x) = y$ in place of $f(x) = \{y\}$.

Moreover, a function $\star$ of X to itself is called a *unary operation* on X. While, a function $*$ of X^2 to X is called a *binary operation* on X. And, for any $x,\ y \in X$, we usually write $x^\star$ and $x * y$ instead of $\star(x)$ and $*\big((x,\ y)\big)$.

If F is a relation on X to Y and $U \subseteq D_F$, then the relation $F \,|\, U = F \cap (U \times Y)$ is called the *restriction* of F to U. Moreover, if F and G are relations on X to Y such that $D_F \subseteq D_G$ and $F = G \,|\, D_F$, then G is called an *extension* of F.

If F is a relation on X to Y, then we can easily see that $F = \bigcup_{x \in X} \{x\} \times F(x)$. Therefore, the values $F(x)$, where $x \in X$, uniquely determine F. Thus, a relation F on X to Y can also be naturally defined by specifying $F(x)$ for all $x \in X$.

For instance, the *complement* F^c and the *inverse* F^{-1} can be defined such that $F^c(x) = Y \setminus F(x)$ for all $x \in X$ and $F^{-1}(y) = \{x \in X : \ y \in F(x)\}$ for all $y \in Y$. Thus, we also have $F^c = X \times Y \setminus F$ and $F^{-1} = \{(y,\ x) \in Y \times X : \ (x,\ y) \in F\}$.

Moreover, if in addition G is a relation on Y to Z, then the *composition* $G \circ F$ can be defined such that $(G \circ F)(x) = G\,[\,F(x)\,]$ for all $x \in X$. Thus, we also have $G \circ F = \{(x,\ z) \in X \times Z : \ \exists\ y \in Y : \ (x,\ y) \in F, \ (y,\ z) \in G\}$.

While, if G is a relation on Z to W, then the *box product* $F \boxtimes G$ can be naturally defined such that $(F \boxtimes G)(x,\ z) = F(x) \times G(z)$ for all $x \in X$ and $z \in Z$. Note that the box product can be defined for any family of relations.

If F is a relation on X to Y, then a function f of D_F to Y is called a *selection* of F if $f \subseteq F$, i. e., $f(x) \in F(x)$ for all $x \in D_F$. Thus, by the Axiom of Choice, every relation has a selection. Moreover, it is the union of its selections.

For any relation F on X to Y, we may naturally define two *set-valued functions* φ_F of X to $\mathscr{P}(Y)$ and Φ_F of $\mathscr{P}(X)$ to $\mathscr{P}(Y)$ such that $\varphi_F(x) = F(x)$ for all $x \in X$ and $\Phi_F(A) = F\,[\,A\,]$ for all $A \subseteq X$.

Functions of X to $\mathscr{P}(Y)$ can be identified with relations on X to Y. While, functions of $\mathscr{P}(X)$ to $\mathscr{P}(Y)$ are more general objects than relations on X to Y. They were briefly called *corelations* on X to Y in [74].

Now, a relation R on X may be briefly defined to be *reflexive* on X if $\Delta_X \subseteq R$, and *transitive* if $R \circ R \subseteq R$. Moreover, R may be briefly defined to be *symmetric* if $R^{-1} \subseteq R$, and *antisymmetric* if $R \cap R^{-1} \subseteq \Delta_X$.

Thus, a reflexive and transitive (symmetric) relation may be called a *preorder (tolerance) relation*. And, a symmetric (antisymmetric) preorder relation may be called an *equivalence (partial order) relation*.

For instance, for $A \subseteq X$, the *Pervin relation* $R_A = A^2 \cup A^c \times X$ is a preorder relation on X. (See [32, 69].) While, for a *pseudo-metric* d on X and $r > 0$, the *surrounding* $B_r^d = \big\{(x,\ y) \in X^2 : \ d(x,\ y) < r\big\}$ is a tolerance relation on X.

Moreover, we may recall that if $\mathscr{A}$ is a *partition* of X, i. e., a family of pairwise disjoint, nonvoid subsets of X such that $X = \bigcup \mathscr{A}$, then $S_{\mathscr{A}} = \bigcup_{A \in \mathscr{A}} A^2$ is an equivalence relation on X, which can, to some extent, be identified with $\mathscr{A}$.

Now, for any relation R on X, we may also naturally define $R^0 = \Delta_X$ and $R^n = R \circ R^{n-1}$ if $n \in \mathbb{N}$. Moreover, we may naturally define $R^\infty = \bigcup_{n=0}^{\infty} R^n$. Thus, R^∞ is just the smallest preorder relation containing R [22].

If R is a relation on X to Y, then the ordered pair $(X, Y)(R) = \big((X, Y), R\big)$ is called a *relational space* (simple relator space [41]) or a *context space* (formal context) [20].

Having in mind an abbreviation of Birkhoff [1], a relational space $X(\leq) = (X, X)(\leq)$ is called a *goset* (generalized ordered set) [79]. And, a preordered set (partially ordered set) is called a *proset (poset)*.

If f is a function of one goset X to another Y and g is a function of Y to X such that, for any $x \in X$ and $y \in Y$, we have

$$f(x) \leq y \quad \Longleftrightarrow \quad x \leq g(y),$$

then we say that the functions f and g establish a *Galos connection* between X and Y or that f is a g-normal function of X to Y [71].

While, if f is a function of one goset X to another Y and φ is an unary operation on X such that, for any $u, v \in X$, we have

$$f(u) \leq g(v) \quad \Longleftrightarrow \quad u \leq \varphi(v),$$

then we say that the functions f and φ establish a *Pataki connection* between X and Y or that f is a φ-regular function of X to Y [68].

If f is a g-normal function of X to Y and $\varphi = g \circ f$, then we can at once see that $f(u) \leq f(v) \iff u \leq g\big(f(v)\big) \iff u \leq (g \circ f)(v) \iff u \leq \varphi(v)$ for all $u, v \in X$. Therefore, f is φ-regular.

Conversely, if f is a φ-regular function of X onto Y and g is a function of Y to X such that $\varphi = g \circ f$, then we can quite similarly see that f is g-normal. Thus, normal functions are more general than the regular ones.

Galois and Pataki connections occur almost every branches of mathematics. They allow of transposing notions and statements from one world of our imagination to another one. (For their theories and applications, see [2, 12, 14, 21].)

Some examples and generalizations of Galois and Pataki connections can also be found in our former papers [5, 62, 78, 80] and [75–77, 83]. However, it is frequently enough to consider such connections only for corelations [67].

3 A Few Basic Theorems on Relations

Concerning relations, we shall frequently need the following theorems.

Theorem 3.1 *For any two relations F and G on X to Y, the following assertions are equivalent:*

(1) $F \subseteq G$,

(2) $F(x) \subseteq G(x)$ *for all* $x \in X$,

(3) $F(x) \subseteq G(x)$ *for all* $x \in D_F$.

Corollary 3.2 *For any relations F and G on X to Y, the following assertions are equivalent:*

(1) $F = G$,

(2) $F(x) = G(x)$ *for all* $x \in X$,

(3) $D_F = D_G$ *and* $F(x) = G(x)$ *for all* $x \in D_F$.

Theorem 3.3 *If F is a relation on X to Y, then for any $A \subseteq X$ and $B \subseteq Y$, the following assertions are equivalent:*

(1) $F[A] \cap B \neq \emptyset$; *(2)* $A \cap F^{-1}[B] \neq \emptyset$.

Corollary 3.4 *If F is a relation on X to Y, then for any $B \subseteq Y$ we have*

$$F^{-1}[B] = \big\{x \in X : \ F(x) \cap B \neq \emptyset\big\}.$$

Remark 3.5 Thus, in particular, for any relation F on X to Y we have

(1) $D_F = F^{-1}[X] = R_{F^{-1}}$; (2) $D_{F^{-1}} = F[X] = R_F$.

Theorem 3.6 *If F and G are relations on X to Y, then*

(1) $(F \setminus G)^{-1} = F^{-1} \setminus G^{-1}$; *(2)* $\big(F^c\big)^{-1} = \big(F^{-1}\big)^c$.

Theorem 3.7 *If $\mathscr{F}$ is a family of relations on X to Y, then*

(1) $\Big(\bigcap \mathscr{F}\Big)^{-1} = \bigcap_{F \in \mathscr{F}} F^{-1}$; *(2)* $\Big(\bigcup \mathscr{F}\Big)^{-1} = \bigcup_{F \in \mathscr{F}} F^{-1}$.

Theorem 3.8 *If F is a relation on X to Y and G is a relation on Y to Z, then for any $A \subseteq X$ we have*

$$(G \circ F)[A] = G\big[F[A]\big].$$

Theorem 3.9 *If F is a relation on X to Y, then*

(1) $F \circ \emptyset = \emptyset = \emptyset \circ F$; *(2)* $F \circ X^2 = X \times R_F$;

(3) $Y^2 \circ F = D_F \times Y$; *(4)* $F \circ \Delta_X = F = \Delta_Y \circ F$.

Theorem 3.10 *If F is a relation on X to Y and G is a relation on Y to Z, then*

$$(G \circ F)^{-1} = F^{-1} \circ G^{-1}.$$

Theorem 3.11 *If F is a relation on X to Y, G is a relation on Y to Z, and H is a relation on Z to W, then*

$$H \circ (G \circ F) = (H \circ G) \circ F.$$

Remark 3.12 From Theorems 3.11 and 3.9, we can see that $\mathscr{P}(X^2)$, with composition, is a monoid (semigroup with identity).

Therefore, by induction, for any relation R on X and $n \in \{0\} \cup \mathbb{N}$ we may naturally define $R^n = \Delta_X$ if $n = 0$ and $R^n = R \circ R^{n-1}$ if $n > 0$.

Theorem 3.13 *If R is a relation on X, then*

$$R^\infty = \bigcup_{n=0}^{\infty} R^n$$

is the smallest preorder relation on X such that $R \subseteq R^\infty$.

Corollary 3.14 *A relation R on X is a preorder on X if and only if $R = R^\infty$.*

Theorem 3.15 *For any relation R on X, we have*

(1) $R^\infty = \left(R^\infty\right)^\infty$; *(2)* $\left(R^\infty\right)^{-1} = \left(R^{-1}\right)^\infty$.

Corollary 3.16 *A relation R on X is a preorder on X if and only if its inverse R^{-1} is a preorder on X.*

Theorem 3.17 *For a relation F on X to Y, the following assertions are equivalent:*
(1) F is total; *(2)* $\Delta_X \subseteq F^{-1} \circ F$;
(3) $F[\,A\,] \subseteq B$ implies $A \subseteq F^{-1}[\,B\,]$ for all $A \subseteq X$ and $B \subseteq Y$.

Corollary 3.18 *A relation F on X to Y is total if and only if the relation $F^{-1} \circ F$ is reflexive on X.*

Theorem 3.19 *For a relation F on X to Y, the following assertions are equivalent:*
(1) F is a function; *(2)* $F \circ F^{-1} \subseteq \Delta_Y$;
(3) $A \subseteq F^{-1}[\,B\,]$ implies $F[\,A\,] \subseteq B$ or all $A \subseteq X$ and $B \subseteq Y$.

Corollary 3.20 *A relation F on X to Y is a function if and only if $F \circ F^{-1} = \Delta_{R_F}$.*

Theorem 3.21 *For a function f on X to Y and a function g on Y to Y, the following assertions are equivalent:*

(1) $g = f^{-1}$; *(2)* $f \circ g = \Delta_{D_g}$ *and* $g \circ f = \Delta_{D_f}$.

Theorem 3.22 *If f is a function of X to Y, then for any $A \subseteq X$ and $B \subseteq Y$ we have*

$$f[\,A\,] \subseteq B \iff A \subseteq f^{-1}[\,B\,].$$

Remark 3.23 This theorem shows that if f is a function on X to Y, then the setfunctions Φ_f and $\Phi_{f^{-1}}$ establish a *Galois connection* between the posets $\mathscr{P}(X)$ and $\mathscr{P}(Y)$.

Note that, if F is a relation on X to Y, then by Theorem 3.3 we can also state that the setfunctions Φ_F and $\Phi_{F^{-1}}$ also establish a Galois connection between the power sets $\mathscr{P}(X)$ and $\mathscr{P}(Y)$. However, for this, we have to write $A \leq B$ if $A \cap B \neq \emptyset$.

4 Some Further Theorems on Relations

Theorem 4.1 *If F is a relation on X to Y, then for any $A, B \subseteq X$ we have*

(1) $F[A] \setminus F[B] \subseteq F[A \setminus B]$; *(2)* $F[A]^c \subseteq F[A^c]$ *if* $Y = R_F$.

Theorem 4.2 *If F is a relation on X to Y, then for any family $\mathscr{A}$ of subsets of X we have*

(1) $F\left[\bigcap \mathscr{A}\right] \subseteq \bigcap_{A \in \mathscr{A}} F[A]$; *(2)* $F\left[\bigcup \mathscr{A}\right] = \bigcup_{A \in \mathscr{A}} F[A]$.

Remark 4.3 If in particular F^{-1} is a function, then the corresponding equalities are also true in the above two theorems.

Remark 4.4 Note that if F is a relation on X to Y, then F^{-1} is a function if and only if $F(x) \cap F(z) \neq \emptyset$ implies $x = z$ for all $x, z \in X$.

Theorem 4.5 *If F and G are relations on X to Y, then for any $A \subseteq X$ we have*

(1) $F[A] \setminus G[A] \subseteq (F \setminus G)[A]$; *(2)* $F[A]^c \subseteq F^c[A]$ *if* $A \neq \emptyset$.

Theorem 4.6 *If $\mathscr{F}$ is a family of relations on X to Y, then for any $A \subseteq X$ we have*

(1) $\left(\bigcap \mathscr{F}\right)[A] \subseteq \bigcap_{F \in \mathscr{F}} F[A]$; *(2)* $\left(\bigcup \mathscr{F}\right)[A] = \bigcup_{F \in \mathscr{F}} F[A]$.

Remark 4.7 If in particular A is a singleton, then the corresponding equalities are also true in the above two theorems.

Remark 4.8 Note that if F is a relation on X onto Y, then by Theorems 4.1 and 4.5, we have $F[A]^c \subseteq F[A^c] \cap F^c[A]$ for all nonvoid subset A of X.

Theorem 4.9 *If F is a relation on X to Y, the for any $A \subseteq X$ we have*

$$F^c[A]^c = \bigcap_{x \in A} F(x).$$

Corollary 4.10 *If F is a relation on X to Y and G is a relation on Y to Z, then for any $x \in X$ we have*

$$\big(G \circ F\big)^c(x) = \bigcap_{y \in F(x)} G^c(y).$$

Corollary 4.11 *If f is a function on X to Y and G is a relation on Y to Z, then $(G \circ f)^c = G^c \circ f$.*

Theorem 4.12 *If F is a relation on X to Y and G is a relation on Y to Z, then*

(1) $(G \circ F)^c \subseteq G^c \circ F$ *if* $X = D_F$; *(2)* $(G \circ F)^c \subseteq G \circ F^c$ *if* $Z = R_G$.

Theorem 4.13 *If F and G are relations on X to Y and H is a relation on Y to Z, then*

$$\big(H\circ F\big)\setminus\big(H\circ G\big)\subseteq H\circ\big(F\setminus G\big).$$

Theorem 4.14 *If F is a relation on X to Y and G and H are relations on Y to Z, then*

$$\big(G\circ F\big)\setminus\big(H\circ F\big)\subseteq\big(G\setminus H\big)\circ F.$$

Theorem 4.15 *If $\mathscr{F}$ is a family of relations on X to Y and G is a relation on Y to Z, then*

(1) $G\circ\bigcap\mathscr{F}\subseteq\bigcap_{F\in\mathscr{F}}G\circ F,$ *(2)* $G\circ\bigcup\mathscr{F}=\bigcup_{F\in\mathscr{F}}G\circ F,.$

Theorem 4.16 *If F is a relation on X to Y and $\mathscr{G}$ is a family of relations on Y to Z, then*

(1) $\Big(\bigcap\mathscr{G}\Big)\circ F\subseteq\bigcap_{G\in\mathscr{G}}G\circ F,$ *(2)* $\Big(\bigcup\mathscr{G}\Big)\circ F=\bigcup_{G\in\mathscr{G}}G\circ F.$

Theorem 4.17 *For any relations F on X to Z and G on Y to W, we have*

$$(\,F\boxtimes G\,)^{-1}=F^{-1}\boxtimes G^{-1}.$$

Theorem 4.18 *If F is a relation on X to Z and G is a relation on Y to W, then for any $R\subseteq X\times Y$ we have*

$$(\,F\boxtimes G\,)\,[\,R\,]=G\circ R\circ F^{-1}.$$

Corollary 4.19 *For any relations F on X to Y and G on Y to Z, we have*

$$G\circ F=\big(F^{-1}\boxtimes G\big)\,[\,\Delta_Y].$$

Corollary 4.20 *If F is a relation on X to Z and G is a relation on Y to W, then for any $x\in X$ and $y\in Y$, we have*

$$(\,F\boxtimes G\,)(x,\ y)=G\circ\{(x,\ y)\}\circ F^{-1}.$$

Remark 4.21 These corollaries show that the box and composition products of two relations are actually equivalent tools.

However, in contrast to the composition product, the box product of relations can be immediately defined for arbitrary families of relations.

5 Increasing and Union-Preserving Corelations

The following terminology was first introduced in our former paper [74].

Definition 5.1 A function U on one power set $\mathscr{P}(X)$ to another $\mathscr{P}(Y)$ is called a *corelation* on X to Y.

Remark 5.2 In particular, a corelation U on X to itself will be simply called a corelation on X.

Moreover, a corelation U on X to Y will be said to be a corelation on X onto Y if it maps $\mathscr{P}(X)$ onto $\mathscr{P}(Y)$.

Note that if a subset A of X is not in the domain of U, then by the corresponding definition for relations we have $U\,(A) = \emptyset$. Therefore, every corelation on X to Y is actually a corelation of X to Y.

Definition 5.3 A corelation U on X to Y is called

(1) *increasing* if $U(A) \subseteq U(B)$ for all $A \subseteq B \subseteq X$;
(2) *quasi-increasing* if $U\big(\{x\}\big) \subseteq U(A)$ for all $x \in A \subseteq X$;
(3) *union-preserving* if $U\big(\bigcup \mathscr{A}\big) = \bigcup_{A \in \mathscr{A}} U(A)$ for all $\mathscr{A} \subseteq \mathscr{P}(X)$.

Remark 5.4 In particular, a corelation U on X may be called *extensive, intensive, involutive* and *idempotent* if $A \subseteq U(A)$, $U(A) \subseteq A$, $U\big(U(A)\big) = A$ and $U\big(U(A)\big) = U(A)$ for all $A \subseteq X$, respectively.

Moreover, an increasing involutive (idempotent) corelation is called a *involution (projection) operation.* While, an extensive (intensive) projection operation is called a *closure (interior) operation.*

Furthermore, an increasing extensive (intensive) corelation is called a *preclosure (preinterior) operation*. And, an extensive (intensive) idempotent corelation is called a *semiclosure (semiinterior)* operation.

Simple reformulations of properties (1) and (2) in Definition 5.3 give the following three theorems.

Theorem 5.5 *For a corelation U on X to Y, the following assertions are equivalent:*

(1) U is quasi-increasing;
(2) $\bigcup_{x \in A} U\big(\{x\}\big) \subseteq U(A)$ *for all* $A \subseteq X$.

Theorem 5.6 *For a corelation U on X to Y, the following assertions are equivalent:*

(1) U is increasing;
(2) $U\big(\bigcap \mathscr{A}\big) \subseteq \bigcap_{A \in \mathscr{A}} U(A)$ *for all* $\mathscr{A} \subseteq \mathscr{P}(X)$;
(3) $U\big(A_1 \cap A_2\big) \subseteq U\big(A_1\big) \cap U\big(A_2\big)$ *for all* $A_1,\ A_2 \subseteq X$.

Theorem 5.7 *For a corelation U on X to Y, the following assertions are equivalent:*

(1) U is increasing;

(2) $\bigcup_{A \in \mathscr{A}} U(A) \subseteq U\big(\bigcup \mathscr{A} \big)$ *for all* $\mathscr{A} \subseteq \mathscr{P}(X)$*;*

(3) $U\big(A_1\big) \cup U(A_2) \subseteq U\big(A_1 \cup A_2\big)$ *for all* $A_1,\ A_2 \subseteq X$*.*

Hence, it is clear that in particular we also have the following

Corollary 5.8 *For a corelation U on X to Y, the following assertions are equivalent:*

(1) U *is union-preserving;*

(2) U *is increasing and* $U\big(\bigcup \mathscr{A} \big) \subseteq \bigcup_{A \in \mathscr{A}} U(A)$ *for all* $\mathscr{A} \subseteq \mathscr{P}(X)$*.*

However, it is now more important to note that we also have the following theorem which has also been proved, in a different way, by Pataki [43].

Theorem 5.9 *For a corelation U on X to Y, the following assertions are equivalent:*

(1) U *is union-preserving;*

(2) $U(A) = \bigcup_{x \in A} U\big(\{x\}\big)$ *for all* $A \subseteq X$*.*

Proof To prove the implication (2) $\Longrightarrow$ (1), note that if (2) holds, then U is increasing. Therefore, by Theorem 5.7, we have $\bigcup_{A \in \mathscr{A}} U(A) \subseteq U\big(\bigcup \mathscr{A} \big)$ for all $\mathscr{A} \subseteq \mathscr{P}(X)$. Thus, to obtain (1), we need only prove the converse inclusion.

For this, note that if $\mathscr{A} \subseteq \mathscr{P}(X)$, then by (2) we have

$$U\big(\bigcup \mathscr{A} \big) = \bigcup_{x \in \bigcup \mathscr{A}} U\big(\{x\}\big).$$

Therefore, if $y \in U\big(\bigcup \mathscr{A} \big)$, then there exists $x \in \bigcup \mathscr{A}$ such that $y \in U\big(\{x\}\big)$. Thus, in particular there exists $A_0 \in \mathscr{A}$ such that $x \in A_0$, and so $\{x\} \subseteq A_0$. Hence, by using the increasingness of U, we can already see that

$$y \in U\big(\{x\}\big) \subseteq U(A_0) \subseteq \bigcup_{A \in \mathscr{A}} U(A).$$

Therefore, $U\big(\bigcup \mathscr{A} \big) \subseteq \bigcup_{A \in \mathscr{A}} U(A)$ also holds.

From this theorem, by Theorem 5.5, it is clear that in particular we also have

Corollary 5.10 *For a corelation U on X to Y, the following assertions are equivalent:*

(1) U *is union-preserving;*

(2) U *is quasi-increasing and* $U(A) \subseteq \bigcup_{x \in A} U\big(\{x\}\big)$ *for all* $A \subseteq X$*.*

Now, by using Theorem 5.9, we can also easily establish the following two examples.

Example 5.11 If R is a reflexive relation on X and

$$U(A) = R^{-1}[\,A\,]$$

for all $A \subseteq X$, then by the reflexivity of R^{-1} and Theorem 5.9 it is clear that U is a union-preserving preclosure operation on X.

Remark 5.12 Note that, for any $x \in X$ and $A \subseteq X$, we have $x \in R^{-1}[\,A\,]$ if and only if there exists $a \in A$ such that $x \in R^{-1}(a)$, and thus $a \in R\,(x)$. Therefore, $U(A)$ is just the R-closure of A.

Example 5.13 Suppose that $\mathscr{C}$ is a T_1-separating cover of a set X in the sense that $X = \bigcup \mathscr{C}$ and for every $x,\ y \in X$, with $x \neq y$, there exist $C \in \mathscr{C}$ such that such that $x \in C$, but $y \notin C$.

For any $A \subseteq X$, define

$$U\,(A) = \bigcap \big\{ C \in \mathscr{C} : \ \ A \subseteq C \big\}.$$

Then, it can be easily seen that U is a preclosure operation on X such that $U\big(\{x\}\big) = \{x\}$ for all $x \in X$.

Therefore, for any $A \subseteq X$, we have $\bigcup_{x \in A} U\big(\{x\}\big) = \bigcup_{x \in A} \{x\} = A$. Thus, by Theorem 5.9, U is union-preserving if and only if $U\,(A) = A$ for all $A \subseteq X$. That is, U is the identity corelation on X.

Remark 5.14 Note that here $\mathscr{C}$ may, in particular, be the family of all closed subsets of a T_1-space X, or the family of all convex subsets (linear subspaces) of a vector space X. Therefore, the most important closure operations fail to be union-preserving.

However, to clarify the importance of union-preserving corelations, in addition to Example 5.11, we can also state

Example 5.15 If U is a normal corelation on X to Y in the sense that U is V-normal for some corelation V on Y to X, then U is union-preserving.

This statement is a practically important particular case of [68, Corollary 7.2].

Remark 5.16 If U is a regular corelation on X onto Y in the sense that U is Φ-regular for some corelation Φ on X, then by [71, Theorem 7.5] the corelation U is already normal.

6 Some Pointwise Operations and an Inequality for Corelations

Here, to distinguish the pointwise complements and differences for corelations from the global ones, we shall use bold notations.

Definition 6.1 If U and V are corelations on X to Y, then for any $A \subseteq X$ we define

$$U^{\boldsymbol{c}}(A) = U(A)^{c} \qquad \text{and} \qquad \big(U \boldsymbol{\setminus} V\big)(A) = U(A) \setminus V\,(A).$$

Remark 6.2 Thus, if in particular $U(A) = Y$ for all $A \subseteq X$, then

$$(U \setminus V)(A) = U(A) \setminus V(A) = Y \setminus V(A) = V(A)^c = V^{\mathbf{c}}(A)$$

for all $A \subseteq X$. Therefore, in this particular case, we have $U \setminus V = V^{\mathbf{c}}$.

Moreover, to distinguish the pointwise intersections and unions for corelations from the global ones, we shall use lattice theoretic notations.

Definition 6.3 If $\mathcal{U}$ is a family of corelations on X to Y, then for any $A \subseteq X$ we define

$$\left(\bigwedge \mathcal{U}\right)(A) = \bigcap_{U \in \mathcal{U}} U(A) \qquad \text{and} \qquad \left(\bigvee \mathcal{U}\right)(A) = \bigcup_{U \in \mathcal{U}} U(A).$$

Remark 6.4 Thus, for any two corelations U and V on X to Y, we also write

$$U \wedge V = \bigwedge \{U,\ V\} \qquad \text{and} \qquad U \vee V = \bigvee \{U,\ V\}.$$

Now, by using the corresponding definitions and Theorem 5.9, we can easily prove the following two theorems.

Theorem 6.5 *If $\mathcal{U}$ is a family of increasing (quasi-increasing) corelations on X to Y, then $\bigwedge \mathcal{U}$ and $\bigvee \mathcal{U}$ are also increasing (quasi-increasing) corelations on X to Y.*

Theorem 6.6 *If $\mathcal{U}$ is a family of union-preserving corelations on X to Y, then $\bigvee \mathcal{U}$ is also an union-preserving corelations on X to Y.*

Proof Under the notation $V = \bigvee \mathcal{U}$, for any $A \subseteq X$ we have

$$V(A) = \left(\bigvee \mathcal{U}\right)(A) = \bigcup_{U \in \mathcal{U}} U(A).$$

Hence, by using that each member of $\mathcal{U}$ is union-preserving, we can see that

$$V(A) = \bigcup_{U \in \mathcal{U}} U(A) = \bigcup_{U \in \mathcal{U}} \bigcup_{x \in A} U\big(\{x\}\big) = \bigcup_{x \in A} \bigcup_{U \in \mathcal{U}} U\big(\{x\}\big) = \bigcup_{x \in A} V\big(\{x\}\big).$$

Therefore, by Theorem 5.9, V is also union-preserving.

The following example shows that the corresponding assertion need not be true for the corelation $\bigwedge \mathcal{U}$.

Example 6.7 Let X be a set such that $\operatorname{card}(X) > 1$, and for any $A \subseteq X$ define

$$U(A) = \Delta_X[\,A\,] \qquad \text{and} \qquad V(A) = \Delta_X^c\,[\,A\,].$$

Then, U and V are union-preserving corelations on X such that the corelation $U \wedge V$ is not union-preserving.

Namely, now we have $U\big(\{x\}\big) = \varDelta_X\big[\{x\}\big] = \{x\}$ and

$$V\big(\{x\}\big) = \varDelta_X^c\,[\,\{x\}\,] = \varDelta_X^c(x) = \varDelta_X(x)^c = \{x\}^c$$

for all $x \in X$.

Therefore, if $x_1,\ x_2 \in X$ such that $x_1 \neq x_2$, and $A = \{x_1,\ x_2\}$, then we can see that $U(A) = \varDelta_X[\,A\,] = A$ and

$$V(A) = \varDelta_X^c\,[\,A\,] = \varDelta_X^c\big(\{x_1,\ x_2\}\big) = \varDelta_X^c\big(\{x_1\}\big) \cup \varDelta_X^c\big(\{x_2\}\big) = \{x_1\}^c \cup \{x_2\}^c.$$

Hence, it is clear that $A = \{x_1,\ x_2\} \subseteq \{x_1\}^c \cup \{x_2\}^c = V(A)$, and thus

$$\big(U \wedge V\big)(A) = U(A) \cap V(A) = A \cap V(A) = A.$$

However, $\big(U \wedge V\big)\big(\{x_i\}\big) = U\big(\{x_i\}\big) \cap V\big(\{x_i\}\big) = \{x_i\} \cap \{x_i\}^c = \emptyset$ for $i = 1,\ 2$, and thus

$$\bigcup_{x \in A} \big(U \wedge V\big)\big(\{x\}\big) = \big(U \wedge V\big)\big(\{x_1\}\big) \cup \big(U \wedge V\big)\big(\{x_2\}\big) = \emptyset.$$

In the sequel, since set inclusion is not, in general, a convenient partial order for functions, we shall use the following

Definition 6.8 For any two sets X and Y, denote by $\mathscr{Q}(X, Y)$ the family of all corelations on X to Y.

Moreover, for any two $U,\ V \in \mathscr{Q}(X, Y)$, define $U \leq V$ if $U(A) \subseteq V(A)$ for all $A \subseteq X$.

Thus, we can easily prove the following

Theorem 6.9 *With the above inequality relation $\leq$, the family $\mathscr{Q}(X, Y)$ forms a complete poset.*

Proof It is clear that the relation $\leq$ considered in Definition 6.8 is a partial order (reflexive, transitive and antisymmetric) relation on $\mathscr{Q}(X, Y)$.

Moreover, if $\mathscr{U} \subseteq \mathscr{Q}(X, Y)$ and $V = \bigvee \mathscr{U}$, i. e.,

$$V(A) = \bigcup_{U \in \mathscr{U}} U(A)$$

for all $A \subseteq X$, then it can be easily seen that $V = \sup\big(\mathscr{U}\big)$. Thus, the poset $\mathscr{Q}(X, Y)$ is sup-complete.

The fact that $\mathscr{Q}(X, Y)$ is inf-complete can be proved quite similarly by showing that $\bigwedge \mathscr{U} = \inf\big(\mathscr{U}\big)$.

Remark 6.10 Note that, by a basic theorem of Birkhoff [1, p. 112], a poset is inf-complete if and only if it is sup-complete.

Moreover, by our former paper [3], this theorem can be extended to an arbitrary goset (generalized ordered set) even with a simpler proof.

Definition 6.11 In the sequel, the families of the quasi-increasing, increasing and union-preserving members of $\mathscr{Q}(X, Y)$ will be denoted by $\mathscr{Q}_1(X, Y)$, $\mathscr{Q}_2(X, Y)$ and $\mathscr{Q}_3(X, Y)$, respectively.

Remark 6.12 Thus, we evidently have

$$\mathscr{Q}_3(X, Y) \subseteq \mathscr{Q}_2(X, Y) \subseteq \mathscr{Q}_1(X, Y) \subseteq \mathscr{Q}(X, Y).$$

Moreover, in addition to Theorem 6.9, we can also prove the following

Theorem 6.13 *With the corresponding restriction of the inequality relation $\leq$ considered in Definition 6.8, the family $\mathscr{Q}_i(X, Y)$, with $i = 1, 2, 3$, is also a complete poset.*

Proof To prove this for $i = 3$, note that if $\mathscr{U} \subseteq \mathscr{Q}_3(X, Y)$, then by Theorem 6.9 there exists $V \in \mathscr{Q}(X, Y)$ such that $V = \sup(\mathscr{U})$.

Moreover, we necessarily have

$$V(A) = \bigcup_{U \in \mathscr{U}} U(A) = \Big(\bigvee \mathscr{U} \Big)(A)$$

for all $A \subseteq X$, and thus $V = \bigvee \mathscr{U}$.

Hence, by using Theorem 6.6, we can infer that $V \in \mathscr{Q}_3(X, Y)$. Therefore, $V = \sup \big(\mathscr{U}\big)$ is also true in $\mathscr{Q}_3(X, Y)$. Thus, $\mathscr{Q}_3(X, Y)$ is also sup-complete.

Remark 6.14 Now, by Remark 6.10, $\inf(\mathscr{U})$ also exists in $\mathscr{Q}_3(X, Y)$. However, because of Example 6.7, it can be strictly smaller than $\bigwedge \mathscr{U}$. Therefore, the latter notation may cause some confusions.

7 Increasingness and Union-Preservingness of Composite Corelations

In addition to Theorems 6.5 and 6.6, it is also worth proving the following

Theorem 7.1 *If U and V are increasing (union-preserving) corelations on X to Y and Y to Z, respectively, then $V \circ U$ is a increasing (union-preserving) corelation on X to Z.*

Proof If for instance U and V are union-preserving, then by the corresponding definitions we have

$$\big(V \circ U\big)(A) = V\big(U(A)\big) = V\Big(\bigcup_{x \in A} U\big(\{x\}\big)\Big) = \bigcup_{x \in A} V\big(U\big(\{x\}\big)\big) = \bigcup_{x \in A} \big(V \circ U\big)\big(\{x\}\big)$$

for all $A \subseteq X$. Therefore, by Theorem 5.9, the corelation $V \circ U$ is also union-preserving.

Remark 7.2 If U is quasi-increasing and V is increasing, then we can at once see that $V \circ U$ is also quasi-increasing.

However, if V is only quasi-increasing, then we cannot state that $V \circ U$ is quasi-increasing, even if U is a constant corelation.

From the above theorem, by induction, we can immediately derive

Corollary 7.3 *If U is an increasing (union-preserving) corelation on X, then U^n is also an increasing (union-preserving) corelation on X for all $n \in \mathbb{N}$.*

Here, to avoid confusion with the ordinary preorder hull $U^\infty = \bigcup_{n=o}^{\infty} U^n$ of a corelation U, we shall use bold notation.

Definition 7.4 For any corelation U on X, we define

$$U^\infty = \bigvee_{n=0}^{\infty} U^n,$$

where in contrast to our former notation U^0 is now to denote the identity corelation on X.

Remark 7.5 Thus, for any $A \subseteq X$, we have

$$U^\infty(A) = \Big(\bigvee_{n=0}^{\infty} U^n \Big)(A) = \bigcup_{n=0}^{\infty} U^n(A).$$

Now, concerning the corelation U^∞, we can easily prove the following theorems.

Theorem 7.6 *If U is a corelation on X, then U^∞ is an extensive corelation on X such that $U \leq U^\infty$.*

Proof Since $U^0(A) = A$ and $U^1(A) = U(A)$ for all $A \subseteq X$, from Remark 7.5, it is clear that $A \subseteq U^\infty(A)$ and $U(A) \subseteq U^\infty(A)$ for all $A \subseteq X$. Therefore, the required assertions are also true.

Theorem 7.7 *If U is an increasing (union-preserving) corelation on X, then U^∞ is also an increasing (union-preserving) corelation on X.*

Proof From Corollary 7.3, we know that the corelation U^n is also increasing (union-preserving) for all $n \in \mathbb{N}$. Hence, by Definition 7.4 and Theorems 6.5 and 6.6, it is clear that the required assertions are also true.

Theorem 7.8 *If U is an union-preserving corelation on X, then U^∞ is an idempotent corelation on X.*

Proof From Corollary 5.8, we know that U is increasing. Moreover, by Theorem 7.6, for any $A \subseteq X$, we have $A \subseteq U^\infty(A)$. Hence, by using Theorem 7.7, we can infer that

$$U^\infty(A) \subseteq U^\infty\big(U^\infty(A)\big) = \big(U^\infty \circ U^\infty\big)(A).$$

Moreover, by using Remark 7.5, we can see that

$$U^{\infty}\big(U^{n}(A)\big)=\bigcup_{k=0}^{\infty}U^{k}\big(U^{n}(A)\big)=\bigcup_{k=0}^{\infty}U^{k+n}(A)\subseteq\bigcup_{l=0}^{\infty}U^{l}(A)=U^{\infty}(A)$$

for all $n\in\{0\}\cup\mathbb{N}$. Hence, by using Theorem 7.7, we can already see that

$$\big(U^{\infty}\circ U^{\infty}\big)(A)=U^{\infty}\big(U^{\infty}(A)\big)=U^{\infty}\Big(\bigcup_{n=0}^{\infty}U^{n}(A)\Big)=\bigcup_{n=0}^{\infty}U^{\infty}\big(U^{n}(A)\big)\subseteq U^{\infty}(A)$$

also holds. Therefore, we actually have $\big(U^{\infty}\circ U^{\infty}\big)(A)=U^{\infty}(A)$ for all $A\subseteq X$, and thus $U^{\infty}\circ U^{\infty}=U^{\infty}$.

Now, as an immediate consequence of Corollary 5.8 and Theorems 7.6, 7.7 and 7.8, we can also state

Theorem 7.9 *If U is an union-preserving corelation on X, then U^{∞} is an union-preserving closure operation on X such that $U\leq U^{\infty}$.*

In addition to this theorem, it is also worth proving the following

Theorem 7.10 *If U is an increasing and V is an extensive corelation on X such that $U\leq V$ and $V^{2}\leq V$, then $U^{\infty}\leq V$.*

Proof By the above assumptions, for any $A\subseteq X$, we have $U^{0}(A)\subseteq V(A)$ and $U(A)\subseteq V(A)$. Moreover, we also have

$$U^{2}(A)=U\big(U(A)\big)\leq U\big(V(A)\big)\subseteq V\big(V(A)\big)\subseteq V^{2}(A)\subseteq V(A).$$

Hence, by induction, it is clear that $U^{n}(A)\subseteq V(A)$ for all $n\in\{0\}\cup\mathbb{N}$. Therefore, by Remark 7.5,

$$U^{\infty}(A)=\bigcup_{n=0}^{\infty}U^{n}(A)\subseteq V(A),$$

and thus $U^{\infty}\leq V$ also holds.

Finally, we note that, in particular, the following theorem is also true.

Theorem 7.11 *If R is a relation on X, and $U(A)=R[A]$ for all $A\subseteq X$, then $U^{\infty}(A)=R^{\infty}[A]$ for all $A\subseteq X$.*

Proof By the corresponding definitions, for any $A\subseteq X$, we have $U^{0}(A)=R^{0}[A]$ and $U(A)=R[A]$. Moreover, by Theorem 3.8, we also have

$$U^{2}(A)=U\big(U(A)\big)=R\big[R[A]\big]=R^{2}[A].$$

Hence, by induction, it is quite obvious that we also have $U^{n}(A)=R^{n}[A]$ for all $n\in\{0\}\cup\mathbb{N}$. Therefore, by Remark 7.5 and Theorem 4.6, we can also state that

$$U^{\infty}(A) = \bigcup_{n=0}^{\infty} U^{n}(A) = \bigcup_{n=0}^{\infty} R^{n}(A) = \Big(\bigcup_{n=0}^{\infty} R^{n} \Big)(A) = R^{\infty}[\, A \,].$$

Remark 7.12 In the next section, we shall show that a corelation U on X is union-preserving if and only if there exists a relation R on X such that $U\,(A) = R\,[A]$ for all $A \subseteq X$. Therefore, Theorem 7.8 can be proved with the help of Theorem 7.11 too.

8 A Partial Galois Connection Between Relations and Corelations

According to the corresponding definitions of Höhle and Kubiak [25] and the notations of Davey and Priestley [12, p. 55], we may also naturally introduce

Definition 8.1 For any relation R and corelation U on X to Y, we define a corelation $R^{\triangleright}$ and a relation $U^{\triangleleft}$ on X to Y such that

$$R^{\triangleright}(A) = R\,[\, A \,] \qquad \text{and} \qquad U^{\triangleleft}(x) = U\big(\{x\}\big)$$

for all $A \subseteq X$ and $x \in X$.

Thus, we can easily prove the following two theorems.

Theorem 8.2 *If U is a corelation on X to Y, then*

$$R^{\triangleright} \leq U \quad \Longrightarrow \quad R \subseteq U^{\triangleleft}$$

for any relation R on X to Y.

Proof If $R^{\triangleright} \leq U$, then in particular we have

$$R\,(x) = R\,\big[\{x\}\big] = R^{\triangleright}\big(\{x\}\big) \subseteq U\big(\{x\}\big) = U^{\triangleleft}(x)$$

for all $x \in X$. Therefore, by Theorem 3.1, $R \subseteq U^{\triangleleft}$ also holds.

Theorem 8.3 *For a corelation U on X to Y, the following assertions are equivalent:*

(1) U is quasi-increasing;
(2) $R \subseteq U^{\triangleleft} \implies R^{\triangleright} \leq U$ for any relation R on X to Y.

Proof If (1) holds and $R \subseteq U^{\triangleleft}$, then

$$R^{\triangleright}(A) = R\,[\, A \,] = \bigcup_{x \in A} R\,(x) \subseteq \bigcup_{x \in A} U^{\triangleleft}(x) = \bigcup_{x \in A} U\big(\{x\}\big) \subseteq U\,(A)$$

for all $A \subseteq X$. Therefore, by Definition 6.8, $R^{\triangleright} \leq U$, and thus (2) also holds.

While, if (2) holds, then because of $U^{\lhd} \subseteq U^{\lhd}$ we have $U^{\lhd\rhd} = \big(U^{\lhd}\big)^{\rhd} \leq U$. Therefore, for any $A \subseteq X$, we have $U^{\lhd\rhd}(A) \subseteq U(A)$. Moreover, by using the corresponding definitions, we can see that

$$U^{\lhd\rhd}(A) = \big(U^{\lhd}\big)^{\rhd}(A) = U^{\lhd}[\,A\,] = \bigcup_{x \in A} U^{\lhd}(x) = \bigcup_{x \in A} U\big(\{x\}\big).$$

Therefore, $\bigcup_{x \in A} U\big(\{x\}\big) \subseteq U(A)$, and thus, by Theorem 5.5, assertion (1) also holds.

Now, as an immediate consequence of the above two theorems, we can also state

Corollary 8.4 *For an arbitrary relation R and a quasi-increasing corelation U on X to Y, we have*

$$R^{\rhd} \leq U \quad \Longleftrightarrow \quad R \subseteq U^{\lhd}.$$

Remark 8.5 This corollary shows that the operation $\rhd$ and the restriction of $\lhd$ to $\mathscr{Q}_1(X, Y)$ establish a Galois connection between the complete posets $\mathscr{P}(X \times Y)$ and $\mathscr{Q}_1(X, Y)$.

Therefore, the extensive theory of Galois connections could be applied here. However, because of the simplicity of Definition 8.1, it seems now more convenient to use some, more elementary, direct proofs.

For instance, by the corresponding definitions, we evidently have the following

Theorem 8.6 *For any two relations R, S and corelations U, V on X to Y,*

(1) $R \subseteq S \implies R^{\rhd} \leq S^{\rhd}$; *(2)* $U \leq V \implies U^{\lhd} \subseteq V^{\lhd}$.

Remark 8.7 By using Corollary 8.4, instead of (2) we could only prove that the restriction of the function $\lhd$ to $\mathscr{Q}_1(X, Y)$ is increasing.

Moreover, we can also easily prove the following theorem whose first statement has also been established by Höhle and Kubiak [25].

Theorem 8.8 *For any two relations R and S on X to Y,*

(1) $R^{\rhd\lhd} = R$; *(2)* $R^{\rhd} \leq S^{\rhd} \implies R \subseteq S$.

Proof By the corresponding definitions, we have

$$R^{\rhd\lhd}(x) = \big(R^{\rhd}\big)^{\lhd}(x) = R^{\rhd}\big(\{x\}\big) = R\big[\,\{x\}\,\big] = R(x)$$

for all $x \in X$. Therefore, (1) is also true.

To prove (2), note that if $R^{\rhd} \leq S^{\rhd}$ holds, then by Theorem 8.6 we also have $R^{\rhd\lhd} \subseteq S^{\rhd\lhd}$. Hence, by using (1), we can see that $R \subseteq S$ also holds.

Remark 8.9 From this theorem, we can see that $\rhd$ is an injective function of $\mathscr{P}(X \times Y)$ to $\mathscr{Q}(X, Y)$.

Moreover, $\lhd$ maps a part of $\mathscr{Q}(X,\ Y)$ onto $\mathscr{P}(X \times Y)$. And, the composition $\rhd \lhd$ is the identity function of $\mathscr{P}(X \times Y)$.

Now, as an immediate consequence of Theorems 8.6 and 8.8, we can also state

Corollary 8.10 *For any two relations R and S on X to Y, we have*

$$R \subseteq S \quad \Longleftrightarrow \quad R^{\rhd} \leq S^{\rhd}.$$

Concerning the operation $\lhd \rhd$, we can only prove the following theorem which, to some extent, has also been established by Höhle and Kubiak [25] and Pataki [43].

Theorem 8.11 *For a corelation U on X to Y, the following assertions are equivalent:*

(1) $U^{\lhd \rhd} = U$;

(2) U *is union-preserving;*

(3) $U = R^{\rhd}$ *for some relation R on X to Y.*

Proof From the proof of Theorem 8.3, we know that

$$U^{\lhd \rhd}(A) = \bigcup_{x \in A} U\big(\{x\}\big)$$

for all $A \subseteq X$. Moreover, if (2) holds, then by Theorem 5.9 we have

$$U\,(A) = \bigcup_{x \in A} U\big(\{x\}$$

for all $A \subseteq X$. Therefore, $U^{\lhd \rhd}(A) = U\,(A)$ for all $A \subseteq X$, and thus (1) also holds.

Now, since (1) trivially implies (3), we need only note that if (3) holds, then

$$U(A) = R^{\rhd}(A) = R\,[\,A\,] = \bigcup_{x \in A} R\,(x) = \bigcup_{x \in A} R\big[\{x\}\big] = \bigcup_{x \in A} R^{\rhd}\big(\{x\}\big) = \bigcup_{x \in A} U\big(\{x\}\big)$$

for all $A \subseteq X$. Therefore, by Theorem 5.9, assertion (2) also holds.

Remark 8.12 From this theorem, we can see that the function $\rhd$ maps $\mathscr{P}(X \times Y)$ onto $\mathscr{Q}_3(X, Y)$.

Moreover, the restriction of $\lhd$ to $\mathscr{Q}_3(X, Y)$ is injective. And the restriction of $\lhd \rhd$ to $\mathscr{Q}_3(X, Y)$ is the identity function of $\mathscr{Q}_3(X, Y)$.

Now, as an immediate consequence of Theorems 8.6 and 8.11, we can also state

Corollary 8.13 *For any two union-preserving corelations U and V on X to Y, we have*

$$U \leq V \quad \Longleftrightarrow \quad U^{\lhd} \subseteq V^{\lhd}.$$

Remark 8.14 Note that, by Remarks 8.9 and 8.12, the Galois connection mentioned in Remark 8.5 is rather particular.

Therefore, it not surprising that, by Corollaries 8.10 and 8.13, the functions $\rhd$ and $\lhd$ with the corresponding identity functions form Pataki connections.

9 The Galois Interior of a Corelation

Definition 9.1 For any corelation U on X to Y, the corelation

$$U^{\circ} = U^{\lhd \rhd}$$

will be called the *Galois interior* of U.

Thus, by Theorem 8.11, we evidently have the following

Theorem 9.2 *If U is a corelation on X to Y, then U° is a union-preserving corelation on X to Y.*

Moreover, by using Theorem 8.8, we can easily establish the following

Theorem 9.3 *For any relation R and corelation U on X to Y, we have*

(1) $R^{\rhd \circ} = R^{\rhd}$; *(2)* $U^{\circ \lhd} = U^{\lhd}$.

Furthermore, by the proof of Theorem 8.3, we can also state the following

Theorem 9.4 *If U is a corelation on X to Y, then for any $A \subseteq X$, we have*

$$U^{\circ}\big(A\big) = \bigcup_{x \in A} U\big(\{x\}\big).$$

Example 9.5 If U is the complementation operation on X, then for any $A \subseteq X$ we have

$$U^{\circ}(A) = \begin{cases} \emptyset & \text{if} \quad \operatorname{card}(A) = 0, \\ A^{c} & \text{if} \quad \operatorname{card}(A) = 1, \\ X & \text{if} \quad \operatorname{card}(A) > 1. \end{cases}$$

Namely, by Theorem 9.4 and De Morgan's law, we have

$$U^{\circ}\big(A\big) = \bigcup_{x \in A} U\big(\{x\}\big) = \bigcup_{x \in A} \{x\}^{c} = \Big(\bigcap_{x \in A} \{x\} \Big)^{c},$$

whence the required equalities immediately follow.

Now, in addition to Theorem 8.11, we can also easily prove the following

Theorem 9.6 *For any corelation U on X to Y, we have*

(1) $U^{\circ\circ} = U^{\circ}$;

(2) $U^{\circ} \leq U \iff U$ *is quasi-increasing;*

(3) $U^{\circ} = U \iff U$ *is union-preserving.*

Proof Assertion (3) follows from Theorem 8.11 by Definition 9.1. Moreover, by Theorem 9.4, it is clear that, for any $A \subseteq X$, we have

$$U^{\circ}(A) \subseteq U\,(A) \iff \bigcup_{x \in A} U\big(\{x\}\big) \subseteq U\,(A).$$

Hence, by Definition 6.8 and Theorem 5.5, it is clear that (2) is true.

Furthermore, by using Theorem 9.4, we can also see that

$$U^{\circ\circ}(A) = \bigcup_{x \in A} U^{\circ}\big(\{x\}\big) = \bigcup_{x \in A} U\big(\{x\}\big) = U^{\circ}(A)$$

for all $A \subseteq X$. Therefore, (1) is also true.

Remark 9.7 From the above theorem, we can see that the function $\circ$ is a modification operation on $\mathscr{Q}(X,\,Y)$ such that its restriction to $\mathscr{Q}_1(X,\,Y)$ is an interior operation. Moreover, $\mathscr{Q}_3(X,\,Y)$ is the family of all open elements of $\mathscr{Q}_1(X,\,Y)$.

Theorem 9.8 *For any two corelations U and V on X to Y, we have*

(1) $U \leq V \implies U^{\circ} \leq V^{\circ}$;

(2) $U^{\circ} \leq V \implies U^{\lhd} \subseteq V^{\lhd}$;

(3) $U^{\lhd} \subseteq V^{\lhd} \implies U^{\circ} \leq V$ *if V is quasi-increasing.*

Proof Assertion (1) follows from Theorem 8.6 by Definition 9.1. Moreover, by Definition 9.1 and Theorem 8.2, it is clear that

$$U^{\circ} \leq V \implies \big(U^{\lhd}\big)^{\rhd} \leq V \implies U^{\lhd} \leq V^{\lhd}.$$

Therefore, (2) is true.

While, if V is quasi-increasing, then by using Theorem 8.3, we can quite similarly see that

$$U^{\lhd} \subseteq V^{\lhd} \implies \big(U^{\lhd}\big)^{\rhd} \leq V \implies U^{\circ} \leq V.$$

Therefore, (3) is also true.

Remark 9.9 The above theorem shows that the functions $\lhd$ and $\circ$ establish a Pataki connection between $\mathscr{Q}_1(X,\,Y)$ and $\mathscr{P}(X \times Y)$.

Now, by using our former results, we can also prove the following two theorems.

Theorem 9.10 *If R is a relation on X to Y and $U = R^{\rhd}$, then*

(1) *U is an union-preserving corelation on X to Y such that* $U^{\lhd} = R$;

(2) U is the smallest quasi-increasing corelation on X to Y such that $R \subseteq U^{\lhd}$;

(3) U is the largest union-preserving corelation on X to Y such that $U^{\lhd} \subseteq R$.

Proof Theorems 8.11 and 8.8 show that U is union-preserving and $U^{\lhd} = R^{\rhd\lhd} = R$. Therefore, (1) is true.

Moreover, if V is a quasi-increasing corelation on X to Y such that $R \subseteq V^{\lhd}$, then by Theorem 8.3 we also have $R^{\rhd} \leq V$, and thus $U \leq V$. Therefore, by (1), assertion (2) is also true.

While, if V is a union-preserving corelation on X to Y such that $V^{\lhd} \subseteq R$, then by Theorems 8.6 and 8.11 we also have $V^{\lhd\rhd} \leq R^{\rhd}$, and thus $V \leq U$. Therefore, by (1), assertion (3) is also true.

Theorem 9.11 *If U is a corelation on X to Y and $R = U^{\lhd}$, then*

(1) $R^{\rhd} \leq U \iff U$ is quasi-increasing;

(2) $R^{\rhd} = U \iff U$ is union-preserving;

(3) if U is quasi-increasing, then R the largest relation on X to Y such that $R^{\rhd} \leq U$;

(4) if U is union-preserving, then R is the smallest relation on X to Y such that $U \leq R^{\rhd}$.

Proof By the corresponding definitions and Theorem 9.6, we have

$$R^{\rhd} \leq U \iff U^{\lhd\rhd} \leq U \iff U^{\circ} \leq U \iff U \text{ is quasi-increasing}$$

and

$$R^{\rhd} = U \iff U^{\lhd\rhd} = U \iff U^{\circ} = U \iff U \text{ is union-preserving.}$$

Therefore, (1) and (2) are true.

On the other hand, if S is a relation on X to Y such that $S^{\rhd} \leq U$, then by Theorem 8.2 we also have $S \subseteq U^{\lhd}$, and thus $S \subseteq R$. Therefore, by (1), assertion (3) is also true.

While, if S is a relation on X to Y such that $U \leq S^{\rhd}$, then by Theorem 8.6, we also have $U^{\lhd} \subseteq S^{\rhd\lhd}$. Hence, by Theorem 8.8, we can see that $R \subseteq S$. Therefore, by (2), assertion (4) is also true.

Remark 9.12 In addition to Definition 8.1 and Theorem 9.11, it is also worth mentioning that if R is relation and U is a corelation on X to Y, then by the corresponding definitions we have

(1) $R^{\rhd}(A) = \mathrm{cl}_{R^{-1}}(A)$ for all $A \subseteq X$;

(2) $R^{\rhd} \leq U \iff A \in \mathrm{Int}_R\big(U(A)\big)$ for all $A \subseteq X$.

Moreover, by using Theorems 9.11 and 4.6, it can be easily seen that if U is quasi-increasing, then under the notation $\mathrm{Int}_{\rhd}(U) = \big\{S \subseteq X \times Y : \quad S^{\rhd} \leq U\big\}$ we have $U^{\lhd} = \max\,(\mathrm{Int}_{\rhd}(U)) = \bigcup \mathrm{Int}_{\rhd}(U)$.

10 Compatibility of $\vartriangleright$ and $\vartriangleleft$ with Set Theoretic Operations

By using the corresponding definitions and some former theorems on relations, we can easily prove the following five theorems.

Theorem 10.1 *If R is a relation on X to Y, then for any $A,\ B \subseteq X$ we have*

(1) $R^{\vartriangleright}(A) \setminus R^{\vartriangleright}(B) \subseteq R^{\vartriangleright}(\,A \setminus B\,)$; *(2)* $R^{\vartriangleright\boldsymbol{c}}(A) \subseteq R^{\vartriangleright}\big(A^c\big)$ *if* $Y = R\,[X]$.

Proof To check (2), note that if $Y = R\,[X]$, then by Theorem 4.1 we have

$$R^{\vartriangleright\boldsymbol{c}}(A) = R^{\vartriangleright}(A)^c = R\,[\,A\,]^c \subseteq R\,[\,A^c\,] = R^{\vartriangleright}\big(A^c\big).$$

Theorem 10.2 *If R is a relation on X to Y, then for any family $\mathscr{A}$ of subsets of X we have*

(1) $R^{\vartriangleright}\big(\bigcap \mathscr{A}\,\big) \subseteq \bigcap\limits_{A \in \mathscr{A}} R^{\vartriangleright}\big(A\big)$; *(2)* $R^{\vartriangleright}\big(\bigcup \mathscr{A}\,\big) = \bigcup\limits_{A \in \mathscr{A}} R^{\vartriangleright}\big(A\big)$.

Proof To check (1) note that, by Theorem 4.2, we have

$$R^{\vartriangleright}\big(\bigcap \mathscr{A}\,\big) = R\Big[\bigcap \mathscr{A}\,\Big] \subseteq \bigcap_{A \in \mathscr{A}} R\,[\,A\,] = \bigcap_{A \in \mathscr{A}} R^{\vartriangleright}(A).$$

Remark 10.3 If in particular R^{-1} is a function, then by Remark 4.3 the corresponding equalities are also true in the above two theorems.

Theorem 10.4 *For any two relations R and S on X to Y, we have*

(1) $R^{\vartriangleright} \setminus S^{\vartriangleright} \leq (R \setminus S)^{\vartriangleright}$; *(2)* $R^{\vartriangleright\boldsymbol{c}}(A) \subseteq R^{c\,\vartriangleright}(A)$ *if* $\emptyset \neq A \subseteq X$.

Proof To check (1), note that by Theorem 4.5, for any $A \subseteq X$, we have

$$\big(R^{\vartriangleright} \setminus S^{\vartriangleright}\big)(A) = R^{\vartriangleright}(A) \setminus S^{\vartriangleright}(A) = R\,[\,A\,] \setminus S\,[\,A\,] \subseteq (R \setminus S)\,[\,A\,] = (R \setminus S)^{\vartriangleright}(A).$$

Theorem 10.5 *For any family $\mathscr{R}$ of relations on X to Y, we have*

(1) $\big(\bigcap \mathscr{R}\big)^{\vartriangleright} \leq \bigwedge\limits_{R \in \mathscr{R}} R^{\vartriangleright}$; *(2)* $\big(\bigcup \mathscr{R}\big)^{\vartriangleright} = \bigvee\limits_{R \in \mathscr{R}} R^{\vartriangleright}$.

Proof To check (1), note that by Theorem 4.6, for any $A \subseteq X$, we have

$$\big(\bigcap \mathscr{R}\big)^{\vartriangleright}(A) = \big(\bigcap \mathscr{R}\big)\,[\,A\,] \subseteq \bigcap_{R \in \mathscr{R}} R\,[\,A\,] = \bigcap_{R \in \mathscr{R}} R^{\vartriangleright}(A) = \Big(\bigwedge_{R \in \mathscr{R}} R^{\vartriangleright}\Big)(A).$$

Theorem 10.6 *If R is a relation on X to Y, then for any $A \subseteq X$ we have*

$$R^{c\,\vartriangleright\boldsymbol{c}}(A) = \bigcap_{x \in A} R\,(x).$$

Proof By Theorem 4.9, we have

$$R^{c\,\vartriangleright\boldsymbol{c}}(A) = R^{c\,\vartriangleright}(A)^c = R^c[\,A\,]^c = \bigcap_{x \in A} R\,(x).$$

Moreover, by the corresponding definitions and some former theorems on relations, it is clear that we also have the following five theorems.

Theorem 10.7 *If U is a corelation on X to Y, then for any $A,\ B \subseteq X$ we have*

(1) $U^{\triangleleft}[\,A\,] \setminus U^{\triangleleft}[\,B\,] \subseteq U^{\triangleleft}[\,A \setminus B\,]$; *(2)* $U^{\triangleleft}[\,A\,]^{c} \subseteq U^{\triangleleft}[\,A^{c}]$ *if* $Y = U^{\triangleleft}[X]$.

Theorem 10.8 *If U is a corelation on X to Y, then for any family $\mathscr{A}$ of subsets of X we have*

(1) $U^{\triangleleft}\big[\bigcap \mathscr{A}\,\big] \subseteq \bigcap\limits_{A\in\mathscr{A}} U^{\triangleleft}[\,A\,]$; *(2)* $U^{\triangleleft}\big[\bigcup \mathscr{A}\,\big] = \bigcup\limits_{A\in\mathscr{A}} U^{\triangleleft}[\,A\,]$.

Remark 10.9 If in particular $U^{\triangleleft-1}$ is a function, then by Remark 4.3 the corresponding equalities are also true in the latter two theorems.

Theorem 10.10 *For any two corelations U and V on X to Y, we have*

(1) $U^{\triangleleft} \setminus V^{\triangleleft} = (U \setminus V)^{\triangleleft}$; *(2)* $U^{\triangleleft}[\,A\,]^{c} \subseteq U^{\triangleleft\boldsymbol{c}}[\,A\,]$ *if* $\emptyset \neq A \subseteq X$.

Proof To check (1), note that by Theorem 4.5 and Remark 4.7, for any $x \in X$, we have

$$\big(U^{\triangleleft} \setminus V^{\triangleleft}\big)(x) = U^{\triangleleft}(x) \setminus V^{\triangleleft}(x) = U\big(\{x\}\big) \setminus V\big(\{x\}\big) = (U \setminus V)\big(\{x\}\big) = (U \setminus V)^{\triangleleft}(x).$$

Theorem 10.11 *For any family $\mathscr{U}$ of corelations on X to Y, we have*

(1) $\big(\bigwedge \mathscr{U}\,\big)^{\triangleleft} = \bigcap\limits_{U\in\mathscr{U}} U^{\triangleleft}$; *(2)* $\big(\bigvee \mathscr{U}\,\big)^{\triangleleft} = \bigcup\limits_{U\in\mathscr{U}} U^{\triangleleft}$.

Proof To check (1), note that, by Theorem 4.6 and Remark 4.7, for any $x \in X$, we have

$$\big(\bigwedge \mathscr{U}\,\big)^{\triangleleft}(x) = \big(\bigwedge \mathscr{U}\,\big)\big(\{x\}\big) = \bigcap_{U\in\mathscr{U}} U\big(\{x\}\big) = \bigcap_{U\in\mathscr{U}} U^{\triangleleft}(x) = \Big(\bigcap_{U\in\mathscr{U}} U^{\triangleleft}\Big)(x).$$

Theorem 10.12 *If U is a corelation on X to Y, then for any $A \subseteq X$ we have*

$$U^{\triangleleft c}[\,A\,]^{c} = \bigcap_{x\in A} U\,\big(\{x\}\big).$$

Finally, we note that, from Theorems 10.1 and 10.2, by writing $U^{\triangleleft}$ in place of R, we can obtain the following two theorems.

Theorem 10.13 *If U is a corelation on X to Y, then for any $A,\ B \subseteq X$ we have*

(1) $U^{\circ}(A) \setminus U^{\circ}(B) \subseteq U^{\circ}(\,A \setminus B\,)$; *(2)* $U^{\circ\boldsymbol{c}}(A) \subseteq U^{\circ}(\,A^{c})$ *if* $Y = U^{\triangleleft}[X]$.

Theorem 10.14 *If U is a corelation on X to Y, then for any family $\mathscr{A}$ of subsets of X we have*

(1) $U^{\circ}\big(\bigcap \mathscr{A}\,\big) \subseteq \bigcap\limits_{A\in\mathscr{A}} U^{\circ}\big(A\big)$; *(2)* $U^{\circ}\big(\bigcup \mathscr{A}\,\big) = \bigcup\limits_{A\in\mathscr{A}} U^{\circ}\big(A\big)$.

Remark 10.15 While, from Theorems 10.4 and 10.5, by writing $U^{\triangleleft}$ and $V^{\triangleleft}$ in place of R and S, respectively, we can only get some less convenient theorems.

However, it is now more important to note that, by using Theorems 10.11 and 10.5, we can also prove the following

Theorem 10.16 *For any family $\mathcal{U}$ of corelations on X to Y, we have*

(2) $\big(\bigwedge \mathcal{U}\big)^{\circ} \leq \bigwedge_{U\in\mathcal{U}} U^{\circ}$; *(2)* $\big(\bigvee \mathcal{U}\big)^{\circ} = \bigvee_{U\in\mathcal{U}} U^{\circ}$.

Proof To check (1), note that, by Theorems 10.11 and 10.5, we have

$$\big(\bigwedge \mathcal{U}\big)^{\circ} = \big(\bigwedge \mathcal{U}\big)^{\triangleleft\triangleright} = \Big(\bigcap_{U\in\mathcal{U}} U^{\triangleleft}\Big)^{\triangleright} \leq \bigwedge_{U\in\mathcal{U}} U^{\triangleleft\triangleright} = \bigwedge_{U\in\mathcal{U}} U^{\circ}$$

11 Compatibility of $\triangleright$ and $\triangleleft$ with the Composition of Relations

The following theorem has also been established by Höhle and Kubiak [25].

Theorem 11.1 *For any two relations R on X to Y and S on Y to Z, we have*

$$\big(S \circ R\big)^{\triangleright} = S^{\triangleright} \circ R^{\triangleright}.$$

Proof By the corresponding definitions and Theorem 3.8, it is clear that

$$\big(S \circ R\big)^{\triangleright}(A) = \big(S \circ R\big)[\,A\,] = S\big[\,R[\,A\,]\,\big] = S^{\triangleright}\,\big(R^{\triangleright}(A)\big) = \big(S^{\triangleright} \circ R^{\triangleright}\big)\,(A)$$

for all $A \subseteq X$. Therefore, the required equality is also true.

From this theorem, by using Theorem 8.11, we can immediately derive the following two corollaries.

Corollary 11.2 *For an arbitrary relation on R on X to Y and a union-preserving corelation V on Y to Z, we have*

$$\big(V^{\triangleleft} \circ R\big)^{\triangleright} = V \circ R^{\triangleright}.$$

Corollary 11.3 *For a union-preserving corelation U on X to Y and an arbitrary relation S on Y to Z, we have*

$$\big(S \circ U^{\triangleleft}\big)^{\triangleright} = S^{\triangleright} \circ U.$$

By Theorem 4.18, in addition to Theorem 11.1, we can also state the following

Theorem 11.4 *If R is a relation on X to Y and S is a relation on Z to W, then for any $A \subseteq X \times Z$ we have*

$$\big(R \boxtimes S\big)^{\triangleright}(A) = S \circ A \circ R^{-1}.$$

Moreover, by using Theorems 10.5 and 11.1, we can also prove the following

Theorem 11.5 *For any relation R on X, we have*

$$R^{\infty\,\triangleright} = R^{\triangleright\,\infty}.$$

Proof From Theorem 11.1, by induction, we can see that $R^{n\triangleright} = R^{\triangleright n}$ for all $n \in \mathbb{N}$. Moreover, we can also note that $R^{0\,\triangleright} = R^{\triangleright 0}$.

Hence, by the corresponding definitions and Theorem 10.5, it is clear that

$$R^{\infty\,\triangleright} = \Big(\bigcup_{n=0}^{\infty} R^{n}\Big)^{\triangleright} = \bigvee_{n=0}^{\infty} R^{n\triangleright} = \bigvee_{n=0}^{\infty} R^{\triangleright n} = R^{\triangleright\,\infty}.$$

In addition to Theorem 11.1, we can also easily prove the following correction of a false statement of Höhle and Kubiak [25].

Theorem 11.6 *For an arbitrary corelation U on X to Y and a union-preserving corelation V on Y to Z, we have*

$$\big(V \circ U\big)^{\triangleleft} = V^{\triangleleft} \circ U^{\triangleleft}.$$

Proof By the corresponding definitions and Theorem 8.11, we have

$$\begin{aligned}(V \circ U)^{\triangleleft}(x) &= (V \circ U)\big(\{x\}\big) = V\,\Big(U\big(\{x\}\big)\Big) \\ &= V\,\big(U^{\triangleleft}(x)\big) = V^{\triangleleft\triangleright}\big(U^{\triangleleft}(x)\big) = V^{\triangleleft}\big[\,U^{\triangleleft}(x)\big] = \big(V^{\triangleleft} \circ U^{\triangleleft}\big)(x)\end{aligned}$$

for all $x \in X$. Therefore, the required equality is also true.

From this theorem, by using Theorems 8.11 and 8.8, we can immediately derive the following two corollaries.

Corollary 11.7 *For a corelation U on X to Y and a relation S on Y to Z, we have*

$$\big(S^{\triangleright} \circ U\big)^{\triangleleft} = S \circ U^{\triangleleft}.$$

Corollary 11.8 *For an arbitrary relation on R on X to Y and a union-preserving corelation V on Y to Z, we have*

$$\big(V \circ R^{\triangleright}\big)^{\triangleleft} = V^{\triangleleft} \circ R.$$

Now, by using Theorems 11.6 and 11.1, we can also easily prove the following

Theorem 11.9 *For an arbitrary corelation U on X to Y and a union-preserving corelation V on Y to Z, we have*

$$\big(V \circ U \big)^{\circ} = V \circ U^{\circ}.$$

Proof By Theorems 11.6, 11.1 and 9.6, we have

$$\big(V \circ U \big)^{\circ} = \big(V \circ U \big)^{\triangleleft \triangleright} = \big(V^{\triangleleft} \circ U^{\triangleleft} \big)^{\triangleright} = V^{\triangleleft \triangleright} \circ U^{\triangleleft \triangleright} = V^{\circ} \circ U^{\circ} = V \circ U^{\circ}.$$

Moreover, by using Theorems 11.6 and 10.11, we can also prove the following

Theorem 11.10 *For any union-preserving corelation U on X, we have*

$$U^{\infty \triangleleft} = U^{\triangleleft \infty}.$$

Proof From Theorem 11.6, by induction, we can see that $U^{n \triangleleft} = U^{\triangleleft n}$ for all $n \in \mathbb{N}$. Moreover, we can also note that $U^{0 \triangleleft} = U^{\triangleleft 0}$.

Hence, by the corresponding definitions and Theorem 10.11, it is clear that

$$U^{\infty \triangleleft} = \Big(\bigvee_{n=0}^{\infty} U^{n} \Big)^{\triangleleft} = \bigcup_{n=0}^{\infty} U^{n \triangleleft} = \bigcup_{n=0}^{\infty} U^{\triangleleft n} = U^{\triangleleft \infty}.$$

Remark 11.11 In the next section, we shall see that in connection with the usual inversion of relations we cannot prove such compatibility theorems.

12 Incompatibility of $\triangleright$ and $\triangleleft$ with the Inversion of Relations

Theorem 12.1 *For a relation R on X to Y, the following assertions are equivalent:*

(1) R is a function on X onto Y;
(2) $R \circ R^{-1} = \Delta_{Y}$; *(3) $R^{-1 \triangleright} \subseteq R^{\triangleright -1}$.*

Proof By Corollary 3.16, R is a function if and only if $R \circ R^{-1} = \Delta_{R[X]}$. Hence, since R is onto Y if and only if $R\,[X] = Y$, it is clear that (1) and (2) are equivalent.

Next, we show that (2) implies (3). For this, note that if $A \subseteq X$ and $B \subseteq Y$ such that $A = R^{-1 \triangleright}(B)$, then by Definition 8.1 and Theorem 3.8 we also have

$$R^{\triangleright}(A) = R\,[\,A\,] = R\,\big[\,R^{-1 \triangleright}(B)\,\big] = R\,\big[\,R^{-1}[\,B\,]\,\big] = \big(R \circ R^{-1} \big)\,[\,B\,].$$

Hence, if (2) holds, we can infer that

$$R^{\triangleright}(A) = \Delta_{Y}[\,B\,] = B, \qquad \text{and thus} \qquad A \in R^{\triangleright -1}[\,B\,].$$

Therefore, (3) also holds.

Now, to complete the proof, it remains to show that (3) also implies (2). For this, note that if (3) holds, then in particular, for any $y \in Y$, we have

$$R^{-1}(y) = R^{-1}\big[\{y\}\big] = R^{-1\triangleright}\big(\{y\}\big) \in R^{\triangleright-1}\big(\{y\}\big).$$

Hence, we can infer that

$$\{y\} = R^{\triangleright}\big(R^{-1}(y)\big) = R\big[R^{-1}(y)\big], \qquad \text{and thus} \qquad \Delta_Y(y) = \big(R \circ R^{-1}\big)(y).$$

Therefore, by Corollary 3.2, assertion (2) also holds.

From this theorem, by writing R^{-1} in place of R, we can immediately derive

Theorem 12.2 *For a relation R on X to Y, the following assertions are equivalent:*

(1) R^{-1} *is a function on Y onto X;*
(2) $R^{-1} \circ R = \Delta_X$; *(3)* $R^{\triangleright-1} \subseteq R^{-1\triangleright}$.

Proof To derive this from Theorem 12.1, note that

$$R^{\triangleright-1} \subseteq R^{-1\triangleright} \iff R^{\triangleright} \subseteq R^{-1\triangleright-1} \iff \big(R^{-1}\big)^{-1\triangleright} \subseteq \big(R^{-1}\big)^{\triangleright-1}.$$

Now, as an immediate consequence of the above two theorems, we can also state

Corollary 12.3 *For a relation R on X to Y, the following assertions are equivalent:*

(1) R *is an injective function of X onto Y;*
(2) $R^{\triangleright-1} = R^{-1\triangleright}$; *(3)* $R^{-1} \circ R = \Delta_X$ *and* $R \circ R^{-1} = \Delta_Y$.

From the above results, by writing $U^{\triangleleft}$ in place of R, and using that $U^{\circ} = U^{\triangleleft\triangleright}$, we can immediately derive the following assertions.

Theorem 12.4 *For a corelation U on X to Y following assertions are equivalent:*

(1) $U^{\triangleleft}$ *is a function on X onto Y;*
(2) $U^{\triangleleft} \circ U^{\triangleleft-1} = \Delta_Y$; *(3)* $U^{\triangleleft-1\triangleright} \subseteq U^{\circ-1}$.

Theorem 12.5 *For a corelation U on X to Y, the following assertions are equivalent:*

(1) $U^{\triangleleft-1}$ *is a function on Y onto X;*
(2) $U^{\triangleleft-1} \circ U^{\triangleleft} = \Delta_X$; *(3)* $U^{\circ-1} \subseteq U^{\triangleleft-1\triangleright}$.

Corollary 12.6 *For a corelation U on X to Y, the following assertions are equivalent:*

(1) $U^{\triangleleft}$ *is an injective function of X onto Y;*
(2) $U^{\circ-1} = U^{\triangleleft-1\triangleright}$; *(3)* $U^{\triangleleft-1} \circ U^{\triangleleft} = \Delta_X$ *and* $U^{\triangleleft} \circ U^{\triangleleft-1} = \Delta_Y$.

Remark 12.7 Note that if in particular U is union-preserving, then by Theorem 9.6 we have $U^{\circ} = U$. Therefore, assertion (2) can be written in the more instructive form that $U^{-1} = U^{\triangleleft-1\triangleright}$.

13 The Relationally Generated Inverse of a Corelation

Since the ordinary inverse U^{-1} of a corelation U on X to Y is a corelation on Y to X if and only if U is injective, it seems necessary to introduce the following

Definition 13.1 For any corelation U on X to Y, the corelation

$$U^{\mathbf{-1}} = U^{\triangleleft -1 \triangleright}$$

will be called the *relationally generated inverse* of U.

Thus, by writing $R^{\triangleright}$ in place of U, we can immediate derive the following

Theorem 13.2 *For any relation R on X to Y, we have*

$$R^{\triangleright \mathbf{-1}} = R^{-1 \triangleright}.$$

Proof Namely, by Definition 13.1 and Theorem 8.8, we have

$$R^{\triangleright \mathbf{-1}} = R^{\triangleright \triangleleft -1 \triangleright} = R^{-1 \triangleright}.$$

Remark 13.3 From the above theorem, by Theorem 8.8, we can infer that

$$R^{\triangleright \mathbf{-1} \triangleleft} = R^{-1 \triangleright \triangleleft} = R^{-1}.$$

Therefore, a corelationally generated inverse of a relation need not be defined.

Moreover, by using Theorems 8.11 and 8.8, we can easily prove the following

Theorem 13.4 *If U is a corelation on X to Y, then $U^{\mathbf{-1}}$ is a union-preserving corelation on Y to X such that*

(1) $U^{\triangleleft -1} = U^{\mathbf{-1} \triangleleft}$; *(2)* $U^{\circ \mathbf{-1}} = U^{\mathbf{-1}}$.

Proof To prove (2), note that by Definitions 9.1 and 13.1 and Theorem 8.8 we have

$$U^{\circ \mathbf{-1}} = U^{\triangleleft \triangleright \triangleleft -1 \triangleright} = U^{\triangleleft -1 \triangleright} = U^{\mathbf{-1}}.$$

Remark 13.5 If U is a corelation on X such that the relation $U^{\triangleleft}$ is symmetric, then by Definitions 13.1 and 9.1 we can also see that

$$U^{\mathbf{-1}} = U^{\triangleleft -1 \triangleright} = U^{\triangleleft \triangleright} = U^{\circ}.$$

Thus, if U is in addition union-preserving, then by Theorem 9.6 we can also state that $U^{\mathbf{-1}} = U$.

By using Remark 13.5, we can also easily establish the following

Example 13.6 If U is the complementation operation on X, then for any $A \subseteq X$ we have

$$U^{\mathbf{-1}}(A) = \begin{cases} \emptyset & \text{if} \quad \operatorname{card}(A) = 0, \\ A^c & \text{if} \quad \operatorname{card}(A) = 1, \\ X & \text{if} \quad \operatorname{card}(A) > 1. \end{cases}$$

Namely, now for any $x,\ y \in X$ we have

$$y \in U^{\triangleleft}(x) \iff y \in U\big(\{x\}\big) \iff y \in \{x\}^c \iff x \neq y$$
$$\iff x \in \{y\}^c \iff x \in U\big(\{y\}\big) \iff x \in U^{\triangleleft}(y) \iff y \in U^{\triangleleft-1}(x).$$

Therefore, $U^{\triangleleft} = U^{\triangleleft-1}$, and thus by Remark 13.5 we have $U^{\mathbf{-1}} = U^{\circ}$. Hence, by Example 9.5, we can see that the required assertion is also true.

Remark 13.7 In addition to Remark 13.5, it is also worth noticing that if U is and union-preserving corelation on X to Y such $U^{\triangleleft}$ is an injective function of X to Y, then by Remark 12.7 we have $U^{\mathbf{-1}} = U^{-1}$.

However, it is now more important to note that, by using Theorems 11.6 and 11.1, we can also prove the following

Theorem 13.8 *For an arbitrary corelation U on X to Y and a union-preserving corelation V on Y to Z, we have*

$$(\, V \circ U\,)^{\mathbf{-1}} = U^{\mathbf{-1}} \circ V^{\mathbf{-1}}.$$

Proof By Definition 13.1 and Theorems 11.6, 3.10 and 11.1, we have

$$(\, V \circ U\,)^{\mathbf{-1}} = (\, V \circ U\,)^{\triangleleft-1\triangleright} = \big(V^{\triangleleft} \circ U^{\triangleleft} \big)^{-1\triangleright}$$
$$= \Big(U^{\triangleleft-1} \circ V^{\triangleleft-1} \Big)^{\triangleright} = U^{\triangleleft-1\triangleright} \circ V^{\triangleleft-1\triangleright} = U^{\mathbf{-1}} \circ V^{\mathbf{-1}}.$$

From this theorem, by using Theorems 8.11 and 13.2, we can immediately derive

Corollary 13.9 *For a corelation U on X to Y and a relation S on Y to Z, we have*

$$(\, S^{\triangleright} \circ U\,)^{\mathbf{-1}} = U^{\mathbf{-1}} \circ S^{-1\triangleright}.$$

Moreover, by using the corresponding definitions, we can easily prove

Theorem 13.10 *If U is a corelation on X to Y, then for any $B \subseteq Y$ we have*

$$U^{\mathbf{-1}}(B) = \big\{ x \in X : \quad U\big(\{x\}\big) \cap B \neq \emptyset \,\big\}.$$

Proof By the corresponding definition, for any $x \in X$, we have

$$x \in U^{\mathbf{-1}}(B) \iff x \in U^{\lhd -1 \rhd}(B) \iff x \in U^{\lhd -1}[\,B\,]$$
$$\iff U^{\lhd}(x) \cap B \neq \emptyset \iff U\big(\{x\}\big) \cap B \neq \emptyset.$$

By the above proof and Theorem 13.4, it is clear that we also have

Theorem 13.11 *If U is a corelation on X to Y, then for any $B \subseteq Y$ we have*

$$U^{\mathbf{-1}}(B) = U^{\lhd -1}[\,B\,] = U^{\mathbf{-1}\lhd}[\,B\,].$$

Remark 13.12 Hence, analogously to Remark 9.12, we can note that

$$U^{\mathbf{-1}}(B) = \mathrm{cl}_{U^{\lhd}}(B)$$

for all $B \subseteq Y$, and thus $U^{\mathbf{-1}} = \mathrm{cl}_{U^{\lhd}}$.

By using Theorem 13.10, we can easily establish the following two examples which also reveal some serious disadvantages of the relationally generated inversion.

Example 13.13 If U is a corelation on X such that $U\big(\{x\}\big) = \{x\}$ for all $x \in X$, then $U^{\mathbf{-1}}$ is the identity corelation on X.

Remark 13.14 Thus, if in particular U is as in Example 5.13, then $U^{\mathbf{-1}}$ is the identity corelation on X.

Example 13.15 If U is a corelation on X to Y and $Z \subseteq Y$ such that $U\big(\{x\}\big) = Z$ for all $x \in X$, then for any $B \subseteq Y$ we have

$$U^{\mathbf{-1}}(B) = \emptyset \quad \text{if} \quad Z \cap B = \emptyset \qquad \text{and} \qquad U^{\mathbf{-1}}(B) = X \quad \text{if} \quad Z \cap B \neq \emptyset.$$

Namely, by Theorem 13.10, for any $x \in X$ we have

$$x \in U^{\mathbf{-1}}(B) \iff U\big(\{x\}\big) \cap B \neq \emptyset \iff Z \cap B \neq \emptyset.$$

Remark 13.16 Thus, if in particular $Z = Y$, then for any $B \subseteq Y$ we have

$$U^{\mathbf{-1}}(B) = \emptyset \quad \text{if} \quad B = \emptyset \qquad \text{and} \qquad U^{\mathbf{-1}}(B) = X \quad \text{if} \quad B \neq \emptyset.$$

14 Some Further Results on the Relationally Generated Inverse

By using Theorem 13.10, we can also easily prove the following three theorems.

Theorem 14.1 *If U is a corelation on X to Y, then*

$$\left(U^{-1}\right)^{-1} = U^{\circ}.$$

Proof By Theorems 13.10 and 9.4, for any $A \subseteq X$ and $y \in Y$, we have

$$\begin{aligned} y \in \left(U^{-1}\right)^{-1}(A) \iff U^{-1}\big(\{y\}\big) \cap A \neq \emptyset \iff \exists\, x \in A: \ x \in U^{-1}\big(\{y\}\big) \\ \iff \exists\, x \in A: \ U\big(\{x\}\big) \cap \{y\} \neq \emptyset \iff \exists\, x \in A: \ y \in U\big(\{x\}\big) \\ \iff y \in \bigcup_{x \in A} U\big(\{x\}\big) \iff y \in U^{\circ}(A). \end{aligned}$$

Therefore, the required equality is also true.

From this theorem, by using Theorem 9.6, we can immediately derive

Corollary 14.2 *For a corelation U on X to Y, the following assertions are equivalent:*

(1) $U = \left(U^{-1}\right)^{-1}$; *(2)* U *is union-preserving.*

Theorem 14.3 *For any two corelations U and V of X to Y, we have*

(1) $U^{-1} \setminus V^{-1} \leq \left(U \setminus V\right)^{-1}$;
(2) $\left(V^{-1}\right)^{c}(B) \subseteq (V^{c})^{-1}(B)$ *if* $\emptyset \neq B \subseteq Y$.

Proof By Definition 6.1 and Theorem 13.10, for any $x \in X$ and $B \subseteq Y$, we have

$$\begin{aligned} x \in \left(U^{-1} \setminus V^{-1}\right)(B) \iff x \in U^{-1}(B) \setminus V^{-1}(B) \iff \\ x \in U^{-1}(B), \quad x \notin V^{-1}(B) \iff U\big(\{x\}\big) \cap B \neq \emptyset, \quad V\big(\{x\}\big) \cap B = \emptyset \\ \implies \big(U\big(\{x\}\big) \setminus V\big(\{x\}\big)\big) \cap B \neq \emptyset \iff \big(U \setminus V\big)\big(\{x\}\big) \cap B \neq \emptyset \\ \iff x \in \big(U \setminus V\big)^{-1}(B). \end{aligned}$$

Therefore, $\left(U^{-1} \setminus V^{-1}\right)(B) \subseteq \left(U \setminus V\right)^{-1}(B)$ for all $B \subseteq Y$, and thus (1) is true.

Moreover, also by Definition 6.1 and Theorem 13.10, for any $x \in X$ and $\emptyset \neq B \subseteq Y$, we have

$$\begin{aligned} x \in \left(V^{-1}\right)^{c}(B) \iff x \in V^{-1}(B)^{c} \iff x \notin V^{-1}(B) \\ \iff V\big(\{x\}\big) \cap B = \emptyset \iff B \subseteq V\big(\{x\}\big)^{c} \implies V\big(\{x\}\big)^{c} \cap B \neq \emptyset \\ \iff V^{c}\big(\{x\}\big) \cap B \neq \emptyset \iff x \in \left(V^{c}\right)^{-1}(B). \end{aligned}$$

Therefore, $\left(V^{-1}\right)^{c}(B) \subseteq (V^{c})^{-1}(B)$, and thus (2) is also true.

Remark 14.4 In addition to the letter direct proof, it is also worth noticing that assertion (2) can actually be derived from (1), by defining $U\,(A)=Y$ for all $A\subseteq X$.

Namely, in this case, by Remark 6.2 we have $U\setminus V=V^c$, and thus

$$\big(U\setminus V\big)^{-\mathbf{1}}(B)=\big(V^c\big)^{-\mathbf{1}}\,(B)$$

for all $B\subseteq Y$.

Moreover, by the corresponding definitions and Remark 13.16, we have

$$\Big(U^{-\mathbf{1}}\setminus V^{-\mathbf{1}}\Big)\,(B)=U^{-\mathbf{1}}(B)\setminus V^{-\mathbf{1}}(B)=X\setminus V^{-\mathbf{1}}(B)=V^{-\mathbf{1}}(B)^c=\Big(V^{-\mathbf{1}}\Big)^c\,(B)$$

for all $\emptyset\neq B\subseteq Y$.

Theorem 14.5 *For any family $\mathscr{U}$ of corelations on X to Y, we have*

(1) $\Big(\bigwedge\limits_{U\in\mathscr{U}}U\Big)^{-\mathbf{1}}\leq\bigwedge\limits_{U\in\mathscr{U}}U^{-\mathbf{1}}$; *(2)* $\Big(\bigvee\limits_{U\in\mathscr{U}}U\Big)^{-\mathbf{1}}=\bigvee\limits_{U\in\mathscr{U}}U^{-\mathbf{1}}$.

Proof To prove (1), note that by Theorem 13.10 and Definition 6.3, for any $x\in X$ and $B\subseteq Y$, we have

$$x\in\Big(\bigwedge_{U\in\mathscr{U}}U\Big)^{-\mathbf{1}}(B)\iff\Big(\bigwedge_{U\in\mathscr{U}}U\Big)\big(\{x\}\big)\cap B\neq\emptyset$$
$$\iff\Big(\bigcap_{U\in\mathscr{U}}U\big(\{x\}\big)\Big)\cap B\neq\emptyset\implies\forall\;U\in\mathscr{U}:\;U\big(\{x\}\big)\cap B\neq\emptyset\iff$$
$$\forall\;U\in\mathscr{U}:\;x\in U^{-\mathbf{1}}(B)\iff x\in\bigcap_{U\in\mathscr{U}}U^{-\mathbf{1}}(B)\iff x\in\Big(\bigwedge_{U\in\mathscr{U}}U^{-\mathbf{1}}\Big)(B).$$

Therefore,

$$\Big(\bigwedge_{U\in\mathscr{U}}U\Big)^{-\mathbf{1}}(B)\subseteq\Big(\bigwedge_{U\in\mathscr{U}}U^{-\mathbf{1}}\Big)(B)$$

for all $B\subseteq Y$, and thus (1) is also true.

Now, analogously to Theorem 11.10, we can also prove the following

Theorem 14.6 *For any union-preserving corelation U on X, we have*

$$U^{\infty-\mathbf{1}}=U^{-\mathbf{1}\infty}.$$

Proof From Theorem 13.8, by induction, we can see that $U^{n-\mathbf{1}}=U^{-\mathbf{1}n}$ for all $n\in\mathbb{N}$.

Hence, by Definition 7.4 and Theorem 13.4, it is clear that

$$U^{\infty-\mathbf{1}}=\Big(\bigvee_{n=0}^{\infty}U^n\Big)^{-\mathbf{1}}=\bigvee_{n=0}^{\infty}U^{n\,-\mathbf{1}}=\bigvee_{n=0}^{\infty}U^{-\mathbf{1}n}=U^{-1\infty}.$$

15 The Relationally Generated Composition of Corelations

Analogously to Definition 13.1, we may also naturally introduce the following

Definition 15.1 For any two corelations U on X to Y and V on Y to Z, the corelation

$$V \bullet U = \left(V^{\lhd} \circ U^{\lhd} \right)^{\rhd}$$

will be called the *relationally generated composition* of V and U.

Thus, by using Theorem 8.8, we can easily establish the following

Theorem 15.2 *For any two relations R on X to Y and S on Y to Z, we have*

$$S^{\rhd} \bullet R^{\rhd} = \left(S \circ R \right)^{\rhd}.$$

Proof Namely, by Definition 15.1 and Theorem 8.8, we have

$$S^{\rhd} \bullet R^{\rhd} = \left(S^{\rhd \lhd} \circ R^{\rhd \lhd} \right)^{\rhd} = \left(S \circ R \right)^{\rhd}.$$

Remark 15.3 From this theorem, by using Theorem 8.8, we can see that

$$\left(S^{\rhd} \bullet R^{\rhd} \right)^{\lhd} = \left(S \circ R \right)^{\rhd \lhd} = S \circ R.$$

Therefore, the corelationally generated composition of relations need not be defined.

Moreover, by using Theorems 8.11 and 11.1, we can easily prove the following

Theorem 15.4 *If U is a corelation on X to Y and V is a corelation on Y to Z, then $V \bullet U$ is a union-preserving corelation on X to Z such that*

$$V \bullet U = V^{\circ} \circ U^{\circ}.$$

Proof To check this equality, note that, by the corresponding definitions and Theorem 11.1, we have

$$V \bullet U = \left(V^{\lhd} \circ U^{\lhd} \right)^{\rhd} = V^{\lhd \rhd} \circ U^{\lhd \rhd} = V^{\circ} \circ U^{\circ}.$$

Thus, in particular, by Theorem 9.6, we can also state

Corollary 15.5 *For any two union-preserving corelations U on X to Y and V on Y to Z, we have*

$$V \bullet U = V \circ U.$$

Moreover, by using Theorems 15.4, we can easily prove the following theorems.

Theorem 15.6 *For any two corelations U on X to Y and V on Y to Z, we have*

$$\big(V \bullet U\big)^{-1} = U^{-1} \circ V^{-1}.$$

Proof By Theorems 15.4, 9.2, 13.8 and 13.4, it is clear that

$$\big(V \bullet U\big)^{-1} = (\,V^{\circ} \circ U^{\circ})^{-1} = U^{\circ -1} \circ V^{\circ -1} = U^{-1} \circ V^{-1}.$$

Theorem 15.7 *For any three corelations U on X to Y, V on Y to Z and W on Z to Ω, we have*

$$W \bullet \big(V \bullet U\big) = \big(W \bullet V\,\big) \bullet U.$$

Proof By Theorem 15.4 and 3.11, it is clear that

$$W \bullet \big(V \bullet U\big) = W^{\circ} \circ (\,V^{\circ} \circ U^{\circ}) = (\,W^{\circ} \circ V^{\circ}) \circ U^{\circ} = \big(W \bullet V\,\big) \bullet U.$$

Theorem 15.8 *For any two corelations U on X to Y and V on Y to Z and $A \subseteq X$, we have*

$$\big(V \bullet U\big)(A) = \bigcup_{x \in A} \; \bigcup_{y \in U(\{x\})} V\big(\{y\}\big).$$

Proof By Theorems 15.4, 3.11, 9.4 and 9.2, it is clear that

$$\begin{aligned}\big(V \bullet U\big)(A) &= (\,V^{\circ} \circ U^{\circ})\,(A) = V^{\circ}\,\big(U^{\circ}\big(A\big)\big)\\ &= V^{\circ}\Big(\bigcup_{x \in A} U\big(\{x\}\big)\Big) = \bigcup_{x \in A} V^{\circ}\,\big(U\big(\{x\}\big)\big) = \bigcup_{x \in A} \; \bigcup_{y \in U(\{x\}\big)} V\big(\{y\}\big).\end{aligned}$$

16 Proximal Interiors and Closures Derived from Corelations

For the origins of the following definition, see Efremovič [16], Smirnov [48] and Száz [50].

Definition 16.1 If U is a corelation on X to Y, then for any $A \subseteq X$ and $B \subseteq Y$ we write

(1) $A \in \mathrm{Int}_U(\,B)$ if $U(A) \subseteq B$;
(2) $A \in \mathrm{Cl}_U(B)$ if $U(A) \cap B \neq \emptyset$.

The relations Cl_U and Int_U will be called the *proximal closure* and *proximal interior* generated by the corelation U, respectively.

Remark 16.2 Relations on $\mathscr{P}(Y)$ to $\mathscr{P}(X)$ can be identified with functions on $\mathscr{P}(Y)$ to $\mathscr{P}\big(\mathscr{P}(X)\big)$.

Therefore, the above relations may also be naturally considered as corelations on Y to $\mathscr{P}(X)$.

The following theorem will already indicate a main advantage of our notations Cl_U and Int_U over the standardized ones δ_U and $\Subset_U$ of Efremovič and Smirnov. (For the historical developments of the subject, see Thron [84] and Naimpally and Warrack [38].)

Theorem 16.3 *If U is a corelation on X to Y, then for any $B \subseteq Y$ we have*

(1) $\mathrm{Cl}_U\,(B) = \mathscr{P}(X) \setminus \mathrm{Int}_U\,(\,Y \setminus\, B)$*;*
(2) $\mathrm{Int}_U\,(B) = \mathscr{P}(X) \setminus \mathrm{Cl}_U\,(\,Y \setminus B\,)$.

Proof If $A \in \mathrm{Cl}_U(B)$, then $U(A) \cap B \neq \emptyset$, and thus $U(A) \not\subseteq Y \setminus B$. Therefore, $A \notin \mathrm{Int}_{\mathscr{U}}(\,Y \setminus B\,)$, and thus $A \in \mathscr{P}(X) \setminus \mathrm{Int}_{\mathscr{U}}(\,Y \setminus\, B)$. Hence, we can already see that $\mathrm{Cl}_U(B) \subseteq \mathscr{P}(X) \setminus \mathrm{Int}_U(\,Y \setminus\, B)$. The converse inclusion can be proved quite similarly by reversing the above argument. Thus, (1) is true.

Now, (2) can be derived from (1) by noticing that (1) implies

$$\mathrm{Int}_U(\,Y \setminus B\,) = \mathscr{P}(X) \setminus \mathrm{Cl}_U(\,B\,)$$

for all $B \subseteq Y$. Hence, by writing $Y \setminus B$ in place B, we can see that (2) also holds.

By using appropriate complementations, the above theorem can be reformulated in a more concise form.

Corollary 16.4 *For any corelation U on X to Y, we have*

(1) $\mathrm{Int}_U = \big(\,\mathrm{Cl}_U \circ \mathscr{C}_Y\big)^c = \big(\,\mathrm{Cl}_U\,\big)^c \circ \mathscr{C}_Y$*;*
(2) $\mathrm{Cl}_U = \big(\,\mathrm{Int}_U \circ \mathscr{C}_Y\big)^c = \big(\,\mathrm{Int}_U\,\big)^c \circ \mathscr{C}_Y$.

Proof To check (1), note that for any $B \subseteq Y$ we have

$$\begin{aligned}\mathrm{Int}_U\,(B) &= \mathscr{P}(X) \setminus \mathrm{Cl}_U\,(\,Y \setminus B\,) = \mathrm{Cl}_U\,\big(B^c\big)^c\\ &= \mathrm{Cl}_U\,\big(\mathscr{C}_Y(B)\big)^c = \big(\,\mathrm{Cl}_U \circ \mathscr{C}_Y\big)(B)^c = \big(\,\mathrm{Cl}_U \circ \mathscr{C}_Y\big)^c(B).\end{aligned}$$

Therefore, $\mathrm{Int}_U = \big(\,\mathrm{Cl}_U \circ \mathscr{C}_Y\big)^c$. Moreover, by Corollary 4.11, we also have $\big(\,\mathrm{Cl}_U \circ \mathscr{C}_Y\big)^c = \big(\,\mathrm{Cl}_U\,\big)^c \circ \mathscr{C}_Y$.

Remark 16.5 The above results show that the relations Cl_U and Int_U are equivalent tools in the corelational space $(\,X,\ Y)(\,U)$.

However, in the sequel, we shall see that there are cases when one of the above relations is a more convenient tool than the other one.

By using Definition 16.1, we can also easily prove the following

Theorem 16.6 *If U is a corelation on X to Y, then*

(1) Cl_U *is union-preserving;* *(2)* Int_U *is intersection-preserving.*

Proof To prove (2), note that, by Definition 16.1, the corelation Int_U is increasing. Therefore, if $\mathscr{B} \subseteq \mathscr{P}(Y)$, then by Theorem 5.6 we can at once state that

$$\mathrm{Int}_U\left(\bigcap \mathscr{B}\right) \subseteq \bigcap_{B \in \mathscr{B}} \mathrm{Int}_U(B).$$

Moreover, if $A \in \bigcap_{B \in \mathscr{B}} \mathrm{Int}_U(B)$, then $A \in \mathrm{Int}_U(B)$, and thus $U(A) \subseteq B$ for all $B \in \mathscr{B}$. Therefore, $U(A) \subseteq \bigcap \mathscr{B}$, and thus $A \in \mathrm{Int}_U\left(\bigcap \mathscr{B}\right)$. Therefore,

$$\bigcap_{B \in \mathscr{B}} \mathrm{Int}_U(B) \subseteq \mathrm{Int}_U\left(\bigcap \mathscr{B}\right),$$

and thus the corresponding equality is also true.

Remark 16.7 Note that, because of Theorem 16.3, assertions (1) and (2) are actually equivalent.

However, by Theorem 5.9, assertion (1) can also be proved by showing only that $\mathrm{Cl}_U(B) = \bigcup_{y \in B} \mathrm{Cl}_U\left(\{y\}\right)$ for all $B \subseteq Y$.

From Theorem 16.6, by taking empty union and intersection, we can immediately derive

Corollary 16.8 *For any corelation U on X to Y, we have*

(1) $\mathrm{Cl}_U(\emptyset) = \emptyset$; *(2)* $\mathrm{Int}_U(Y) = \mathscr{P}(X)$.

Moreover, from Theorem 16.6, by Theorems 5.6 and 5.7, it is also clear that in particular we also have

Corollary 16.9 *If U is a corelation on X to Y, then the corelations* Cl_U *and* Int_U *are increasing.*

Hence, by using Theorems 5.6 and 5.7, we can immediately derive

Theorem 16.10 *If U is a corelation on X to Y, then for any family $\mathscr{B}$ of subsets of Y, we have*

(1) $\mathrm{Cl}_U\left(\bigcap \mathscr{B}\right) \subseteq \bigcap_{B \in \mathscr{B}} \mathrm{Cl}_U(B)$; *(2)* $\bigcup_{B \in \mathscr{B}} \mathrm{Int}_U(B) \subseteq \mathrm{Int}_U\left(\bigcup \mathscr{B}\right)$.

The following example, together with Theorem 16.3, shows that the corresponding equalities need not be true even in a very simple case.

Example 16.11 If in particular $X = B_1 \cup B_2$, with $B_i = \{i\}$, and U is the identity corelation on X, then by Definition 16.1 it is clear that

$$\mathrm{Int}_U(B_1 \cup B_2) = \mathscr{P}(X), \qquad \text{but} \qquad \mathrm{Int}_U(B_1) \cup \mathrm{Int}_U(B_2) = \mathscr{P}(X) \setminus \{\{X\}\}.$$

Moreover, it is also worth noticing that now we also have

$$\mathrm{Int}_{U^c}(B_1 \cup B_2) = \mathscr{P}(X), \quad \text{but} \quad \mathrm{Int}_{U^c}(B_1) \cup \mathrm{Int}_{U^c}(B_2) = \mathscr{P}(X) \setminus \{\{\emptyset\}\}.$$

By using Corollary 16.9, we can also easily prove the following

Theorem 16.12 *If U is a corelation on X to Y, then the corelations* Cl_U^{-1} *and* Int_U^{-1} *are ascending-valued.*

Proof To prove the first statement, suppose that $A \subseteq X$, $B_1 \in \mathrm{Cl}_U^{-1}(A)$ and $B_1 \subseteq B_2 \subseteq Y$. Then, by the definition of the inverse relation, we have $A \in \mathrm{Cl}_U(B_1)$. Moreover, by Corollary 16.9, we have $\mathrm{Cl}_U(B_1) \subseteq \mathrm{Cl}_U(B_1)$. Therefore, we also have $A \in \mathrm{Cl}_U(B_2)$, and thus $B_1 \in \mathrm{Cl}_U^{-1}(A)$. This shows that $\mathrm{Cl}_U^{-1}(A)$ is an ascending family of subsets of Y.

Remark 16.13 Later, we shall see that if in particular U is union-preserving, then Cl_U^{-1} is increasing, but Int_U^{-1} is decreasing.

By using the corresponding definitions, we can also easily prove the following

Theorem 16.14 *For any two corelations U and V on X to Y, the following assertions are equivalent:*

(1) $U \leq V$; *(2)* $\mathrm{Cl}_U \subseteq \mathrm{Cl}_V$ *(3)* $\mathrm{Int}_V \subseteq \mathrm{Int}_U$.

Proof To check the implication (3) $\Longrightarrow$ (1), note that by Definition 16.1 for any $A \subseteq X$ we have $A \in \mathrm{Int}_V\big(V(A)\big)$, and thus $\big(V(A),\ A\big) \in \mathrm{Int}_V$. Therefore, if (3) holds, then we also have $\big(V(A),\ A\big) \in \mathrm{Int}_U$, and thus $A \in \mathrm{Int}_U\big(V(A)\big)$. Hence, by Definition 16.1, we can see that $U(A) \subseteq V(A)$, and thus (1) also holds.

Now, as an immediate consequence of this theorem, we can also state

Corollary 16.15 *For any two corelations U and V on X to Y, each of the equalities* $\mathrm{Cl}_U = \mathrm{Cl}_V$ *and* $\mathrm{Int}_U = \mathrm{Int}_V$ *implies $U = V$.*

Remark 16.16 By Theorem 16.14, for any two corelations U and V on X to Y, we have

$$\mathrm{Cl}_U \subseteq \mathrm{Cl}_V \iff U \leq V.$$

Therefore, the mappings

$$U \longmapsto \mathrm{Cl}_U \qquad \text{and} \qquad U \longmapsto U,$$

where U is a corelation on X to Y, establish a Pataki connection.

Moreover, this Pataki connection is very particular since the first mapping is injective and the second one is injective and onto.

17 Proximal Interiors and Closures Derived from Relations

By using Definitions 8.1 and 16.1, we can naturally introduce the following

Definition 17.1 For any relation R on X to Y, the relations

$$\operatorname{Cl}_R = \operatorname{Cl}_{R^{\vartriangleright}} \qquad \text{and} \qquad \operatorname{Int}_R = \operatorname{Int}_{R^{\vartriangleright}}$$

will be called the *the proximal closure* and *proximal interior* generated by the relation R, respectively.

Thus, by the results of Sect. 16, we evidently have the following assertions.

Theorem 17.2 *If R is a relation on X to Y, then for any $B \subseteq Y$ we have*

(1) $\operatorname{Cl}_R\,(B) = \mathcal{P}(X) \setminus \operatorname{Int}_R\,(\,Y \setminus\; B)$;
(2) $\operatorname{Int}_R\,(B) = \mathcal{P}(X) \setminus \operatorname{Cl}_R\,(\,Y \setminus B\,)$.

Corollary 17.3 *For any relation R on X to Y, we have*

(1) $\operatorname{Int}_R = \big(\operatorname{Cl}_R \circ \mathcal{C}_Y\big)^c = \big(\operatorname{Cl}_R\big)^c \circ \mathcal{C}_Y$;
(2) $\operatorname{Cl}_R = \big(\operatorname{Int}_R \circ \mathcal{C}_Y\big)^c = \big(\operatorname{Int}_R\big)^c \circ \mathcal{C}_Y$.

Remark 17.4 The above results show that the relations Cl_R and Int_R are equivalent tools in the relational space $(\,X,\; Y)(\,R)$.

Theorem 17.5 *If R is a relation on X to Y, then*

(1) Cl_R *is union-preserving;* *(2)* Int_R *is intersection-preserving.*

Corollary 17.6 *For any relation R on X to Y, we have*

(1) $\operatorname{Cl}_R(\emptyset) = \emptyset$; *(2)* $\operatorname{Int}_R(\,Y\,) = \mathcal{P}(X)$.

Corollary 17.7 *If R is a relation on X to Y, then the corelations Cl_R and Int_R are increasing.*

Theorem 17.8 *If R is a relation on X to Y, then for any family $\mathcal{B}$ of subsets of Y, we have*

(1) $\operatorname{Cl}_R\big(\bigcap \mathcal{B}\big) \subseteq \bigcap\limits_{B \in \mathcal{B}} \operatorname{Cl}_R(B)$; *(2)* $\bigcup\limits_{B \in \mathcal{B}} \operatorname{Int}_R(B) \subseteq \operatorname{Int}_R\big(\bigcup \mathcal{B}\big)$.

The fact that the corresponding equalities need not be true even in the most simple cases is apparent from the following modification of Example 16.11.

Example 17.9 If in particular $X = B_1 \cup B_2$, with $B_i = \{i\}$, and R is the identity relation on X, then

$$\operatorname{Int}_R(\,B_1 \cup B_2) = \mathcal{P}(X), \qquad \text{but} \qquad \operatorname{Int}_R(B_1) \cup \operatorname{Int}_R(\,B_2) = \mathcal{P}(X) \setminus \{\{X\}\}.$$

Moreover, it is also worth noticing that now the above equalities also hold with R^c in place of R.

Theorem 17.10 *If R is a relation on X to Y, then the corelations Cl_R^{-1} and Int_R^{-1} are ascending-valued.*

By the corresponding definitions, it is clear that we also have the following

Theorem 17.11 *If R is a relation on X to Y, then for any $A \subseteq X$ and $B \subseteq Y$ we have*

(1) $A \in \mathrm{Int}_R(B) \iff R[A] \subseteq B$;
(2) $A \in \mathrm{Cl}_R(B) \iff R[A] \cap B \neq \emptyset$.

Moreover, since $U^{\circ} = U^{\lhd \rhd}$, we can also at once state the following

Theorem 17.12 *For any corelation U on X to Y, we have*

(1) $\mathrm{Cl}_{U^{\circ}} = \mathrm{Cl}_{U^{\lhd}}$; *(2)* $\mathrm{Int}_{U^{\circ}} = \mathrm{Int}_{U^{\lhd}}$.

The following theorem, together with Definition 17.1, shows that corelations can generate more general proximal closures and interiors than relations.

Theorem 17.13 *For any corelation U on X to Y, the following assertions are equivalent:*

(1) *U is union-preserving;*
(2) $\mathrm{Cl}_U = \mathrm{Cl}_{U^{\lhd}}$; *(3)* $\mathrm{Int}_U = \mathrm{Int}_{U^{\lhd}}$;
(4) $\mathrm{Cl}_U = \mathrm{Cl}_R$ *for some relation R on X to Y;*
(5) $\mathrm{Int}_U = \mathrm{Int}_R$ *for some relation R on X to Y.*

Proof If (1) holds, then by Theorem 9.6 we have $U = U^{\circ}$. Hence, by Theorem 17.12, we can see that $\mathrm{Int}_U = \mathrm{Int}_{U^{\circ}} = \mathrm{Int}_{U^{\lhd}}$. Therefore, (3) and thus (5) also holds.

While, if (5) holds, then by Definition 17.1 we also have $\mathrm{Int}_U = \mathrm{Int}_{R^{\rhd}}$. Hence, by Corollary 16.15, it follows that $U = R^{\rhd}$. Therefore, by Theorem 8.11, (1) also holds.

By using this theorem, we can easily establish the following

Example 17.14 Let $X = \{1, 2\}$, and for any $A \subseteq X$ define

$$U(A) = A \quad \text{if} \quad A \neq X \qquad \text{and} \qquad U(A) = \{1\} \quad \text{if} \quad A = X.$$

Then, U is a corelation on X such that, for any relation R on X, we have
(1) $\mathrm{Cl}_U \neq \mathrm{Cl}_R$; (2) $\mathrm{Int}_U \neq \mathrm{Int}_R$.

Note that now U is not even quasi-increasing, therefore by Theorem 17.13 the required assertions are true.

However, $U^{\lhd}$ is just the identity relation on X. Therefore, besides the automatic equalities $\mathrm{Int}_U(X) = \mathscr{P}(X) = \mathrm{Int}_{U^{\lhd}}(X)$, we still have

$$\mathrm{Int}_U\big(\{2\}\big) = \big\{\emptyset,\ \{2\}\big\} = \mathrm{Int}_{U^{\lhd}}\big(\{2\}\big).$$

Moreover, it is also worth noticing that now $U^{\circ} = U^{\lhd \rhd}$ is just the identity corelation on X. Therefore, $U \leq U^{\circ}$, but in accordance with Theorem 9.6 the converse inequality fails to hold.

The fact that corelations can generate more general proximal closures and interiors is also apparent from the following

Theorem 17.15 *For any relation R on X to Y, we have*

(1) $\mathrm{Cl}_R^{-1} = \mathrm{Cl}_{R^{-1}}$*;* *(2)* $\mathrm{Int}_R^{-1} = \mathscr{C}_Y \circ \mathrm{Int}_{R^{-1}} \circ \mathscr{C}_X$.

Proof By Theorems 17.11 and 3.3, for any $A \subseteq X$ and $B \subseteq Y$, we have

$$\begin{aligned} B \in \mathrm{Cl}_R^{-1}(A) \iff A \in \mathrm{Cl}_R(B) \iff R\,[\,A\,] \cap B \neq \emptyset \\ \iff R^{-1}(B) \cap A \neq \emptyset \iff B \in \mathrm{Cl}_{R^{-1}}(A). \end{aligned}$$

Therefore, (1) is true.

By using Theorem 17.11, assertion (2) can be proved somewhat more tediously. Therefore, it is better to derive it from assertion (1) by using Corollary 17.3. For this, we have to note only that

$$\begin{aligned} \mathrm{Int}_R^{-1} = \big(\,\mathrm{Cl}_R^{c} \circ \mathscr{C}_Y\big)^{-1} = \mathscr{C}_Y^{-1} \circ \mathrm{Cl}_R^{c\,-1} \\ = \mathscr{C}_Y \circ \mathrm{Cl}_R^{-1\,c} = \mathscr{C}_Y \circ \mathrm{Cl}_{R^{-1}}^{c} = \mathscr{C}_Y \circ \mathrm{Int}_{R^{-1}} \circ \mathscr{C}_X . \end{aligned}$$

Remark 17.16 By using the above theorem, the properties of the relations Cl_R^{-1} and Int_R^{-1} can be easily derived from those of Cl_R and Int_R.

For instance, from Theorem 17.5, by using Theorem 17.15, we can easily derive

Theorem 17.17 *If R is a relation on X to Y, then*

(1) Cl_R^{-1} *is union-preserving;* *(2)* Int_R^{-1} *is union-reversing.*

Proof To prove (2), note that, by Theorem 17.15, we have

$$\mathrm{Int}_R^{-1}(A) = \big(\mathscr{C}_Y \circ \mathrm{Int}_{R^{-1}} \circ \mathscr{C}_X\big)(A) = \mathscr{C}_Y\left[\mathrm{Int}_{R^{-1}}\left(A^c\right)\right]$$

for all $A \subseteq X$.

Hence, by using Theorem 17.5 and DeMorgan's law, we can see that

$$\begin{aligned} \mathrm{Int}_R^{-1}\left(\bigcup \mathscr{A}\right) = \mathscr{C}_Y\Big[\,\mathrm{Int}_{R^{-1}}\left(\left(\bigcup \mathscr{A}\right)^c\right)\Big] = \mathscr{C}_Y\Big[\,\mathrm{Int}_{R^{-1}}\Big(\bigcap_{A \in \mathscr{A}} A^c\Big)\Big] \\ = \mathscr{C}_Y\Big(\bigcap_{A \in \mathscr{A}} \mathrm{Int}_{R^{-1}}\left(A^c\right)\Big) = \bigcap_{A \in \mathscr{A}} \mathscr{C}_Y\left[\mathrm{Int}_{R^{-1}}\left(A^c\right)\right] = \bigcap_{A \in \mathscr{A}} \mathrm{Int}_R^{-1}(A) \end{aligned}$$

for all $\mathscr{A} \subseteq \mathscr{P}(X)$.

Remark 17.18 The assertions of the above theorem can be reformulated in the more direct forms that:

(1) for any $\mathscr{A} \subseteq \mathscr{P}(X)$ and $B \subseteq Y$ we have

$$\bigcup \mathscr{A} \in \mathrm{Cl}_R(B) \iff \exists\ A \in \mathscr{A}: \quad A \in \mathrm{Cl}_R(B).$$

(2) for any $\mathscr{A} \subseteq \mathscr{P}(X)$ and $B \subseteq Y$ we have

$$\bigcup \mathscr{A} \in \mathrm{Int}_R(B) \iff \forall\ A \in \mathscr{A}: \quad A \in \mathrm{Int}_R(B).$$

To check (2), note that, by Theorem 17.17, we have

$$\bigcup \mathscr{A} \in \mathrm{Int}_R(B) \iff B \in \mathrm{Int}_R^{-1}\left(\bigcup \mathscr{A}\right) \iff B \in \bigcap_{A \in \mathscr{A}} \mathrm{Int}_R^{-1}(A)$$
$$\iff \forall\ A \in \mathscr{A}: \quad B \in \mathrm{Int}_R^{-1}(A) \iff \forall\ A \in \mathscr{A}: \quad A \in \mathrm{Int}_R(B).$$

From Theorem 17.17, by taking empty union, we can immediately derive

Corollary 17.19 *For any relation R on X to Y, we have*

(1) $\mathrm{Cl}_R^{-1}(\emptyset) = \emptyset$; *(2)* $\mathrm{Int}_R^{-1}(\emptyset) = \mathscr{P}(Y)$.

Remark 17.20 The assertions of this corollary can be reformulated in the more direct forms that:

(1) $\emptyset \notin \mathrm{Cl}_R(B)$ for all $B \subseteq Y$; (2) $\emptyset \in \mathrm{Int}_R(B)$ for all $B \subseteq Y$;

From Theorem 17.17, it is clear that in particular we can also state

Corollary 17.21 *If R is a relation on X, then the corelation Cl_R^{-1} is increasing, while the corelation Int_R^{-1} is decreasing.*

Thus, analogously to Theorem 17.8, we can also prove

Theorem 17.22 *If R is a relation on X to Y, then for any family $\mathscr{A}$ of subsets of X, we have*

(1) $\mathrm{Cl}_R^{-1}\left(\bigcap \mathscr{A}\right) \subseteq \bigcap_{A \in \mathscr{A}} \mathrm{Cl}_R^{-1}(A)$; *(2)* $\bigcup_{A \in \mathscr{A}} \mathrm{Int}_R^{-1}(A) \subseteq \mathrm{Int}_R^{-1}\left(\bigcap \mathscr{A}\right)$.

Finally, we note that, by using Theorem 16.14, we can also prove

Theorem 17.23 *For any two relations R and S on X to Y, the following assertions are equivalent:*

(1) $R \subseteq S$; *(2)* $\mathrm{Cl}_R \subseteq \mathrm{Cl}_S$ *(3)* $\mathrm{Int}_S \subseteq \mathrm{Int}_R$.

Proof To check the implication (3) $\Longrightarrow$ (1), note that if (3) holds, then by Definition 17.1 we have $\mathrm{Int}_{S^{\triangleright}} \subseteq \mathrm{Int}_{S^{\triangleright}}$. Hence, by Theorem 16.14, it follows that $R^{\triangleright} \subseteq S^{\triangleright}$. Therefore, by Theorem 8.8, inclusion (1) also holds.

Now, as an immediate consequence of this theorem, we can also state

Corollary 17.24 *For any two relations R and S on X to Y, each of the equalities* $\mathrm{Cl}_R = \mathrm{Cl}_S$ *and* $\mathrm{Int}_R = \mathrm{Int}_S$ *implies that* $R = S$.

Remark 17.25 By Theorem 17.22, for any two relations R and S on X to Y, we have

$$\mathrm{Cl}_R \subseteq \mathrm{Cl}_S \quad \Longleftrightarrow \quad R \subseteq V.$$

Therefore, the mappings

$$R \longmapsto \mathrm{Cl}_R \qquad \text{and} \qquad R \longmapsto R,$$

where R is a relation on X to Y, establish a Pataki connection.

Moreover, this Pataki connection is very particular since the first mapping is injective and the second one is injective and onto.

18 Topological Interiors and Closures Derived from Corelations

Definition 18.1 If U is a corelation on X to Y, then for any $x \in X$ and $B \subseteq Y$ we write

(1) $x \in \mathrm{cl}_U(B)$ if $\{x\} \in \mathrm{Cl}_U(B)$;
(2) $x \in \mathrm{int}_U(B)$ if $\{x\} \in \mathrm{Int}_U(B)$.

The relations cl_U and int_U will be called the *topological closure* and *topological interior* generated by the corelation U, respectively.

Remark 18.2 Relations on $\mathscr{P}(Y)$ to X can be identified with functions on $\mathscr{P}(Y)$ to $\mathscr{P}(X)$.

Therefore, the above relations may also be naturally considered as corelations on Y to X.

By using the above definition, from the results of Sect. 16, we can easily derive the following assertions.

Theorem 18.3 *If U is a corelation on X to Y, then for any $B \subseteq Y$ we have*

(1) $\mathrm{cl}_U(B) = X \setminus \mathrm{int}_U(Y \setminus B)$; *(2)* $\mathrm{int}_U(B) = X \setminus \mathrm{cl}_U(Y \setminus B)$.

Proof To check (1), note that by Definition 18.1 and Theorem 16.3, for any $x \in X$ we have

$$x \in \mathrm{cl}_U(B) \iff \{x\} \in \mathrm{Cl}_U(B) \iff \{x\} \notin \mathrm{Int}_U(B^c) \iff x \notin \mathrm{int}_U(B^c).$$

Corollary 18.4 *For any corelation U on X to Y, we have*

(1) $\operatorname{int}_U = \left(\operatorname{cl}_U \circ \mathscr{C}_Y \right)^c = \left(\operatorname{cl}_U \right)^c \circ \mathscr{C}_Y$*;*

(2) $\operatorname{cl}_U = \left(\operatorname{int}_U \circ \mathscr{C}_Y \right)^c = \left(\operatorname{int}_U \right)^c \circ \mathscr{C}_Y$.

Remark 18.5 The above results show that the relations cl_U and int_U are also equivalent tools in the corelational space $(X, Y)(U)$.

The fact that the relations Cl_U and Int_U are, in general, much better tools than cl_U and int_U is already apparent from the following

Example 18.6 If U is a corelation on X such that $U\big(\{x\}\big) = \{x\}$ for all $x \in X$, then cl_U and int_U are already the identity corelations on X.

To check this, note that now, for any $x \in X$ and $A \subseteq X$, we have

$$x \in \operatorname{int}_U(A) \iff \{x\} \in \operatorname{Int}_U(A) \iff U\big(\{x\}\big) \subseteq A \iff \{x\} \subseteq A \iff x \in A.$$

Therefore, $\operatorname{int}_U(A) = A$ for all $A \subseteq X$.

Theorem 18.7 *If U is a corelation on X to Y, then*

(1) cl_U *is union-preserving;* *(2)* int_U *is intersection-preserving.*

Corollary 18.8 *For any corelation U on X to Y, we have*

(1) $\operatorname{cl}_U(\emptyset) = \emptyset$*;* *(2)* $\operatorname{int}_U(Y) = X$.

Corollary 18.9 *If U is a corelation on X to Y, then the corelations* cl_U *and* int_U *are increasing.*

Theorem 18.10 *If U is a corelation on X to Y, then for any family $\mathscr{B}$ of subsets of Y, we have*

(1) $\operatorname{cl}_U\left(\bigcap \mathscr{B} \right) \subseteq \bigcap_{B \in \mathscr{B}} \operatorname{cl}_U(B)$*;* *(2)* $\bigcup_{B \in \mathscr{B}} \operatorname{int}_U(B) \subseteq \operatorname{int}_U\left(\bigcup \mathscr{B} \right)$.

The fact that the corresponding equalities need not be true even in the most simple cases is apparent from the following modification of Example 16.11.

Example 18.11 If in particular $X = B_1 \cup B_2$, with $B_i = \{i\}$, and U is a corelation on X such that $U(B_1) = B_1$ and $U(B_2) = X$, then

$$\operatorname{int}_U(B_1 \cup B_2) = X, \qquad \text{but} \qquad \operatorname{int}_U(B_1) \cup \operatorname{int}_U(B_2) = \emptyset.$$

However, it now more important to note that, by the corresponding definitions, we also have the following

Theorem 18.12 *If U is a corelation on X to Y, then for any $x \in X$ and $B \subseteq Y$ we have*

(1) $x \in \operatorname{int}_U(B) \iff U^{\triangleleft}(x) \subseteq B$*;*

(2) $x \in \operatorname{cl}_U(B) \iff U^{\triangleleft}(x) \cap B \neq \emptyset$.

Proof To check (1), note that, by Definitions 18.1, 16.1 and 8.1, we have

$$x \in \operatorname{int}_U(B) \iff \{x\} \in \operatorname{Int}_U(B) \iff U(\{x\}) \subseteq B \iff U^{\lhd}(x) \subseteq B.$$

Hence, by using Definition 13.1 and Theorem 18.3, we can immediately derive

Corollary 18.13 *If U is a corelation on X to Y, then for any $B \subseteq Y$ we have*

(1) $\operatorname{cl}_U(B) = U^{-1}(B)$; *(2)* $\operatorname{int}_U(B) = U^{-1}(B^c)^c$.

Proof To check (1), note that, by Theorem 18.12 and the corresponding definitions, for any $x \in X$ we have

$$x \in \operatorname{cl}_U(B) \iff U^{\lhd}(x) \cap B \neq \emptyset \iff x \in U^{\lhd -1}[B]$$
$$\iff x \in U^{\lhd -1 \rhd}(B) \iff x \in U^{-1}(B).$$

Remark 18.14 Therefore, for any corelation U on X to Y, we actually have

(1) $\operatorname{cl}_U = U^{-1}$; (2) $\operatorname{int}_U = U^{-1c} \circ \mathscr{C}_Y$.

By using Theorem 18.12, we can also easily prove the following

Theorem 18.15 *If U is a quasi-increasing corelation on X to Y, then for any $A \subseteq X$ and $B \subseteq Y$ we have*

(1) $A \in \operatorname{Int}_U(B) \implies A \subseteq \operatorname{int}_U(B)$;
(2) $A \cap \operatorname{cl}_U(B) \neq \emptyset \implies A \in \operatorname{Cl}_U(B)$.

Proof Assertion (2) can, in principle, be immediately derived from (1) by using Theorems 16.3 and 18.3. However, it can certainly be more easily proved by using the corresponding definitions.

For this, it is enough to note only that if $A \cap \operatorname{cl}_U(B) \neq \emptyset$, then there exists $x \in A$ such that $x \in \operatorname{cl}_U(B)$. Therefore, by Theorem 18.12, we also have $U^{\lhd}(x) \cap B \neq \emptyset$. Hence, since now $U^{\lhd}(x) = U\big(\{x\}\big) \subseteq U(A)$, we can infer that $U(A) \cap B \neq \emptyset$, and thus $A \in \operatorname{Cl}_U(B)$ also holds.

Now, by using Theorem 18.12, we can also easily prove the following

Theorem 18.16 *For any two corelations U and V on X to Y, the following assertions are equivalent:*

(1) $U^{\lhd} \subseteq V^{\lhd}$; *(2)* $\operatorname{cl}_U \subseteq \operatorname{cl}_V$ *(3)* $\operatorname{int}_V \subseteq \operatorname{int}_U$.

Proof To check the implication (3) $\implies$ (1), note that by the Theorem 18.12, for any $x \in X$ we have $x \in \operatorname{int}_V\big(V^{\lhd}(x)\big)$. Therefore, if (3) holds, then we also have $x \in \operatorname{int}_U\big(V^{\lhd}(x)\big)$. Hence, by using Theorem 18.12, we can already infer that $U^{\lhd}(x) \subseteq V^{\lhd}(x)$. Therefore, by Theorem 3.1, assertion (1) also holds.

Now, as an immediate consequence of this theorem, we can also state

Corollary 18.17 *For any two corelations U and V on X to Y, the following assertions are equivalent:*

(1) $U^{\lhd} = V^{\lhd}$*;* *(2)* $\operatorname{cl}_U = \operatorname{cl}_V$ *(3)* $\operatorname{int}_V = \operatorname{int}_U$.

Remark 18.18 Note that if both U and V are union-preserving, then by Corollary 8.13 we have $U \leq V \iff U^{\lhd} \subseteq V^{\lhd}$.

While, if U is an arbitrary and V is a quasi-increasing corelation on X to Y, then by Theorem 9.8 we have $U^{\circ} \leq V \iff U^{\lhd} \subseteq V^{\lhd}$.

Remark 18.19 Hence, by Theorem 18.16, we can see that if U is an arbitrary and V is a quasi-increasing corelation on X, then

$$\operatorname{cl}_V \supseteq \operatorname{cl}_U \iff V \geq U^{\circ}.$$

Therefore, the mappings

$$V \longmapsto \operatorname{cl}_V \qquad \text{and} \qquad V \longmapsto V^{\circ},$$

where V is a quasi-increasing corelation on X to Y, establish a Pataki connection with respect to the inverses of the relations $\leq$ and $\subseteq$.

19 Topological Interiors and Closures Derived from Relations

Now, analogously to Definition 17.1, we may also naturally have the following

Definition 19.1 For any relation R on X to Y, the relations

$$\operatorname{cl}_R = \operatorname{cl}_{R^{\rhd}} \qquad \text{and} \qquad \operatorname{int}_R = \operatorname{int}_{R^{\rhd}}$$

will be called the *topological closure* and *topological interior* generated by the relation R, respectively.

Thus, by the corresponding results of Sect. 18, we can at once state the following assertions.

Theorem 19.2 *If R is a relation on X to Y, then for any $B \subseteq Y$ we have*

(1) $\operatorname{cl}_R(B) = X \setminus \operatorname{int}_R(Y \setminus B)$*;* *(2)* $\operatorname{int}_R(B) = X \setminus \operatorname{cl}_R(Y \setminus B)$.

Corollary 19.3 *For any relation R on X to Y, we have*

(1) $\operatorname{int}_R = \big(\operatorname{cl}_R \circ \mathscr{C}_Y\big)^c = \big(\operatorname{cl}_R\big)^c \circ \mathscr{C}_Y$*;*

(2) $\operatorname{cl}_R = \big(\operatorname{int}_R \circ \mathscr{C}_Y\big)^c = \big(\operatorname{int}_R\big)^c \circ \mathscr{C}_Y$.

Remark 19.4 The above results show that the relations cl_R and int_R are equivalent tools in the relational space $(X, Y)(R)$.

Theorem 19.5 *If R is a relation on X to Y, then*

(1) cl_R *is union-preserving;* *(2)* int_R *is intersection-preserving.*

Corollary 19.6 *For any relation R on X to Y, we have*

(1) $\mathrm{cl}_R(\emptyset) = \emptyset$; *(2)* $\mathrm{int}_R(Y) = X$.

Corollary 19.7 *If R is a relation on X to Y, then the corelations* cl_R *and* int_R *are increasing.*

Theorem 19.8 *If R is a relation on X to Y, then for any family $\mathscr{B}$ of subsets of Y, we have*

(1) $\mathrm{cl}_R\Big(\bigcap \mathscr{B}\Big) \subseteq \bigcap_{B \in \mathscr{B}} \mathrm{cl}_R(B)$; *(2)* $\bigcup_{B \in \mathscr{B}} \mathrm{int}_R(B) \subseteq \mathrm{int}_R\Big(\bigcup \mathscr{B}\Big)$.

The fact that the corresponding equalities need not be true even in the most simple cases is apparent from the following counterpart of Example 18.11.

Example 19.9 If in particular $X = B_1 \cup B_2$, with $B_i = \{i\}$, and R is a relation on X such that $R(1) = \{1\}$ and $R(2) = X$, then

$$\mathrm{int}_R(B_1 \cup B_2) = X, \qquad \text{but} \qquad \mathrm{int}_U(B_1) \cup \mathrm{int}_U(B_2) = \emptyset.$$

However, it is now more important to note that, by the corresponding definitions, we also have the following

Theorem 19.10 *If R is a relation on X to Y, then for any $x \in X$ and $B \subseteq Y$ we have*

(1) $x \in \mathrm{cl}_R(B) \iff \{x\} \in \mathrm{Cl}_R(B)$;
(2) $x \in \mathrm{int}_R(B) \iff \{x\} \in \mathrm{Int}_R(B)$.

Proof To check (2), note that by Definitions 19.1, 18.1 and 17.1 we have

$$x \in \mathrm{int}_R(B) \iff x \in \mathrm{int}_{R^{\triangleright}}(B) \iff \{x\} \in \mathrm{Int}_{R^{\triangleright}}(B) \iff \{x\} \in \mathrm{Int}_R(B).$$

From this theorem, by using Theorem 17.11 and the fact that $R\big(\{x\}\big) = R(x)$, we can immediately derive

Corollary 19.11 *If R is a relation on X to Y, then for any $x \in X$ and $B \subseteq Y$ we have*

(1) $x \in \mathrm{int}_R(B) \iff R(x) \subseteq B$;
(2) $x \in \mathrm{cl}_R(B) \iff R(x) \cap B \neq \emptyset$.

By using the corresponding definitions and Theorem 19.2, the latter corollary can be reformulated in the following more concise form.

Corollary 19.12 *If R is a relation on X to Y, then for any $B \subseteq Y$ we have*

(1) $\operatorname{cl}_R(B) = R^{-1}[\,B\,]$, *(2)* $\operatorname{int}_R(B) = R^{-1}[\,B^c\,]^c$.

Remark 19.13 Therefore, for any relation R on X to Y, we actually have

(1) $\operatorname{cl}_R = R^{-1\triangleright}$; (2) $\operatorname{int}_R = R^{-1\triangleright c} \circ \mathscr{C}_Y$.

The following theorem shows that, in contrast to Example 17.14, corelations cannot generate more general topological closures and interiors than relations.

Therefore, our results on the relations cl_R and int_R, established in several former papers, cannot be generalized by using corelations instead of relations.

Theorem 19.14 *For any corelation U on X to Y, we have*

(1) $\operatorname{cl}_U = \operatorname{cl}_{U^\circ} = \operatorname{cl}_{U^\lhd}$*;* *(2)* $\operatorname{int}_U = \operatorname{int}_{U^\circ} = \operatorname{int}_{U^\lhd}$.

Proof From Theorem 17.12, by using Definition 18.1, we can at once see that $\operatorname{int}_{U^\circ} = \operatorname{int}_{U^\lhd}$.

Moreover, by Theorem 18.12 and Corollary 19.11, it is clear that, for any $x \in X$ and $B \subseteq Y$, we have

$$x \in \operatorname{int}_U(B) \iff U^\lhd(x) \subseteq B \iff x \in \operatorname{int}_{U^\lhd}(B).$$

Therefore, $\operatorname{int}_U(B) = \operatorname{int}_{U^\lhd}(B)$ for all $B \subseteq Y$, and thus $\operatorname{int}_U = \operatorname{int}_{U^\lhd}$. This shows that assertion (2) is true.

Assertion (1) can now be immediately derived from assertion (2), by using Theorems 18.3 and 19.2.

In this respect, it is also worth noticing that we also have the following

Theorem 19.15 *If U is a corelation on X to Y, then for any $A \subseteq X$ and $B \subseteq Y$ we have*

(1) $A \subseteq \operatorname{int}_U(B) \iff A \in \operatorname{Int}_{U^\lhd}(B)$*;*
(2) $A \cap \operatorname{cl}_U(B) \neq \emptyset \iff A \in \operatorname{Cl}_{U^\lhd}(B)$.

Proof By Theorems 17.12, 18.15 and 19.14, it is clear that

$$A \in \operatorname{Int}_{U^\lhd}(B) \implies A \in \operatorname{Int}_{U^\circ}(B) \implies A \subseteq \operatorname{int}_{U^\circ}(B) \implies A \subseteq \operatorname{int}_U(B).$$

Namely, from Theorem 9.2 and Corollary 5.10, we know that U° is quasi-increasing.

Conversely, if $A \subseteq \operatorname{int}_U(B)$, then for each $x \in A$ we have $x \in \operatorname{int}_U(B)$. Hence, by Theorem 18.12, we can infer that $U^\lhd(x) \subseteq B$. Therefore, we have

$$U^\lhd[\,A\,] = \bigcup_{x \in A} U^\lhd(x) \subseteq B.$$

Hence, by Theorem 17.11, we can infer that $A \in \operatorname{Int}_{U^\lhd}(B)$. Therefore, assertion (1) is true.

Assertion (2) can now be easily derived from assertion (1) by using Theorems 18.3 and 16.3 and 17.2.

Remark 19.16 The above two theorems show that, for instance, the relations int_U, $\operatorname{int}_{U^\circ}$, $\operatorname{int}_{U^\lhd}$, $\operatorname{Int}_{U^\circ}$ and $\operatorname{Int}_{U^\lhd}$ are also equivalent tools in the corelational space $(X, Y)(U)$.

Now, by using Theorem 18.16, we can also easily prove the following

Theorem 19.17 *For any two relations R and S on X to Y, the following assertions are equivalent:*

(1) $R \subseteq S$; *(2)* $\operatorname{cl}_R \subseteq \operatorname{cl}_S$ *(3)* $\operatorname{int}_S \subseteq \operatorname{int}_R$.

Proof To check the equivalence of (1) and (2), note that by Theorems 8.8 and 18.16 and Definition 19.1 we have

$$R \subseteq S \iff R^{\rhd\lhd} \subseteq S^{\rhd\lhd} \iff \operatorname{cl}_{R^\rhd} \subseteq \operatorname{cl}_{S^\rhd} \iff \operatorname{cl}_R \subseteq \operatorname{cl}_S\,.$$

Remark 19.18 Note that U and V are corelations on X to Y such that $\operatorname{cl}_U \subseteq \operatorname{cl}_V$, then by Theorem 19.14 we also have $\operatorname{cl}_{U^\lhd} \subseteq \operatorname{cl}_{V^\lhd}$. Hence, by using the above theorem, we can infer that $U^\lhd \subseteq U^\lhd$. Therefore, Theorem 18.16 and 19.17 are actually equivalent.

Now, as an immediate consequence of Theorem 19.17, we can also state

Corollary 19.19 *For any two relations R and S on X to Y, each of the equalities* $\operatorname{cl}_R = \operatorname{cl}_S$ *and* $\operatorname{int}_R = \operatorname{int}_S$ *implies $R = S$.*

Remark 19.20 By Theorem 19.17, for any two relations R and S on X to Y, we have

$$\operatorname{cl}_R \subseteq \operatorname{cl}_S \iff U \subseteq S.$$

Therefore, the mappings

$$R \longmapsto \operatorname{cl}_R \qquad \text{and} \qquad R \longmapsto R,$$

where R is a relation on X to Y, establish a Pataki connection.

Moreover, this Pataki connection is very particular since the first mapping is injective and the second one is injective and onto.

20 Fat and Dense Sets Derived from Corelations

Definition 20.1 For any corelation U on X to Y, the members of the families

$$\mathscr{E}_U = \big\{ B \subseteq Y : \ \operatorname{int}_U(B) \neq \emptyset \big\} \qquad \text{and} \qquad \mathscr{D}_U = \big\{ B \subseteq Y : \ \operatorname{cl}_U(B) = X \big\}$$

will be called the *fat sets* and *dense sets* generated by the corelation U, respectively.

Thus, by Corollary 18.13 and Theorem 18.12, we evidently have

Theorem 20.2 *If U is a corelation on X to Y, then for any $B \subseteq Y$ we have*

(1) $B \in \mathscr{D}_U \iff U^{-1}[\,B\,] = X$; *(2)* $B \in \mathscr{E}_R \iff U^{-1}[\,B^c\,] \neq X$.

Theorem 20.3 *If U is a corelation on X to Y, then for any $B \subseteq Y$ we have*

(1) $B \in \mathscr{E}_U$ *if and only if* $U^{\lhd}(x) \subseteq B$ *for some* $x \in X$*;*
(2) $B \in \mathscr{D}_U$ *if and only if* $U^{\lhd}(x) \cap B \neq \emptyset$ *for all* $x \in X$.

From Theorem 20.2, we can immediately derive the following theorem which can also be easily proved with the help of Theorem 18.3.

Theorem 20.4 *If U is a corelation on X to Y, then for any $B \subseteq Y$ we have*

(1) $B \in \mathscr{E}_U \iff B^c \notin \mathscr{D}_U$; *(2)* $B \in \mathscr{D}_U \iff B^c \notin \mathscr{E}_U$.

By using appropriate complementations, this theorem can be written in the following more concise form.

Corollary 20.5 *For any corelation U on X to Y, we have*
(1) $\mathscr{E}_U = \mathscr{C}_Y[\,\mathscr{D}_U^c\,]$; *(2)* $\mathscr{D}_U = \mathscr{C}_Y[\,\mathscr{E}_U^c\,]$.

Remark 20.6 By the corresponding definitions, for any family $\mathscr{B}$ of subsets of Y, we have $\mathscr{B}^c = \mathscr{P}(X) \setminus \mathscr{B}$ and $\mathscr{C}_Y[\,\mathscr{B}\,] = \big\{B^c : \ B \in \mathscr{B}\big\}$ with $B^c = Y \setminus B$.

Thus, the ordinary and elementwise complements of $\mathscr{B}$ are quite different sets. However, sometimes the family $\mathscr{C}_Y[\,\mathscr{B}\,]$ may also be naturally denoted by $\mathscr{B}^c$.

The following theorem can, in principle, be derived from Theorem 20.4. However, it can be more easily proved with the help of Theorem 20.3. Here, we shall give a mixed proof.

Theorem 20.7 *If U is a corelation on X to Y, then for any $B \subseteq Y$ we have*

(1) $B \in \mathscr{E}_U$ *if and only if* $B \cap D \neq \emptyset$ *for all* $D \in \mathscr{D}_U$*;*
(2) $B \in \mathscr{D}_U$ *if and only if* $B \cap E \neq \emptyset$ *for all* $E \in \mathscr{E}_U$.

Proof To check (2), note that if $B \in \mathscr{D}_U$ and $E \in \mathscr{E}_U$, then by Theorem 20.3 there exists $x \in X$ such that $U^{\lhd}(x) \subseteq E$. Moreover, by Theorem 18.12, we have $U^{\lhd}(x) \cap B \neq \emptyset$. Therefore, $B \cap E \neq \emptyset$ also holds.

While, if $B \cap E \neq \emptyset$ for all $E \in \mathscr{E}_U$, then by using Theorem 20.4 we can easily see that $B \in \mathscr{D}_U$. Namely, if $B \notin \mathscr{D}_U$, then by Theorem 20.4 we have $B^c \in \mathscr{E}_U$. Therefore, $B \cap B^c \neq \emptyset$, which is a contradiction.

Remark 20.8 The above two theorems shows that the families $\mathscr{E}_U$ and $\mathscr{D}_U$ are equivalent tools in the corelational space $(X,\ Y)(U)$.

The following theorem is also more immediate from Theorem 20.3 than from Corollary 18.9 or Theorem 20.7.

Theorem 20.9 *If U is a corelation on X to Y, then $\mathscr{E}_U$ and $\mathscr{D}_U$ are ascending families in $\mathscr{P}(Y)$.*

Remark 20.10 A subfamily $\mathscr{A}$ of the poset $\mathscr{P}(X)$ is called ascending if $A \in \mathscr{A}$ and $A \subseteq B \in \mathscr{P}\,(X)$ imply $B \in \mathscr{A}$. That is, $\mathscr{P}^{-1}[\,\mathscr{A}\,] \subseteq \mathscr{A}$, and thus $\mathscr{A} = \mathscr{P}^{-1}[\,\mathscr{A}\,]$.

In the literature, an ascending subfamily $\mathscr{A}$ of $\mathscr{P}(X)$ is usually are called a *stack* in X. Moreover, it is called proper if $\emptyset \notin \mathscr{A}$, or equivalently $\mathscr{A} \neq \mathscr{P}(X)$.

A stack $\mathscr{A}$ in X is called a *filter* if $A_1, A_2 \in \mathscr{A}$ implies $A_1 \cap A_2 \in \mathscr{A}$. While, $\mathscr{A}$ is called a *grill* if $A_1 \cup A_2 \in \mathscr{A}$ implies that either $A_1 \in \mathscr{A}$ or $A_2 \in \mathscr{A}$. They are usually assumed to be proper and nonvoid.

Several interesting facts on the discoveries and applications of stacks, filters and grills and can be found in the fundamental works [84, 85] of Thron. In particular, it is noteworthy that nets and filters are essentially equivalent tools.

By using Theorems 20.3 and 20.4, in addition to Theorem 20.7, we can also easily prove the following two theorems.

Theorem 20.11 *For any corelation U on X to Y, the following assertions are equivalent:*

(1) $\mathscr{E}_U \neq \emptyset$; *(2)* $Y \in \mathscr{E}_U$;
(3) $\emptyset \notin \mathscr{D}_U$; *(4)* $\mathscr{D}_U \neq \mathscr{P}(Y)$; *(5)* $X \neq \emptyset$.

Theorem 20.12 *For any corelation U on X to Y, the following assertions are equivalent:*

(1) $\emptyset \notin \mathscr{E}_U$; *(2)* $\mathscr{E}_U \neq \mathscr{P}(Y)$;
(3) $\mathscr{D}_U \neq \emptyset$; *(4)* $X \in \mathscr{D}_U$; *(5)* $X = U^{\mathbf{-1}}(Y)$.

Proof To check the equivalence of (1) and (5), note that, by Theorem 20.3, we have

$$\begin{aligned} \emptyset \notin \mathscr{E}_U \iff \forall\; x \in X:\;\; U^{\lhd}(x) \not\subseteq \emptyset \iff \forall\; x \in X:\;\; U^{\lhd}(x) \neq \emptyset \\ \iff \forall\; x \in X:\;\; U^{\lhd}(x) \cap Y \neq \emptyset \iff \forall\; x \in X:\;\; x \in U^{\lhd\,-1}[Y] \\ \iff X = U^{\lhd\,-1}[Y] \iff X = U^{\lhd\,-1\,\rhd}(Y) \iff X = U^{\mathbf{-1}}(Y) \end{aligned}$$

Remark 20.13 A subfamily $\mathscr{B}$ of a stack $\mathscr{A}$ in X is called a *base* of $\mathscr{A}$ if for each $A \in \mathscr{A}$ there exists $B \in \mathscr{B}$ such that $B \subseteq A$.

Note that, for any family $\mathscr{B}$ of subsets of X, the family

$$\mathscr{B}^{*} = \big\{ A \subseteq X : \quad \exists\; B \in \mathscr{B} : \quad B \subseteq A \big\}$$

is a stack in X such that $\mathscr{B}$ is a base of $\mathscr{B}^{*}$.

Moreover, it is also noteworthy that now, by Corollary 19.11, for any $A \subseteq X$ we have

$$A \in \mathscr{B}^{*} \iff \exists\; B \in \mathscr{B} :\;\; B \in \mathscr{P}(A) \iff \mathscr{P}(A) \cap \mathscr{B} \neq \emptyset \iff A \in \mathrm{cl}_{\mathscr{P}}(\mathscr{B}).$$

Therefore, $\mathscr{B}^{*} = \mathrm{cl}_{\mathscr{P}}(\mathscr{B})$.

Now, as an important addition to Theorem 20.9, we can also prove the following

Theorem 20.14 *For any corelation U on X to Y, the stack $\mathscr{E}_U$ has a base $\mathscr{B}$ with* card $(\mathscr{B}) \leq$ card (X).

Proof By Theorem 20.3, it is clear that the family $\mathscr{B} = \{ U^{\triangleleft}(x) : \quad x \in X \}$ is a base of the stack $\mathscr{E}_U$.

Moreover, the function f, defined by $f(x) = U^{\triangleleft}(x)$ for all $x \in X$, is onto $\mathscr{B}$. Hence, by the axiom of choice, the cardinality condition follows.

Namely, now f^{-1} is a relation of $\mathscr{B}$ to X. Hence, by choosing a selection φ of f^{-1}, we can see that φ is an injection of $\mathscr{B}$ to X.

Remark 20.15 Now, a corresponding property of the family $\mathscr{D}_U$ could, in principle, be derived from the above theorem by using either Theorem 20.4 or 20.7.

However, it is now more important to note that, as an interesting counterpart of Theorem 18.16, we can also prove the following

Theorem 20.16 *For any two corelations U and V on X to Y, the following assertions are equivalent:*

(1) $\mathscr{E}_U \subseteq \mathscr{E}_V$; *(2)* $\mathscr{D}_V \subseteq \mathscr{D}_U$;
(3) $\left(V^{\triangleleft -1} \circ U^{\triangleleft c} \right)(x) \neq X$ *for all* $x \in X$;
(4) $V^{\triangleleft} \circ \varphi \subseteq U^{\triangleleft}$ *for some function φ of X to itself;*
(5) $V^{\triangleleft} \circ \Phi \subseteq U^{\triangleleft}$ *for some relation Φ of X to itself;*
(6) for every $x \in X$ there exists $v \in X$ such that $V^{\triangleleft}(v) \subseteq U^{\triangleleft}(x)$.

Proof It is clear that (4) $\Longrightarrow$ (5) $\Longrightarrow$ (6). Moreover, by using the axiom of choice, we can also at once see that (6) implies (4). Therefore, assertions (4), (5) and (6) are equivalent.

Furthermore, by using Theorem 20.3, we can easily see that (1) and (2) are also equivalent. Moreover, by Theorem 20.3, it is clear that (6) implies (1).

On the other hand, if $x \in X$, then by Theorem 20.3 we have $U^{\triangleleft}(x) \in \mathscr{E}_U$. Therefore, if (1) holds, then we also have $U^{\triangleleft}(x) \in \mathscr{E}_V$. Thus, by Theorem 20.3, there exists $v \in X$ such that $V^{\triangleleft}(v) \subseteq U^{\triangleleft}(x)$. Therefore, (6) also holds.

While, if (6) holds, then for any $x \in X$ there exists $v \in X$ such that

$$V^{\triangleleft}(v) \cap U^{\triangleleft}(x)^c = \emptyset, \qquad \text{and thus} \qquad V^{\triangleleft}(v) \cap U^{\triangleleft c}(x) = \emptyset.$$

Hence, we can infer that

$$v \notin V^{\triangleleft -1}[\, U^{\triangleleft c}(x)\,], \qquad \text{and thus} \qquad v \notin \left(V^{\triangleleft -1} \circ U^{\triangleleft c} \right)(x).$$

Therefore, (3) also holds.

Now, to complete the proof, we need only note that the converse implication (3) $\Longrightarrow$ (6) can be proved quite similarly by reversing the above argument.

21 Fat and Dense Sets Derived from Relations

Definition 21.1 For any relation R on X to Y, the members of the families

$$\mathscr{E}_R = \mathscr{E}_{R^{\triangleright}} \qquad \text{and} \qquad \mathscr{D}_R = \mathscr{D}_{R^{\triangleright}}$$

will be called the *fat sets* and *dense sets* generated by the relation R, respectively.

Thus, by the corresponding definitions, we evidently have the following

Theorem 21.2 *If R is a relation on X to Y, then for any $B \subseteq Y$ we have*

(1) $B \in \mathscr{E}_R \iff \operatorname{int}_R(B) \neq \emptyset$; *(2)* $B \in \mathscr{D}_R \iff \operatorname{cl}_R(B) = X$.

Proof To check (1), note that by Definitions 21.1, 20.1 and 19.1, we have

$$B \in \mathscr{E}_R \iff B \in \mathscr{E}_{R^{\triangleright}} \iff \operatorname{int}_{R^{\triangleright}}(B) \neq \emptyset \iff \operatorname{int}_R(B) \neq \emptyset.$$

From this theorem, by using Corollaries 19.12 and 19.11, we can immediately derive the following two corollaries.

Corollary 21.3 *If R is a relation on X to Y, then for any $B \subseteq Y$ we have*

(1) $B \in \mathscr{D}_R \iff R^{-1}[\,B\,] = X$;
(2) $B \in \mathscr{E}_R \iff R^{-1}[\,B^c\,] \neq X$.

Corollary 21.4 *If R is a relation on X to Y, then for any $B \subseteq Y$ we have*

(1) $B \in \mathscr{E}_R$ *if and only if* $R\,(x) \subseteq B$ *for some* $x \in X$;
(2) $B \in \mathscr{D}_R$ *if and only if* $R\,(x) \cap B \neq \emptyset$ *for all* $x \in X$.

Moreover, by using Theorems 21.2 and 19.14, we can also prove the following

Theorem 21.5 *For any corelation U on X to Y, we have*

(1) $\mathscr{E}_U = \mathscr{E}_{U^{\circ}} = \mathscr{E}_{U^{\triangleleft}}$; *(2)* $\mathscr{D}_U = \mathscr{D}_{U^{\circ}} = \mathscr{D}_{U^{\triangleleft}}$.

Proof To check (1), note that by Definitions 9.1 and 21.1 we have

$$\mathscr{E}_{U^{\circ}} = \mathscr{E}_{U^{\triangleleft\triangleright}} = \mathscr{E}_{U^{\triangleleft}}.$$

Moreover, by Theorems 21.2 and 19.14, for any $B \subseteq Y$, we have

$$B \in \mathscr{E}_{U^{\triangleleft}} \iff \operatorname{int}_{U^{\triangleleft}}(B) \neq \emptyset \iff \operatorname{int}_U(B) \neq \emptyset \iff B \in \mathscr{E}_U.$$

Therefore, $\mathscr{E}_{U^{\triangleleft}} = \mathscr{E}_U$ is also true.

Remark 21.6 This theorem shows that corelations cannot generate more fat and dense sets than relations.

From the results of Sect. 20, by Definition 21.1, it is clear that we can also state the following theorems.

Theorem 21.7 *If R is a relation on X to Y, then for any $B \subseteq Y$ we have*

(1) $B \in \mathscr{E}_R \iff B^c \notin \mathscr{D}_R$; *(2)* $B \in \mathscr{D}_R \iff B^c \notin \mathscr{E}_R$.

Corollary 21.8 *For any relation R on X to Y, we have*
(1) $\mathscr{E}_R = \mathscr{C}_Y[\,\mathscr{D}_R^c]$; *(2)* $\mathscr{D}_R = \mathscr{C}_Y[\,\mathscr{E}_R^c]$.

Theorem 21.9 *If R is a relation on X to Y, then for any $B \subseteq Y$ we have*

(1) $B \in \mathscr{E}_R$ *if and only if* $B \cap D \neq \emptyset$ *for all* $D \in \mathscr{D}_R$;
(2) $B \in \mathscr{D}_R$ *if and only if* $B \cap E \neq \emptyset$ *for all* $E \in \mathscr{E}_R$.

Remark 21.10 The above two theorems shows that the families $\mathscr{E}_R$ and $\mathscr{D}_R$ are equivalent tools in the relational space $(X,\ Y)(R)$.

Theorem 21.11 *If U is a corelation on X to Y, then $\mathscr{E}_U$ and $\mathscr{D}_U$ are stacks in Y.*

Theorem 21.12 *For any relation R on X to Y, the following assertions are equivalent:*

(1) $\mathscr{E}_R \neq \emptyset$; *(2)* $Y \in \mathscr{E}_R$;
(3) $\emptyset \notin \mathscr{D}_R$; *(4)* $\mathscr{D}_R \neq \mathscr{P}(Y)$; *(5)* $X \neq \emptyset$.

Theorem 21.13 *For any relation R on X to Y, the following assertions are equivalent:*

(1) $\emptyset \notin \mathscr{E}_R$; *(2)* $\mathscr{E}_R \neq \mathscr{P}(Y)$;
(3) $\mathscr{D}_R \neq \emptyset$; *(4)* $X \in \mathscr{D}_R$; *(5)* $X = R^{-1}[\,Y\,]$.

Theorem 21.14 *For any relation R on X to Y, the stack $\mathscr{E}_R$ has a base $\mathscr{B}$ with* card$\,(\mathscr{B}) \leq$ card$\,(X)$.

Remark 21.15 The importance of the study of the cardinalities of the bases of the stack of all fat sets in a relator space, concerning a problem of mine on paratopologically simple relators, was first recognized by J. Deák (1994) and G. Pataki (1998). (For the corresponding results, see [41].)

Theorem 21.16 *For any two relations R and S on X to Y, the following assertions are equivalent:*

(1) $\mathscr{E}_R \subseteq \mathscr{E}_S$; *(2)* $\mathscr{D}_S \subseteq \mathscr{D}_R$;
(3) $\left(S^{-1} \circ R^c\right)(x) \neq X$ *for all* $x \in X$;
(4) $S \circ \varphi \subseteq R$ *for some function φ of X to itself;*
(5) $S \circ \Phi \subseteq R$ *for some relation Φ of X to itself;*
(6) for every $x \in X$ there exists $v \in X$ such that $S(v) \subseteq R(x)$.

Proof To derive this from Theorem 20.16, note that by Definition 21.1 we have

$$\mathscr{E}_R \subseteq \mathscr{E}_S \iff \mathscr{E}_{R^{\triangleright}} \subseteq \mathscr{E}_{S^{\triangleright}}.$$

Moreover, by Theorem 8.8, we have $R^{\triangleright\triangleleft} = R$ and $S^{\triangleright\triangleleft} = S$.

Remark 21.17 In our former papers [51, 57, 58], motivated by a standard definition for families of sets [10, p. 339], a relator $\mathscr{S}$ on X was said to be uniformly refined by a relator $\mathscr{R}$ on X if for every $S \in \mathscr{S}$ there exist $R \in \mathscr{R}$ and a function φ of X to itself such that $R \subseteq S \circ \varphi$.

22 Open and Closed Sets Derived from Corelations

Definition 22.1 For any corelation U on X, the members of the families

$$\mathscr{T}_U = \big\{ A \subseteq X : \quad A \subseteq \operatorname{int}_U(A) \big\} \qquad \text{and} \qquad \mathscr{F}_U = \big\{ A \subseteq X : \quad \operatorname{cl}_U(A) \subseteq A \big\}$$

will be called the *open sets* and *closed sets* generated by the corelation U, respectively.

Thus, by Theorem 18.3, we evidently have the following

Theorem 22.2 *If U is a corelation on X, then for any $A \subseteq X$ we have*

(1) $A \in \mathscr{T}_U \iff A^c \in \mathscr{F}_U$; *(2)* $A \in \mathscr{F}_U \iff A^c \in \mathscr{T}_U$.

Now, by using elementwise complementation, we can also state

Corollary 22.3 *For any corelation U on X, we have*

(1) $\mathscr{T}_U = \mathscr{F}_U^{\,c}$; *(2)* $\mathscr{F}_U = \mathscr{T}_U^{\,c}$.

From Theorems 18.7 and 18.10, by using Definition 22.1, we can easily derive

Theorem 22.4 *For any corelation U on X, the families $\mathscr{T}_U$ and $\mathscr{F}_U$ are closed under arbitrary unions and intersections.*

Proof To check this for the family $\mathscr{T}_U$, note that if $\mathscr{A} \subseteq \mathscr{T}_U$, then by Definition 22.1 we have $A \subseteq \operatorname{int}_U(A)$ for all $A \in \mathscr{A}$. Hence, by using Theorems 18.7 and 18.10, we can infer that

$$\bigcap \mathscr{A} \subseteq \bigcap_{A \in \mathscr{A}} \operatorname{int}_U(A) = \operatorname{int}_U\Big(\bigcap \mathscr{A}\Big) \quad \text{and}$$

$$\bigcup \mathscr{A} \subseteq \bigcup_{A \in \mathscr{A}} \operatorname{int}_U(A) \subseteq \operatorname{int}_U\Big(\bigcup \mathscr{A}\Big).$$

Therefore, by Definition 22.1, the inclusions $\bigcap \mathscr{A} \in \mathscr{T}_U$ and $\bigcup \mathscr{A} \in \mathscr{T}_U$ also hold.

From this theorem, by taking empty union and intersection, we obtain

Corollary 22.5 *For any corelation U on X, we have* $\{\emptyset,\ X\} \subseteq \mathscr{T}_U \cap \mathscr{F}_U$.

By Definition 22.1 and Theorem 19.14, we can also at once state

Theorem 22.6 *If U is a corelation on X, then for any $A \subseteq X$ we have*

(1) $A \in \mathscr{F}_U \iff \mathrm{cl}_{U^\circ}(A) \subseteq A \iff \mathrm{cl}_{U^\triangleleft}(A) \subseteq A$;
(2) $A \in \mathscr{T}_U \iff A \subseteq \mathrm{int}_{U^\circ}(A)\,\} \iff A \subseteq \mathrm{int}_{U^\triangleleft}(A)$.

Thus, by Definition 22.1, we can also state

Corollary 22.7 *For any corelation U on X, we have*

(1) $\mathscr{T}_U = \mathscr{T}_{U^\circ}$; *(2)* $\mathscr{F}_U = \mathscr{F}_{U^\circ}$.

Moreover, from Theorem 22.6, by Corollary 19.11, it is clear that we also have

Theorem 22.8 *If U is a corelation on X, then for any $A \subseteq X$ we have*

(1) $A \in \mathscr{T}_U$ *if and only if* $U^\triangleleft(x) \subseteq A$ *for all* $x \in A$;
(2) $A \in \mathscr{F}_U$ *if and only if* $A \cap U^\triangleleft(x) \neq \emptyset$ *implies* $x \in A$ *for all* $x \in X$.

The latter assertions can be written in the following more concise forms.

Corollary 22.9 *If U is a corelation on X, then for any $A \subseteq X$ we have*

(1) $A \in \mathscr{T}_U \iff U^\triangleleft[\,A\,] \subseteq A$; *(2)* $A \in \mathscr{F}_U \iff U^{\triangleleft -1}[\,A\,] \subseteq A$.

Hence, by using Theorems 17.11 and 17.12, we can easily derive

Theorem 22.10 *If U is a corelation on X, then for any $A \subseteq X$ we have*

(1) $A \in \mathscr{T}_U \iff A \in \mathrm{Int}_{U^\triangleleft}(A) \iff A \in \mathrm{Int}_{U^\circ}(A)$;
(2) $A \in \mathscr{F}_U \iff A^c \notin \mathrm{Cl}_{U^\triangleleft}(A) \iff A^c \notin \mathrm{Cl}_{U^\circ}(A)$.

Proof To prove assertion (2), note that by Theorem 22.2, assertion (1) and Theorems 16.3 and 17.12 we have

$$A \in \mathscr{F}_U \iff A^c \in \mathscr{T}_U \iff A^c \in \mathrm{Int}_{U^\triangleleft}(A^c)$$
$$\iff A^c \notin \mathrm{Cl}_{U^\triangleleft}(A) \iff A^c \notin \mathrm{Cl}_{U^\circ}(A).$$

Now, in particular, we can also state

Corollary 22.11 *If U is a corelation on X, then for any $A \subseteq X$ we have*

(1) $A \in \mathscr{T}_U \iff U^\circ(A) \subseteq A$; *(2)* $A \in \mathscr{F}_U \iff A \cap U^\circ(A^c) = \emptyset$.

By using Theorem 18.15, in addition to Theorem 22.10, we can also prove

Theorem 22.12 *If U is a quasi-increasing corelation on X, then for any $A \subseteq X$*

(1) $A \in \mathrm{Int}_U(A)$ *implies* $A \in \mathscr{T}_U$;
(2) $A \in \mathscr{F}_U$ *and* $\mathrm{cl}_U(A) \neq \emptyset$ *imply* $A \in \mathrm{Cl}_U(A)$.

Proof To prove assertion (2), note that if $A \in \mathscr{F}_U$, then by Definition 22.1 we have $\mathrm{cl}_U(A) \subseteq A$, and thus $A \cap \mathrm{cl}_U(A) = \mathrm{cl}_U(A)$. Hence, if $\mathrm{cl}_U(A) \neq \emptyset$, by using Theorem 18.15 we can infer that $A \in \mathrm{Cl}_U(A)$.

Moreover, by Definition 22.1 and Corollary 18.13, it is clear that we also have

Theorem 22.13 *If U is a corelation on X, then for any $A \subseteq X$ we have*

(1) $A \in \mathscr{F}_U \iff U^{-1}(A) \subseteq A$; *(2)* $A \in \mathscr{T}_U \iff A \subseteq U^{-1}(A^c)^c$.

Thus, by Definition 16.1 and Theorems 22.2 and 16.3, we can also state

Corollary 22.14 *If U is a corelation on X, then for any $A \subseteq X$ we have*

(1) $A \in \mathscr{F}_U \iff A \in \operatorname{Int}_{U^{-1}}(A)$; *(2)* $A \in \mathscr{T}_U \iff A^c \notin \operatorname{Cl}_{U^{-1}}(A)$.

By using Definitions 20.1 and 22.1, we can also easily prove the following

Theorem 22.15 *For any corelation U on X, we have*

(1) $\mathscr{T}_U \setminus \{\emptyset\} \subseteq \mathscr{E}_U$; *(2)* $\mathscr{D}_U \cap \mathscr{F}_U \subseteq \{X\}$.

Proof To prove (2), note that if $A \in \mathscr{D}_U \cap \mathscr{F}_U$, then $A \in \mathscr{D}_U$ and $A \in \mathscr{F}_U$. Therefore, by Definitions 20.1 and 22.1, we have $\operatorname{cl}_U(A) = X$ and $\operatorname{cl}_U(A) \subseteq A$, and thus $X \subseteq A$. Hence, it follows that $A = X$, and thus $A \in \{X\}$.

From assertion (2), we can immediately derive

Corollary 22.16 *For any corelation U on X, we have*

(1) $\mathscr{F}_U \subseteq \big(\mathscr{P}(X) \setminus \mathscr{D}_U\big) \cup \{X\}$; *(2)* $\mathscr{D}_U \subseteq \big(\mathscr{P}(X) \setminus \mathscr{F}_U\big) \cup \{X\}$.

Proof To prove (1), note that if $A \subseteq X$ such that $A \notin \big(\mathscr{P}(X) \setminus \mathscr{D}_U\big) \cup \{X\}$, then $A \notin \big(\mathscr{P}(X) \setminus \mathscr{D}_U\big)$ and $A \notin \{X\}$. Therefore, $A \in \mathscr{D}_U$, and thus by assertion (2) of Theorem 22.15 we necessarily have $A \notin \mathscr{F}_U$. Therefore, inclusion (1) is true.

Moreover, by using Theorems 20.9, 22.15, 20.4 and 22.2, we can also prove

Theorem 22.17 *If U is a corelation on X, then for any $A \subseteq X$ we have*

(1) $A \in \mathscr{E}_U$ *if* $\Omega \subseteq A$ *for some* $\Omega \in \mathscr{T}_U \setminus \{\emptyset\}$;
(2) $A \in \mathscr{D}_U$ *only if* $A \setminus W \neq \emptyset$ *for all* $W \in \mathscr{F}_U \setminus \{X\}$.

Proof From Theorem 20.9 we know that $\mathscr{E}_U$ is ascending in $\mathscr{P}(X)$. Therefore, assertion (1) is an immediate consequence of assertion (1) of Theorem 22.15.

To check assertion (2), note that if $A \in \mathscr{D}_U$, then by Theorem 20.4 we have $A^c \notin \mathscr{E}_U$. Therefore, by assertion (1), for any $\Omega \in \mathscr{T}_U \setminus \{\emptyset\}$ we have $\Omega \not\subseteq A^c$, and thus $A \cap \Omega \neq \emptyset$.

Now, if $W \in \mathscr{F}_U \setminus \{X\}$, then by Theorem 22.2 we can see that $W^c \in \mathscr{T}_U \setminus \{\emptyset\}$. Therefore, by our former observation, we can state that $A \cap W^c \neq \emptyset$, and thus $A \setminus W \neq \emptyset$.

23 Open and Closed Sets Derived from Relations

Definition 23.1 For any relation R on X, the members of the families

$$\mathscr{T}_R = \mathscr{T}_{R^{\triangleright}} \qquad \text{and} \qquad \mathscr{F}_R = \mathscr{F}_{R^{\triangleright}}$$

will be called the *open sets* and *closed sets* generated by the relation R, respectively.

Thus, by Definition 19.1, we evidently have the following

Theorem 23.2 *If R is a relation on X, then for any $A \subseteq X$, we have*

(1) $A \in \mathscr{F}_R \iff \operatorname{cl}_R(A) \subseteq A \iff \operatorname{cl}_{R^{\triangleright}}(A) \subseteq A$;
(2) $A \in \mathscr{T}_R \iff A \subseteq \operatorname{int}_R(A) \iff A \subseteq \operatorname{int}_{R^{\triangleright}}(A)$.

Proof To prove (1), note that, by Definitions 23.1 and 19.1, we have

$$A \in \mathscr{F}_R \iff A \in \mathscr{F}_{R^{\triangleright}} \iff \operatorname{cl}_{R^{\triangleright}}(A) \subseteq A \iff \operatorname{cl}_R(A) \subseteq A .$$

Now, by Theorem 19.14 and Definition 22.1, we can also state

Theorem 23.3 *For any corelation U on X, we have*

(1) $\mathscr{T}_U = \mathscr{T}_{U^{\triangleleft}}$; *(2)* $\mathscr{T}_U = \mathscr{T}_{U^{\triangleleft}}$.

Moreover, from the results of Sect. 22, we can immediately derive the following theorems.

Theorem 23.4 *If R is a relation on X, then for any $A \subseteq X$, we have*

(1) $A \in \mathscr{T}_R \iff A^c \in \mathscr{F}_R$; *(2)* $A \in \mathscr{F}_R \iff A^c \in \mathscr{T}_R$.

Corollary 23.5 *For any relation R on X, we have*

(1) $\mathscr{T}_R = \mathscr{F}_R^c$; *(2)* $\mathscr{F}_R = \mathscr{T}_R^c$.

Theorem 23.6 *For any relation R on X, the families $\mathscr{T}_R$ and $\mathscr{F}_R$ are closed under arbitrary unions and intersections.*

Corollary 23.7 *For any relation R on X, we have* $\{\emptyset,\ X\} \subseteq \mathscr{T}_R \cap \mathscr{F}_R$.

Theorem 23.8 *If R is a relation on X, then for any $A \subseteq X$ we have*

(1) $A \in \mathscr{T}_R$ *if and only if* $R\,(x) \subseteq A$ *for all* $x \in A$;
(1) $A \in \mathscr{F}_R$ *if and only if* $A \cap R\,(x) \neq \emptyset$ *implies* $x \in A$ *for all* $x \in X$.

Proof To prove (1), note that, by Definition 23.1 and Theorems 22.8 and 8.8,

$$\begin{aligned} A \in \mathscr{T}_R \iff A \in \mathscr{T}_{R^{\triangleright}} &\iff \forall\ x \in A : \quad R^{\triangleright\triangleleft}(x) \subseteq A \\ &\iff \forall\ x \in A : \quad R\,(x) \subseteq A . \end{aligned}$$

The above assertions can be reformulated in the following more concise forms.

Corollary 23.9 *If R is a relation on X, then for any $A \subseteq X$ we have*

(1) $A \in \mathscr{T}_R \iff R\,[\,A\,] \subseteq A$;

(2) $A \in \mathscr{F}_R \iff R^{-1}[\,A\,] \subseteq A$.

Hence, it is clear that in particular we can also state

Corollary 23.10 *If R is a relation on X, then*
(1) $\mathscr{T}_R = \mathscr{F}_{R^{-1}}$; *(2)* $\mathscr{F}_R = \mathscr{T}_{R^{-1}}$.

Remark 23.11 Note that U is a corelation on X, then in general U^{-1} is not a corelation on X. Therefore, the corresponding theorem for U^{-1} cannot be stated.

However, concerning the relationally generated inverse $U^{\mathbf{-1}}$ of U, we can prove

Theorem 23.12 *For any corelation U on X, we have*

(1) $\mathscr{T}_{U^{\mathbf{-1}}} = \mathscr{F}_{U^{\lhd}} = \mathscr{F}_{U^{\circ}}$; *(2)* $\mathscr{F}_{U^{\mathbf{-1}}} = \mathscr{T}_{U^{\lhd}} = \mathscr{T}_{U^{\circ}}$.

Proof To prove (1), by using Definitions 13.1 and 23.1, Corollary 23.10 and Definition 9.1, we can easily see that

$$\mathscr{T}_{U^{\mathbf{-1}}} = \mathscr{T}_{U^{\lhd -1 \rhd}} = \mathscr{T}_{U^{\lhd -1}} = \mathscr{F}_{U^{\lhd}} = \mathscr{F}_{U^{\lhd \rhd}} = \mathscr{F}_{U^{\circ}}.$$

Now, by using Theorem 22.10, we can also prove the following

Theorem 23.13 *If R is a relation on X, then for any $A \subseteq X$ we have*

(1) $A \in \mathscr{T}_R \iff A \in \operatorname{Int}_R(A) \iff A \in \operatorname{Int}_{R^{\rhd}}(A)$;
(2) $A \in \mathscr{F}_R \iff A^c \notin \operatorname{Cl}_R(A) \iff A^c \notin \operatorname{Cl}_{R^{\rhd}}(A)$.

Proof To prove (1), note that, by Definition 23.1 and Theorem 22.10, for any $A \subseteq X$ we have

$$A \in \mathscr{T}_R \iff A \in \mathscr{T}_{R^{\rhd}} \iff A \in \operatorname{Int}_{R^{\rhd\lhd}}(A) \iff A \in \operatorname{Int}_{R^{\rhd\circ}}(A).$$

Moreover, by Theorems 8.8 and 9.3, we have $R^{\rhd\lhd} = R$ and $R^{\rhd\circ} = R^{\rhd}$. Therefore, the required implications are also true.

From this theorem, by using Definition 16.1, we can also get

Corollary 23.14 *If R is a relation on X, then for any $A \subseteq X$ we have*
(1) $A \in \mathscr{T}_R \iff R^{\rhd}(A) \subseteq A$; *(2)* $A \in \mathscr{F}_R \iff A \cap R^{\rhd}(A) = \emptyset$.

Moreover, by using Theorem 22.12, we can also prove the following

Theorem 23.15 *If R is a relation on X, then for any $A \subseteq X$*

(1) $A \in \operatorname{Int}_R(A)$ *implies* $A \in \mathscr{T}_R$;
(2) $A \in \mathscr{F}_R$ *and* $\operatorname{cl}_R(A) \neq \emptyset$ *imply* $A \in \operatorname{Cl}_R(A)$.

Proof From Theorems 8.11 and Corollary 5.10, we know that the corelation $R^{\rhd}$ is quasi-increasing. Therefore, to prove (2), by Theorem 22.12, we can state that

$$A \in \mathscr{F}_{\mathscr{R}^{\rhd}} \quad \text{and} \quad \operatorname{cl}_{R^{\rhd}}(A) \neq \emptyset \quad \text{implies} \quad A \in \operatorname{Cl}_{R^{\rhd}}(A).$$

Hence, by Definitions 23.1, 19.1 and 17.1, we can see that the required implication is also true.

Remark 23.16 To see the necessity of the condition $\mathrm{cl}_R(A) \neq \emptyset$ in the above theorem, note that if $\mathrm{cl}_R(A) = \emptyset$, then $\mathrm{cl}_R(A) \subseteq A$, and thus $A \in \mathscr{F}_R$ by Theorem 23.2.

Moreover, if $\mathrm{cl}_R(A) = \emptyset$, then by Corollary 19.12, we also have $R^{-1}[\,A\,] = \emptyset$, and thus $A \cap R^{-1}[\,A\,] = \emptyset$. Therefore, $R[\,A\,] \cap A = \emptyset$, and thus $A \notin \mathrm{Cl}_R(A)$ by Theorem 17.11.

Finally, we note that, from Theorems 22.15 and 22.17, by Definitions 23.1 and 21.1, it is clear that we also have the following two theorems.

Theorem 23.17 *For any relation R on X, we have*

(1) $\mathscr{T}_R \setminus \{\emptyset\} \subseteq \mathscr{E}_R$; *(2)* $\mathscr{D}_R \cap \mathscr{F}_R \subseteq \{X\}$.

Corollary 23.18 *For any relation R on X, we have*
(1) $\mathscr{F}_R \subseteq \big(\mathscr{P}(X) \setminus \mathscr{D}_R\big) \cup \{X\}$; *(2)* $\mathscr{D}_R \subseteq \big(\mathscr{P}(X) \setminus \mathscr{F}_R\big) \cup \{X\}$.

Theorem 23.19 *If R is a relation on X, then for any $A \subseteq X$ we have*

(1) $A \in \mathscr{E}_R$ *if* $\Omega \subseteq A$ *for some* $\Omega \in \mathscr{T}_R \setminus \{\emptyset\}$;
(2) $A \in \mathscr{D}_R$ *only if* $A \setminus W \neq \emptyset$ *for all* $W \in \mathscr{F}_R \setminus \{X\}$.

24 Some Further Results on Open and Closed Sets

The origin of the following theorem and the use of the preorder closure in the theory of relator spaces go back to Mala [33]. (See also [22, 34, 44].)

Theorem 24.1 *For any two relations R and S on X, the following assertions are equivalent:*

(1) $R \subseteq S^{\infty}$; *(2)* $R^{\infty} \subseteq S^{\infty}$; *(3)* $\mathscr{T}_S \subseteq \mathscr{T}_R$; *(4)* $\mathscr{F}_S \subseteq \mathscr{F}_R$.

Proof By Theorem 3.13, it is clear that (1) and (2) are equivalent. Moreover, from Theorems 23.4, it is clear that (3) and (4) are also equivalent. Therefore, it is enough to prove only the equivalence of (1) and (3).

For this, note that if $A \in \mathscr{T}_S$, then by Corollary 23.9 we have $S[\,A\,] \subseteq A$. Hence, by induction, we can infer that $S^n[\,A\,] \subseteq A$ for all $n \in \mathbb{N}$. Thus, since $S^0[\,A\,] = \Delta_X[\,A\,] = A$ also holds, by Theorem 4.6 we can also state that

$$S^{\infty}[\,A\,] = \Big(\bigcup_{n=0}^{\infty} S^n\Big)[\,A\,] = \bigcup_{n=0}^{\infty} S^n[\,A\,] \subseteq A\,.$$

Therefore, if (1) holds, then we also have

$$R\,[\,A\,] \subseteq S^{\infty}\,[\,A\,] \subseteq A\,.$$

Hence, by Corollary 23.9, we can see that $A \in \mathscr{T}_R$. Therefore, (1) implies (3).

Moreover, if $x \in X$, then by using Theorems 3.13 and 3.15, we can see that

$$S\left[\,S^{\infty}(x)\right] \subseteq S^{\infty}\left[\,S^{\infty}(x)\right] = \left(S^{\infty} \circ S^{\infty}\right)(x) = S^{\infty}(x).$$

Hence, by using Corollary 23.9, we can infer that $S^{\infty}(x) \in \mathscr{T}_S$. Therefore, if (3) holds, then we also have $S^{\infty}(x) \in \mathscr{T}_R$. Hence, by using $x \in S^{\infty}(x)$ and Corollary 23.9, we can infer that

$$R(x) \subseteq R\left[\,S^{\infty}(x)\,\right] \subseteq S^{\infty}(x).$$

Therefore, by Theorem 3.1, assertion (1) also holds.

Remark 24.2 From the above proof, we can see that if R is a relation on X, then $R^{\infty}(x) \in \mathscr{T}_R$ for all $x \in X$. Thus, R^{∞} is an open-valued preorder relation on X.

Moreover, from the above proof, we can also see that $\mathscr{T}_R \subseteq \mathscr{T}_{R^{\infty}}$. Hence, by using that $R \subseteq R^{\infty}$, and thus $\mathscr{T}_{R^{\infty}} \subseteq \mathscr{T}_R$, we can already infer that $\mathscr{T}_R = \mathscr{T}_{R^{\infty}}$.

However, the latter fact can more easily be derived from the following immediate consequence of the above theorem by using the equality $R^{\infty} = R^{\infty\infty}$.

Corollary 24.3 *For any two relations R and S on X, the following assertions are equivalent:*

(1) $R^{\infty} = S^{\infty}$; *(2)* $\mathscr{T}_R = \mathscr{T}_S$; *(3)* $\mathscr{F}_R = \mathscr{F}_S$.

From Theorem 24.1, by Corollary 3.14, it is clear that in particular we also have

Theorem 24.4 *If R is an arbitrary and S is preorder relation on X, then the following assertions are equivalent:*

(1) $R \subseteq S$; *(2)* $R^{\infty} \subseteq S$; *(3)* $\mathscr{T}_S \subseteq \mathscr{T}_R$; *(4)* $\mathscr{F}_S \subseteq \mathscr{F}_R$.

Thus, by this theorem or Corollary 24.3, we can also state

Corollary 24.5 *If R is an arbitrary and S is preorder relation on X, then the following assertions are equivalent:*

(1) $R^{\infty} = S$; *(2)* $\mathscr{T}_R = \mathscr{T}_S$; *(3)* $\mathscr{F}_R = \mathscr{F}_S$.

Hence, it is clear that, in particular, we also have

Corollary 24.6 *For any two preorder relations R and S on X, each of the equalities $\mathscr{T}_R = \mathscr{T}_S$ and $\mathscr{F}_R = \mathscr{F}_S$ implies that $R = S$.*

Moreover, by using Corollary 24.5, we can also easily prove the following

Theorem 24.7 *For a relation R on X, we have*

(1) $R^{\infty} = \Delta_X \iff \mathscr{T}_R = \mathscr{P}(X) \iff \mathscr{F}_R = \mathscr{P}(X)$;

(2) $R^{\infty} = X^2 \iff \mathscr{T}_R = \{\emptyset,\ X\} \iff \mathscr{F}_R = \{\emptyset,\ X\}$.

Proof To check (1), note that, by Corollary 23.9, we have $\mathscr{T}_{\Delta_X} = \mathscr{P}(X)$ and $\mathscr{F}_{\Delta_X} = \mathscr{P}(X)$. Moreover, Δ_X is a preorder relation on X. Therefore, Corollary 24.5 can be applied to obtain the required equivalences.

Remark 24.8 Note that, for a reflexive relation R on X, the inclusion $R \subseteq \Delta_X$ already implies that $R = \Delta_X$.

While, for an arbitrary relation R on X, the inclusion $X^2 \subseteq R^\infty$ means only that, for any $x,\ y \in X$, with $x \neq y$, there exists a finite sequence $(x_i)_{i=0}^n$ in X such that $x_0 = x$, $x_n = y$ and $x_i \in R\,(x_{i-1})$ for all $i = 1,\ 2,\ \ldots, n$.

In our former papers [31, 44], a relator $\mathscr{R}$ on X was called well-chained if, under the plausible notation $\mathscr{R}^\infty = \{R^\infty :\ R \in \mathscr{R}\}$, we have $\mathscr{R}^\infty = \{X^2\}$. Thus, well-chainedness is a particular case of simplicity [41].

Moreover, it was shown that connectedness a particular case of well-chainedness [44]. In this respect, it noteworthy that compactness is a particular case of total boundedness [60]. While, "convergent" and "Cauchy" are actually equivalent notions [54].

Note that, analogously to the identity relation Δ_X and the universal relation X^2, the Davis–Pervin relation $R_A = A^2 \cup A^c \times X$, where $A \subseteq X$, is also an important preorder relation on X.

Therefore, in addition to Theorem 24.7, it is also worth proving the following

Theorem 24.9 *For any $A \subseteq X$ and relation R on X, the following assertions are equivalent:*

(1) $R^\infty = R_A$; *(2)* $\mathscr{T}_R = \big\{\emptyset,\ A,\ X\big\}$; *(3)* $\mathscr{F}_R = \big\{\emptyset,\ A^c,\ X\big\}$.

Proof For any $\Omega \subseteq X$, we have $R_A[\,\Omega\,] = \emptyset$ if $\Omega = \emptyset$,

$$R_A[\,\Omega\,] = \Omega \quad \text{if} \quad \emptyset \neq \Omega \subseteq A \qquad \text{and} \qquad R_A[\,\Omega\,] = X \quad \text{if} \quad \Omega \not\subseteq A.$$

Hence, by using Corollary 23.9 and Theorem 23.4, we can see that

$$\mathscr{T}_{R_A} = \big\{\emptyset,\ A,\ X\big\}; \qquad \text{and} \qquad \mathscr{F}_{R_A} = \big\{\emptyset,\ A^c,\ X\big\}.$$

Moreover, since R_A is a preorder relation on X, by Corollary 3.14 we can state that $R_A^\infty = R_A$. Therefore, Corollary 24.5 can again be applied to obtain the required equivalences.

Remark 24.10 Note that $R_\emptyset = X^2$ and $R_X = X^2$. Thus, in particular, Theorem 24.9 is a substantial generation of assertion (2) of Theorem 24.7.

Concerning the relation R_A, it also worth noticing that, by Corollary 23.10, we have $\mathscr{T}_{R_{A^c}} = \mathscr{F}_{R_A} = \mathscr{T}_{R_A^{-1}}$. Therefore, by Corollary 24.6, we also have $R_{A^c} = R_A^{-1}$.

The importance of the relations R_A, with $A \subseteq X$, is also apparent from

Theorem 24.11 *For any relation R on X, we have*

$$R^{\infty} = \bigcap_{A \in \mathscr{T}_R} R_A .$$

Proof Define $S = \bigcap_{A \in \mathscr{T}_R} R_A$. Then, for any $x \in X$, we have

$$S(x) = \bigcap_{A \in \mathscr{T}_R} R_A(x) = \bigcap \big\{ A \in \mathscr{T}_R : \quad x \in A \big\}.$$

Hence, by Remark 24.2, we can see that $S(x) \subseteq R^{\infty}(x)$. Therefore, $S \subseteq R^{\infty}$.

On the other hand, if $A \in \mathscr{T}_R$, then by the corresponding definitions, for any $x \in A$, we have $R(x) \subseteq A = R_A(x)$. Hence, it is clear that $R \subseteq R_A$, and thus $R^{\infty} \subseteq R_A^{\infty} = R_A$ for all $A \in \mathscr{T}_A$. Therefore, $R^{\infty} \subseteq S$, and thus also $S = R^{\infty}$.

Remark 24.12 From Theorem 24.1, we can see that the mappings

$$R \longmapsto \mathscr{T}_R \qquad \text{and} \qquad R \longmapsto R^{\infty},$$

where R is a relation on X, establish a Pataki connection with respect to the relation $\subseteq$ and $\supseteq$.

Therefore, it is not surprising that we also have the following

Theorem 24.13 *If R is a relation on X, then $S = R^{\infty}$ is the largest relation on X such that $\mathscr{T}_R \subseteq \mathscr{T}_S$ $\big(\mathscr{T}_R = \mathscr{T}_S \big)$, or equivalently $\mathscr{F}_R \subseteq \mathscr{F}_S$ $\big(\mathscr{F}_R = \mathscr{F}_S \big)$.*

Proof By Theorem 3.15, we have $R^{\infty} = R^{\infty\infty} = S^{\infty}$. Hence, by using Corollary 24.3, we can infer that $\mathscr{T}_R = \mathscr{T}_S$.

Moreover, if Ω is a relation on X such that $\mathscr{T}_R \subseteq \mathscr{T}_{\Omega}$, then by Theorem 24.1 we have $\Omega \subseteq R^{\infty} = S$. Therefore, S has the required properties.

From Theorem 24.1, by using Theorem 23.3, we can also easily derive

Theorem 24.14 *For any two corelations U and V on X to Y, the following assertions are equivalent:*

(1) $U^{\lhd} \subseteq V^{\lhd \infty}$; *(2)* $U^{\lhd \infty} \subseteq V^{\lhd \infty}$; *(3)* $\mathscr{T}_V \subseteq \mathscr{T}_U$; *(4)* $\mathscr{F}_V \subseteq \mathscr{F}_U$.

Proof By Theorem 24.1, we have

$$U^{\lhd} \subseteq V^{\lhd \infty} \iff U^{\lhd \infty} \subseteq V^{\lhd \infty} \iff \mathscr{T}_{V^{\lhd}} \subseteq \mathscr{T}_{U^{\lhd}} \iff \mathscr{F}_{V^{\lhd}} \subseteq \mathscr{F}_{U^{\lhd}}.$$

Moreover, by Theorem 23.3, we have $\mathscr{T}_{U^{\lhd}} = \mathscr{T}_U$ and $\mathscr{F}_{U^{\lhd}} = \mathscr{F}_U$.

Now, as an immediate consequence of this theorem, we can also state

Corollary 24.15 *For any two corelations U and V on X to Y, the following assertions are equivalent:*

(1) $U^{\lhd \infty} = V^{\lhd \infty}$; *(2)* $\mathscr{T}_U = \mathscr{T}_V$; *(3)* $\mathscr{F}_U = \mathscr{F}_V$.

25 Generalizations to Corelators and Relators

Analogously to our former terminology on *relators* and *relator spaces* [63, 65], a family $\mathscr{U}$ of corelations on X to Y will be called a *corelator* on X to Y, and the ordered pair $(X, Y)(\mathscr{U}) = \big((X, Y), \mathscr{U}\big)$ will be called a *corelator space*.

In particular, a corelator $\mathscr{U}$ on X to itself will be called a corelator on X, and the notation $X(\mathscr{U}) = (X, X)(\mathscr{U})$ will be used. Moreover, a corelator space $(X, Y)(\mathscr{U})$ will be called *simple* if the corelator $\mathscr{U}$ is simple in the sense that it is a singleton $\{U\}$. Singleton are usually identified with their elements.

Since corelations on X to Y are more general objects than relations on X to Y, it is clear that corelators on X to Y are also more general objects than relators on X to Y. In particular, instead of a relator $\mathscr{R}$ on X to X we may always naturally consider the associated corelator $\mathscr{R}^{\triangleright} = \{R^{\triangleright} : \ \ R \in \mathscr{R}\}$. Therefore, in the sequel we shall mainly be dealing with corelators.

Now, for a corelator $\mathscr{U}$ on X to Y, we may, for instance, naturally define

$$\operatorname{Int}_{\mathscr{U}} = \bigcup_{U\in\mathscr{U}} \operatorname{Int}_U, \qquad \operatorname{int}_{\mathscr{U}} = \bigcup_{U\in\mathscr{U}} \operatorname{int}_U \qquad \text{and} \qquad \mathscr{E}_{\mathscr{U}} = \bigcup_{U\in\mathscr{U}} \mathscr{E}_U .$$

Namely, thus for any $x \in X$ and $B \subseteq Y$, we also have

(1) $x \in \operatorname{int}_{\mathscr{U}}(B) \iff \{x\} \in \operatorname{Int}_{\mathscr{U}}(B)$;

(2) $B \in \mathscr{E}_{\mathscr{U}} \iff \operatorname{int}_{\mathscr{U}}(B) \neq \emptyset$.

However, in connection with the generated open sets we must be more careful. Namely, if $\mathscr{U}$ is a corelator on X, then the family

$$\tau_{\mathscr{U}} = \bigcup_{U\in U} \mathscr{T}_U$$

of all *proximally open sets* generated by $\mathscr{U}$ is, in general, only a proper subfamily of the family

$$\mathscr{T}_{\mathscr{U}} = \big\{A \subseteq X : \ \ A \subseteq \operatorname{int}_{\mathscr{U}}(A)\big\}$$

of all *topologically open sets* generated by $\mathscr{U}$.

To see this, note that if for instance $x_1, x_2 \in X$ such that $x_1 \neq x_2$, and for each $i = 1, 2$ we define

$$R_i = \{x_i\}^2 \cup \big(X \setminus \{x_i\}\big)^2,$$

then $\mathscr{R} = \{R_1, R_2\}$ is an equivalence relator on X such that, under the notation $A = \{x_1, x_2\}$, we have $A \in \mathscr{T}_{\mathscr{R}} \setminus \tau_{\mathscr{R}}$ whenever $X \neq A$.

The above mentioned property of the topologically open sets causes the serious inconvenience that if $\mathscr{U}$ is a corelator on X, then in general there does not exist a largest corelator $\mathscr{U}^{\square}$ on X such that $\mathscr{T}_{\mathscr{U}} = \mathscr{T}_{\mathscr{U}^{\square}}$. The latter fact for relators was first proved by Mala [33, Example 5.3] by considering the singleton relator

$\mathscr{R} = \{X^2\}$ with $\operatorname{card}(X) > 2$. This simple relator was later more deeply investigated by Pataki [42, Example 7.2]. (See also [67, Example 10.11].)

The notion of a fat set, the duality of fat and dense sets, and the fact that the fat and dense sets in a relator space are frequently more important tools than the open and closed ones was first revealed by the present author in [52, 53]. The participants of the symposium, considering me as an outsider, were unwilling to acknowledge this.

Despite that, to realized the validity of my claims, it is enough to note only that if for instance $\leq$ is a certain order relation on X, then $\mathscr{T}_{\leq}$ and $\mathscr{E}_{\leq}$ are just the families of all ascending subsets and residual subsets of the ordered set $X(\leq)$, respectively. And, to note that the residual subsets are more important tools than the ascending ones. Namely, they can be used to define coherences and convergences of nets [49, 50].

Moreover, if for instance R is a relation on $\mathbb{R}$ such that

$$R(x) = \{x-1\} \cup [\,x,\ +\infty\,[$$

for all $x \in \mathbb{R}$, then $\mathscr{T}_R = \mathscr{T}_{R^{\triangleright}} = \{\emptyset,\ \mathbb{R}\}$, but $\mathscr{E}_R = \mathscr{E}_{R^{\triangleright}}$ is quite a large family of subsets of $\mathbb{R}$. Namely, it contains all supersets of the sets $R(x)$ with $x \in X$. Therefore, in this particular case, $\mathscr{E}_R$ is a much better tool than $\mathscr{T}_R$.

To let the reader feel the appropriateness of the term "fat", it is also worth mentioning that a subset A of a corelator space $X(\mathscr{U})$ may be called *rare* if $\operatorname{cl}_{\mathscr{U}}(A) \notin \mathscr{E}_{\mathscr{U}}$. Moreover, the subset A may be called *meager* if there exists a sequence $(A_n)_{n=1}^{\infty}$ of rare subsets of $X(\mathscr{U})$ such that $A = \bigcup_{n=1}^{\infty} A_n$. Thus, the corelator space $X(\mathscr{U})$ may be called *Baire* if the fat subsets of $X(\mathscr{U})$ are not meager, or equivalently the meager subsets of $X(\mathscr{U})$ are not fat. (The corresponding subject for relator spaces has been worked out in [64, 66].)

In this respect, it is also worth mentioning that if F and G are relations on a corelator space $X(\mathscr{U})$ to Y and Z, respectively, and $\mathscr{V}$ is a corelator on Y to Z, then we may also naturally write

(1) $F \in \underline{\operatorname{Lim}}_{\mathscr{V}}(G)$ if $\big\{x \in X:\ G(x) \subseteq V\big(F(x)\big)\big\} \in \mathscr{E}_{\mathscr{U}}$ for all $V \in \mathscr{V}$;

(2) $F \in \overline{\operatorname{Adh}}_{\mathscr{V}}(G)$ if $\big\{x \in X:\ G(x) \cap V\big(F(x)\big) \neq \emptyset\big\} \in \mathscr{D}_{\mathscr{U}}$ for all $V \in \mathscr{V}$.

Now, by taking $F_B = X \times B$ for some $B \subseteq Y$, we may also naturally write:

(3) $B \in \underline{\operatorname{lim}}_{\mathscr{V}}(G)$ if $F_B \in \underline{\operatorname{Lim}}_{\mathscr{V}}(G)$, (4) $B \in \overline{\operatorname{adh}}_V(G)$ if $F_B \in \overline{\operatorname{Adh}}_V(G)$.

However, to have some sufficiently powerful tools in corelator spaces, it is usually enough to consider only the particular case when $X(\mathscr{U})$ is a preordered set, F and G are functions and B is a singleton. That is, it is usually enough to consider the convergence and adherence of preordered nets to preordered nets and points.

Finally, we note several useful algebraic tools can also defined in corelator spaces. First of all, the results of our former paper [65] on upper and lower bound relations generated by relators should also be extended to corelators.

Acknowledgements The work of the author has been supported by the Hungarian Scientific Research Fund (OTKA) Grant K-111651.

The author is indebted to Zsolt Páles for calling his attention to the LaTeX command \pmb (pure man's bold) whose repeated applications could have given an even more visible bold effect in our notations.

References

1. G. Birkhoff, *Lattice Theory*. American Mathematical Society Colloquium Publications, vol. 25 (American Mathematical Society, Providence, RI, 1967)
2. T.S. Blyth, M.F. Janowitz, *Residuation Theory* (Pergamon Press, Oxford, 1972)
3. Z. Boros, Á. Száz, Infimum and supremum completeness properties of ordered sets without axioms. An. St. Univ. Ovid. Constanta **16**, 1–7 (2008)
4. N. Bourbaki, *General Topology*, Chaps. 1–4 (Springer, Berlin, 1989)
5. S. Buglyó, Á. Száz, A more important Galois connection between distance functions and inequality relations. Sci. Ser. A Math. Sci. (N.S.) **18**, 17–38 (2009)
6. D. Bushaw, *Bibliography on Uniform Topology and Its Supplements I–VI* (Washington State University, Washington, 1965–1971)
7. S.C. Carlson, C. Votaw, Para-uniformities, para-proximities, and H-Closed extensions. Rocky Mt. J. Math. **16**, 805–835 (1986)
8. E. Čech, *Topological Spaces* (Academia, Prague, 1966)
9. Á. Császár, *Foundations of General Topology* (Pergamon Press, London, 1963)
10. Á. Császár, *General Topology* (Akadémiai Kiadó, Budapest, 1978)
11. D.W. Curtis, J.C. Mathews, Generalized uniformities for pairs of spaces, in *Topology Conference* (Arizona State University, Tempe, 1968), pp. 212–246
12. B.A. Davey, H.A. Priestley, *Introduction to Lattices and Order* (Cambridge University Press, Cambridge, 2002)
13. A.S. Davis, Indexed systems of neighbordoods for general topological spaces. Am. Math. Mon. **68**, 886–893 (1961)
14. K. Denecke, M. Erné, S.L. Wismath (eds.) *Galois Connections and Applications* (Kluwer Academic, Dordrecht, 2004)
15. D. Doičinov, A unified theory of topological spaces, proximity spaces and uniform spaces. Dokl. Acad. Nauk SSSR **156**, 21–24 (1964) (Russian)
16. V.A. Efremovič, The geometry of proximity. Mat. Sb. **31**, 189–200 (1952) (Russian)
17. V.A. Efremović, A.S. Švarc, A new definition of uniform spaces. Metrization of proximity spaces. Dokl. Acad. Nauk. SSSR **89**, 393–396 (1953) (Russian)
18. R. Engelking, *General Topology* (Polish Scientific Publishers, Warszawa, 1977)
19. P. Fletcher, W.F. Lindgren, *Quasi-Uniform Spaces* (Marcel Dekker, New York, 1982)
20. B. Ganter, R. Wille, *Formal Concept Analysis* (Springer, Berlin, 1999)
21. G. Gierz, K.H. Hofmann, K. Keimel, J.D. Lawson, M. Mislove, D.S. Scott, *A Compendium of Continuous Lattices* (Springer, Berlin, 1980)
22. T. Glavosits, Generated preorders and equivalences. Acta Acad. Paed. Agrienses, Sect. Math. **29**, 95–103 (2002)
23. V. Gregori, J. Ferrer, Quasi-metrization and completion for Pervin's quasi-uniformity. Stohastica **6**, 151–156 (1982)
24. H. Herlich, Topological structeres. Math. Centre Tracts **52**, 59–122 (1974)
25. U. Höhle, T. Kubiak, On regularity of sup-preserving maps: generalizing Zareckiĭ's theorem. Semigroup Forum **83**, 313–319 (2011)
26. W. Hunsaker, W. Lindgren, Construction of quasi-uniformities. Math. Ann. **188**, 39–42 (1970)
27. J.R. Isbell, *Uniform Spaces*. Mathematical Surveys, vol. 12 (American Mathematical Society, Providence, 1964)

28. J.L. Kelley, *General Topology* (Van Nostrand Reinhold Company, New York, 1955)
29. H.-J. Kowalsky, *Topologische Räumen* (Birkhäuser, Basel, 1960)
30. K. Kuratowski, *Introduction to Set Theory and Topology* (Pergamon Press, New York, 1972)
31. J. Kurdics, Á. Száz, Well-chained relator spaces. Kyungpook Math. J. **32**, 263–271 (1992)
32. N. Levine, On Pervin's quasi uniformity. Math. J. Okayama Univ. **14**, 97–102 (1970)
33. J. Mala, Relators generating the same generalized topology. Acta Math. Hungar. **60**, 291–297 (1992)
34. J. Mala, Á. Száz, Modifications of relators. Acta Math. Hungar. **77**, 69–81 (1997)
35. Z.P. Mamuzič, *Introduction to General Topology* (Noordhoff, Groningen, 1963)
36. M.G. Murdeshwar, S.A. Naimpally, *Quasi-Uniform Topological Spaces* (Noordhoff, Groningen, 1966)
37. F. Mynard, E. Pearl (eds.), *Beyond Topology*. Contemporary Mathematics, vol. 486 (American Mathematical Society, Providence, 2009)
38. S.A. Naimpally, B.D. Warrack, *Proximity Spaces* (Cambridge University Press, Cambridge, 1970)
39. H. Nakano, K. Nakano, Connector theory. Pac. J. Math. **56**, 195–213 (1975)
40. J.F. Nash, Jr., M.Th. Rassias (eds.), *Open Problems in Mathematics* (Springer, New York, 2016)
41. G. Pataki, Supplementary notes to the theory of simple relators. Radovi Mat. **9**, 101–118 (1999)
42. G. Pataki, On the extensions, refinements and modifications of relators. Math. Balk. **15**, 155–186 (2001)
43. G. Pataki, On a generalized infimal convolution of set functions. Ann. Math. Sil. **27**, 99–106 (2013)
44. G. Pataki, Á. Száz, A unified treatment of well-chainedness and connectedness properties. Acta Math. Acad. Paedagog. Nyházi. (N.S.) **19**, 101–165 (2003)
45. W.J. Pervin, Quasi-uniformization of topological spaces. Math. Ann. **147**, 316–317 (1962)
46. R. Pöschel, M. Rössinger, A general Galois theory for cofunctions and corelations. Algebra Univers. **43**, 331–345 (2000)
47. W. Sierpinski, *General Topology*. Mathematical Expositions, vol. 7 (University of Toronto Press, Toronto, 1956)
48. Yu. M. Smirnov, On proximity spaces. Math. Sb. **31**, 543–574 (1952) (Russian)
49. Á. Száz, Coherences instead of convergences, in *Proceedings of Conference in Convergence and Generalized Functions (Katowice, 1983)* (Institute of Mathematics of the Polish Academy of Sciences, Warsaw, 1984), pp. 141–148
50. Á. Száz, Basic tools and mild continuities in relator spaces. Acta Math. Hungar. **50**, 177–201 (1987)
51. Á. Száz, Lebesgue relators. Monatsh. Math. **110**, 315–319 (1990)
52. Á. Száz, The fat and dense sets are more important than the open and closed ones, in *Abstracts of the Seventh Prague Topological Symposium* (Inst. Math. Czechoslovak Acad. Sci., Prague, 1991), p. 106
53. Á. Száz, Structures derivable from relators. Singularité **3**, 14–30 (1992)
54. Á. Száz, Cauchy nets and completeness in relator spaces. Colloq. Math. Soc. János Bolyai **55**, 479–489 (1993)
55. Á. Száz, Refinements of relators. Technical Report, Inst. Math., Univ. Debrecen, 1993/76, 19 pp. (1993)
56. Á. Száz, Neighbourhood relators. Bolyai Soc. Math. Stud. **4**, 449–465 (1995)
57. Á. Száz, Relations refining and dividing each other. Pure Math. Appl. Ser. B **6**, 385–394 (1995)
58. Á. Száz, Connectednesses of refined relators. Technical Report, Inst. Math., Univ. Debrecen, 1996/14, 6 pp.
59. Á. Száz, Topological characterizations of relational properties. Grazer Math. Ber. **327**, 37–52 (1996)
60. Á. Száz, Uniformly, proximally and topologically compact relators. Math. Pannon. **8**, 103–116 (1997)
61. Á. Száz, An extension of Kelley's closed relation theorem to relator spaces. Filomat **14**, 49–71 (2000)

62. Á. Száz, A Galois connection between distance functions and inequality relations. Math. Bohem. **127**, 437–448 (2002)
63. Á. Száz, Somewhat continuity in a unified framework for continuities of relations. Tatra Mt. Math. Publ. **24**, 41–56 (2002)
64. Á. Száz, An extension of Baire's category theorem to relator spaces. Math. Morav. **7**, 73–89 (2003)
65. Á. Száz, Upper and lower bounds in relator spaces. Serdica Math. J. **29**, 239–270 (2003)
66. Á. Száz, Rare and meager sets in relator spaces. Tatra Mt. Math. Publ. **28**, 75–95 (2004)
67. Á. Száz, Galois-type connections on power sets and their applications to relators. Technical Report, Inst. Math., Univ. Debrecen, 2005/2, 38 pp. (2005)
68. Á. Száz, Supremum properties of Galois-type connections. Comment. Math. Univ. Carolin. **47**, 569–583 (2006)
69. Á. Száz, Minimal structures, generalized topologies, and ascending systems should not be studied without generalized uniformities. Filomat **21**, 87–97 (2007)
70. Á. Száz, Applications of relations and relators in the extensions of stability theorems for homogeneous and additive functions. Aust. J. Math. Anal. Appl. **6**, 1–66 (2009)
71. Á. Száz, Galois type connections and closure operations on preordered sets. Acta Math. Univ. Comen. **78**, 1–21 (2009)
72. Á. Száz, Foundations of the theory of vector relators. Adv. Stud. Contemp. Math. **20**, 139–195 (2010)
73. Á. Száz, Lower semicontinuity properties of relations in relator spaces. Adv. Stud. Contemp. Math. (Kyungshang) **23**, 107–158 (2013)
74. Á. Száz, A particular Galois connection between relations and set functions. Acta Univ. Sapientiae Math. **6**, 73–91 (2014)
75. Á. Száz, Galois and Pataki connections on generalized ordered sets. Technical Report, Inst. Math., Univ. Debrecen, 2014/3, 27 pp. (2014)
76. Á. Száz, Generalizations of Galois and Pataki connections to relator spaces. J. Int. Math. Virtual Inst. **4**, 43–75 (2014)
77. Á. Száz, A unifying framework for studying continuity, increasingness, and Galois connections. MathLab J. **1**, 154–173 (2018)
78. Á. Száz, The closure-interior Galois connection and its applications to relational inclusions and equations. J. Int. Math. Virt. Inst. **8**, 181–224 (2018)
79. Á. Száz, Basic tools, increasing functions, and closure operations in generalized ordered sets, in *Contributions in Mathematics and Engineering: In Honor of Constantion Caratheodory*, ed. by P.M. Pardalos, Th.M. Rassias (Springer, Cham, 2016), pp. 551–616
80. Á. Száz, A natural Galois connection between generalized norms and metrics. Acta Univ. Sapientiae Math. **9**, 360–373 (2017)
81. Á. Száz, An answer to the question "What is the essential difference between Algebra and Topology?" of Shukur Al-aeashi. Technical Report, Inst. Math., Univ. Debrecen 2017/2, 6 pp. (2017)
82. Á. Száz, Four general continuity properties, for pairs of functions, relations and relators, whose particular cases could be investigated by hundreds of mathematicians. Technical Report, Inst. Math., Univ. Debrecen, 2017/1, 17 pp. (2017)
83. Á. Száz, A. Zakaria, Mild continuity properties of relations and relators in relator spaces, in *Essays in Mathematics and its Applications: In Honor of Vladimir Arnold*, ed. by P.M. Pardalos, Th.M. Rassias (Springer, Cham, 2016), pp. 439–511
84. W.J. Thron, *Topological Structures* (Holt, Rinehart and Winston, New York, 1966)
85. W.J. Thron, Proximity structures and grills. Math. Ann. **206**, 35–62 (1973)
86. H. Tietze, Beiträge zur allgemeinen Topologie I. Axiome für verschiedene Fassungen des Umgebungsbegriffs. Math. Ann. **88**, 290–312 (1923)
87. A. Weil, Sur les espaces á structure uniforme et sur la topologie générale. Actual. Sci. Ind. vol. 551 (Herman and Cie, Paris, 1937)

Rational Contractions and Coupled Fixed Points

Mihai Turinici

1 Introduction

Let X be a nonempty set. By a *sequence* in X, we mean any mapping $x : N \to X$, where $N = \{0, 1, \ldots\}$ is the set of *natural* numbers. For simplicity reasons, we will denote it as $(x(n); n \geq 0)$, or $(x_n; n \geq 0)$; moreover, when no confusion can arise, one further simplifies this notation as $(x(n))$ or (x_n), respectively. Also, any sequence $(y_n := x_{i(n)}; n \geq 0)$ with

$(i(n); n \geq 0)$ is *strictly ascending* (whence: $i(n) \to \infty$ as $n \to \infty$)

will be referred to as a *subsequence* of $(x_n; n \geq 0)$. Finally, call the subset Y of X, *almost singleton* (in short: *asingleton*) provided $y_1, y_2 \in Y$ implies $y_1 = y_2$; and *singleton* if, in addition, Y is nonempty; note that in this case, we have the representation $Y = \{y\}$, for some $y \in X$.

Take a metric $d : X \times X \to R_+ := [0, \infty[$ over X; as well as a selfmap $T \in \mathscr{F}(X)$. [Here, for each couple A, B of nonempty sets, $\mathscr{F}(A, B)$ stands for the class of all functions from A to B; when $A = B$, we write $\mathscr{F}(A)$ in place of $\mathscr{F}(A, A)$.] Denote $\text{Fix}(T) = \{x \in X; x = Tx\}$; each point of this set is referred to as *fixed* under T. The determination of such points is to be performed in the context below, comparable with the one in Rus [35, Ch 2, Sect 2.2]:

pic-0) We say that T is *fix-asingleton*, when $\text{Fix}(T)$ is asingleton; likewise, we say that T is *fix-singleton* when $\text{Fix}(T)$ is singleton

pic-1) We say that $x \in X$ is a *Picard point* (modulo $(d; T)$) if the iterative sequence $(T^n x; n \geq 0)$ is d-Cauchy; when this property holds for all $x \in X$, then T is called a *Picard operator* (modulo d)

M. Turinici (✉)
"A. Myller" Mathematical Seminar, "A. I. Cuza" University, Iaşi, Romania
e-mail: mturi@uaic.ro

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization
and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_26

pic-2) We say that $x \in X$ is a *strong Picard point* (modulo $(d; T)$) if the iterative sequence $(T^n x; n \geq 0)$ is d-convergent and $\lim_n(T^n x) \in \mathrm{Fix}(T)$; when this property holds for all $x \in X$, then T is called a *strong Picard operator* (modulo d).

The basic result in this area was obtained in 1922 by Banach [2]. Call $T : X \to X$, (d, μ)-*contractive* (where $\mu \geq 0$), provided

(B-con) $d(Tx, Ty) \leq \mu d(x, y), \forall x, y \in X.$

Theorem 1 *Assume that T is (d, μ)-contractive, for some $\mu \in [0, 1[$. In addition, let X be d-complete. Then,*

(11-a) T is fix-singleton: $\mathrm{Fix}(T) = \{z\}$*, for some $x \in X$*
(11-b) T is strong Picard (modulo d): $\lim_n T^n x = z$*, for each $x \in X$.*

This result (referred to as: *Banach's contraction principle*) found some basic applications to the operator equations theory. As a consequence, many extensions for it were proposed. The most general ones have the *implicit* form

(si-con) $(d(Tx, Ty), d(x, y), d(x, Tx), d(y, Ty), d(x, Ty), d(Tx, y)) \in \mathscr{M}$,
for all $x, y \in X, x\nabla y$;

where $\mathscr{M} \subseteq R^6_+$ is a (nonempty) subset, and ∇ is a *relation* over X. In particular, when $\mathscr{M}$ is the zero-section of a certain function $F : R^6_+ \to R$, the implicit contractive condition above has the familiar form:

(fi-con) $F(d(Tx, Ty), d(x, y), d(x, Tx), d(y, Ty), d(x, Ty), d(Tx, y)) \leq 0$,
for all $x, y \in X, x\nabla y$.

For the explicit trivial relation case of it, characterized as

(fe-con) $d(Tx, Ty) \leq G(d(x, y), d(x, Tx), d(y, Ty), d(x, Ty), d(Tx, y))$,
for all $x, y \in X$

(where $G : R^5_+ \to R_+$ is a function), some consistent lists of such contractions may be found in the survey papers by Rhoades [33] or Collaco and E Silva [11], as well as the references therein; these, in particular, include some outstanding results in the area due to Boyd and Wong [5], Reich [32], and Matkowski [21]. And, for the implicit setting above, certain technical aspects have been considered by Leader [20] and Turinici [43]. On the other hand, in the case of ∇ being a *(partial) order* on X, some early statements were obtained in the 1986 papers by Turinici [44, 45]; two decades later, these results have been re-discovered—at the level of Banach contractive maps—by Ran and Reurings [31]; see also Nieto and Rodriguez-Lopez [29]. Further, an extension—to the same framework—of Leader's contribution was performed in Agarwal et al [1]; and, since then, the number of such papers increased rapidly. Finally, the case of ∇ being *amorphous* (i.e.: it has no regularity properties at all) has been discussed (via graph techniques) in Jachymski [18]; and (from a general perspective) by Samet and Turinici [38].

A basic particular case of the implicit contractive property above is

(2i-con) $(d(Tx, Ty), d(x, y)) \in \mathscr{M}$, for all $x, y \in X$, $x\nabla y$;

where $\mathscr{M} \subseteq R_+^2$ is a (nonempty) subset. The classical example over this direction (again in the trivial relation setting) is due to Meir and Keeler [23]; further refinements of the method were proposed by Matkowski [22] and Ćirić [9]. Having these precise, it is our aim in the following to propose an *analytic* perspective for the study of Meir-Keeler contractions over quasi-ordered metric spaces; this, in particular, includes the old (metrical) contractions due to Boyd and Wong [5] or Matkowski [21], as well as the recent ones introduced by Dutta and Choudhury [12]. Finally, as an application of the obtained facts, a rational type coupled fixed point theorem over quasi-ordered metric spaces is established, which includes a recent statement obtained (via rather different methods) by Nashine and Kadelburg [27]. Further aspects will be delineated elsewhere.

2 Dependent Choice Principle

Throughout this exposition, the axiomatic system in use is Zermelo-Fraenkel's (abbreviated: ZF), as described by Cohen [10, Ch 2]. The notations and basic facts to be considered in this system are more or less standard. Some important ones are discussed below.

(A) Let X be a nonempty set. By a *relation* over X, we mean any nonempty part $\mathscr{R} \subseteq X \times X$; for simplicity, we sometimes write $(x, y) \in \mathscr{R}$ as $x\mathscr{R}y$. Note that $\mathscr{R}$ may be regarded as a mapping between X and $\exp[X]$ (=the class of all subsets in X). To verify this, denote for $x \in X$:

$X(x, \mathscr{R}) = \{y \in X; x\mathscr{R}y\}$ (the *section* of $\mathscr{R}$ through x);

then, the desired mapping representation is $[\mathscr{R}(x) = X(x, \mathscr{R}), x \in X]$. A basic example of such object is

$\mathscr{I} = \{(x, x); x \in X\}$ [the *identical relation* over X].

Given the relations $\mathscr{R}$, $\mathscr{S}$ over X, define their *product* $\mathscr{R} \circ \mathscr{S}$ as

$(x, z) \in \mathscr{R} \circ \mathscr{S}$, if there exists $y \in X$ with $(x, y) \in \mathscr{R}$, $(y, z) \in \mathscr{S}$.

Also, for each relation $\mathscr{R}$ in X, denote

$\mathscr{R}^{-1} = \{(x, y) \in X \times X; (y, x) \in \mathscr{R}\}$ (the *inverse* of $\mathscr{R}$).

Finally, given the relations $\mathscr{R}$ and $\mathscr{S}$ on X, let us say that $\mathscr{R}$ is *coarser* than $\mathscr{S}$ (or, equivalently: $\mathscr{S}$ is *finer* than $\mathscr{R}$), provided

$\mathscr{R} \subseteq \mathscr{S}$; i.e.: $x\mathscr{R}y$ implies $x\mathscr{S}y$.

Given a relation $\mathscr{R}$ on X, the following properties are to be discussed here:

(P1) $\mathscr{R}$ is *reflexive*: $\mathscr{I} \subseteq \mathscr{R}$
(P2) $\mathscr{R}$ is *irreflexive*: $\mathscr{R} \cap \mathscr{I} = \emptyset$
(P3) $\mathscr{R}$ is *transitive*: $\mathscr{R} \circ \mathscr{R} \subseteq \mathscr{R}$
(P4) $\mathscr{R}$ is *symmetric*: $\mathscr{R}^{-1} = \mathscr{R}$
(P5) $\mathscr{R}$ is *antisymmetric*: $\mathscr{R}^{-1} \cap \mathscr{R} \subseteq \mathscr{I}$.

This yields the classes of relations to be used; the following ones are important for our developments:

(C0) $\mathscr{R}$ is *amorphous* (i.e.: it has no specific properties)
(C1) $\mathscr{R}$ is a *quasi-order* (reflexive and transitive)
(C2) $\mathscr{R}$ is a *strict order* (irreflexive and transitive)
(C3) $\mathscr{R}$ is an *equivalence* (reflexive, transitive, symmetric)
(C4) $\mathscr{R}$ is a *(partial) order* (reflexive, transitive, antisymmetric)
(C5) $\mathscr{R}$ is *trivial* (i.e.: $\mathscr{R} = X \times X$).

Remember that, by a *sequence* in X we mean any map $x : N \to X$; also denoted as $(x(n))$ or (x_n). Take such an object; as well as a relation $\mathscr{R}$ on X.

I) Let us say that (x_n) is *$\mathscr{R}$-ascending*, if

$x_n \mathscr{R} x_{n+1}$, for all $n \geq 0$.

Note that this property is not hereditary; i.e.: it cannot hold for a subsequence. However, when $\mathscr{R}$ is transitive, this ascending property may be written as

$x_n \mathscr{R} x_m$, whenever $n < m$;

wherefrom, it is hereditary.

II) Let us say that (x_n) is *bounded above* by an element $u \in X$, when

$x_n \mathscr{R} u$, for all n; written as: $(x_n)\mathscr{R}u$.

Clearly, this property is hereditary:

$(x_n)\mathscr{R}u$ implies $(y_n)\mathscr{R}u$, for each subsequence (y_n) of (x_n).

The converse inclusion is not in general true; i.e.,

$((y_n)\mathscr{R}u$ for some subsequence (y_n) of $(x_n))$ does not imply $(x_n)\mathscr{R}u$.

(B) Remember that, an outstanding part of (ZF) is the *Axiom of Choice* (abbreviated: AC); which, in a convenient manner, may be written as

(AC) For each couple (J, X) of nonempty sets and each function
$F : J \to \exp(X)$, there exists a (selective) function
$f : J \to X$, with $f(\nu) \in F(\nu)$, for each $\nu \in J$.

(Here, $\exp(X)$ stands for the class of all nonempty elements in $\exp[X]$). Sometimes, when the ambient set X is endowed with denumerable type structures, the case of

$J = N$ will suffice for handling choice reasonings; and, existence of such a selective function may be determined by using a weaker form of (AC), called: *Dependent Choice* principle (in short: DC). Call the relation $\mathscr{R}$, *proper* when

$(X(x, \mathscr{R}) =)\mathscr{R}(x)$ is nonempty, for each $x \in X$.

Note that, in this case, $\mathscr{R}$ is to be viewed as a mapping between X and $\exp(X)$; the couple $(X, \mathscr{R})$ will be then referred to as a *proper relational structure*. Given $a \in X$, let us say that the sequence $(x_n; n \geq 0)$ in X is $(a; \mathscr{R})$-*iterative*, provided

$x_0 = a$ and $x_n \mathscr{R} x_{n+1}$ (i.e.: $x_{n+1} \in \mathscr{R}(x_n)$), $\forall n$.

Proposition 1 *Let the relational structure $(X, \mathscr{R})$ be proper. Then, for each $a \in X$ there is at least an $(a, \mathscr{R})$-iterative sequence in X.*

This principle—proposed, independently, by Bernays [3] and Tarski [42]—is deductible from (AC), but not conversely; cf. Wolk [48]. Moreover, by the developments in Moskhovakis [25, Ch 8] and Schechter [41, Ch 6], the *reduced* system (ZF-AC+DC) it comprehensive enough so as to cover the "usual" mathematics; see also Moore [24, Appendix 2].

Let $(\mathscr{R}_n; n \geq 0)$ be a sequence of relations on X. Given $a \in X$, let us say that the sequence $(x_n; n \geq 0)$ in X is $(a; (\mathscr{R}_n; n \geq 0))$-*iterative*, provided

$x_0 = a$ and $x_n \mathscr{R}_n x_{n+1}$ (i.e.: $x_{n+1} \in \mathscr{R}_n(x_n)$), $\forall n$.

The following *Diagonal Dependent Choice* principle (in short: DDC) is available.

Proposition 2 *Let $(\mathscr{R}_n; n \geq 0)$ be a sequence of proper relations on X. Then, for each $a \in X$ there exists at least one $(a; (\mathscr{R}_n; n \geq 0))$-iterative sequence in X.*

Clearly, (DDC) includes (DC); to which it reduces when $(\mathscr{R}_n; n \geq 0)$ is constant. The reciprocal of this is also true. In fact, letting the premises of (DDC) hold, put $P = N \times X$; and let $\mathscr{S}$ be the relation over P introduced as

$\mathscr{S}(i, x) = \{i + 1\} \times \mathscr{R}_i(x), \ (i, x) \in P.$

It will suffice applying (DC) to $(P, \mathscr{S})$ and $b := (0, a) \in P$ to get the conclusion in our statement; we do not give details.

Summing up, (DDC) is provable in (ZF-AC+DC). This is valid as well for its variant, referred to as: *Selected Dependent Choice* principle (in short: SDC).

Proposition 3 *Let the map $F : N \to \exp(X)$ and the relation $\mathscr{R}$ over X fulfill*

($\forall n \in N$): $\mathscr{R}(x) \cap F(n + 1) \neq \emptyset, \ \forall x \in F(n)$.

Then, for each $a \in F(0)$ there exists a sequence $(x(n); n \geq 0)$ in X, with

$$x(0) = a, \ x(n) \in F(n), \ x(n + 1) \in \mathscr{R}(x(n)), \ \ \forall n.$$

As before, (SDC) $\Longrightarrow$ (DC) ($\Longleftrightarrow$ (DDC)); just take $[F(n) = X, n \in N]$. But, the reciprocal is also true, in the sense: (DDC) $\Longrightarrow$ (SDC). This follows from

Proof (Proposition 3) Let the premises of (SDC) be true. Define a sequence of relations $(\mathscr{R}_n; n \geq 0)$ over X as: for each $n \geq 0$,

$\mathscr{R}_n(x) = \mathscr{R}(x) \cap F(n+1)$, if $x \in F(n)$,
$\mathscr{R}_n(x) = \{x\}$, otherwise $(x \in X \setminus F(n))$.

Clearly, $\mathscr{R}_n$ is proper, for all $n \geq 0$. So, by (DDC), it follows that for the starting $a \in F(0)$, there exists an $(a, (\mathscr{R}_n; n \geq 0))$-iterative sequence $(x(n); n \geq 0)$ in X. Combining with the very definition above, it follows that conclusion in the statement is effectively holding.

In particular, when $\mathscr{R} = X \times X$, the regularity condition imposed in (SDC) holds. The corresponding variant of our underlying statement is just (AC(N)) (=the *Denumerable Axiom of Choice*). Precisely, we have

Proposition 4 *Let $F : N \to \exp(X)$ be a function. Then, for each $a \in F(0)$ there exists a function $f : N \to X$ with $f(0) = a$ and $f(n) \in F(n)$, $\forall n \in N$.*

As a consequence of the discussed facts, (DC) $\Longrightarrow$ (AC(N)) in (ZF-AC). A direct verification of this is obtainable by taking $A = N \times X$ and introducing the relation $\mathscr{R}$ over it, according to:

$\mathscr{R}(n, x) = \{n+1\} \times F(n+1), \ n \in N, x \in X;$

we do not give details. The reciprocal of this last inclusion is not true; see Moskhovakis [25, Ch 8, Sect 8.25] for details.

3 Conv-Cauchy Structures

Let X be a nonempty set.

(A) Denote by $\mathscr{S}(X)$ the class of all sequences (x_n) in X. By a (sequential) *convergence structure* on X we mean, as in Kasahara [19], any part $\mathscr{C}$ of $\mathscr{S}(X) \times X$ with the properties

(conv-1) $\mathscr{C}$ is hereditary:
$((x_n); x) \in \mathscr{C} \Longrightarrow ((y_n); x) \in \mathscr{C}$, for each subsequence (y_n) of (x_n)
(conv-2) $\mathscr{C}$ is *reflexive*:
$(\forall u \in X)$: the constant sequence $(x_n = u; n \geq 0)$ fulfills $((x_n); u) \in \mathscr{C}$;

in this case, the couple $(X, \mathscr{C})$ will be referred to as a *convergence space*. For simplicity, the relation $((x_n); x) \in \mathscr{C}$ will be denoted $x_n \xrightarrow{\mathscr{C}} x$; and reads: x is the *$\mathscr{C}$-limit* of (x_n); the set of all such points will be denoted as

$\mathscr{C} - \lim_n(x_n)$; or, $\lim_n(x_n)$ when $\mathscr{C}$ is understood;

when it is not empty, we say that (x_n) is *$\mathscr{C}$-convergent*. The following optional condition about the convergence structure $\mathscr{C}$ is to be considered

(conv-3) $\mathscr{C}$ is *separated*:
$\mathscr{C} - \lim_n(x_n)$ is an asingleton, for each sequence $(x_n; n \geq 0)$ in X;

some concrete cases will be given a bit further.

Likewise, by a (sequential) *Cauchy structure* on X we shall mean, as in Turinici [46], any part $\mathscr{H}$ of $\mathscr{S}(X)$ with

(Cauchy-1)$\mathscr{H}$ is *hereditary*:
$(x_n) \in \mathscr{H} \Longrightarrow (y_n) \in \mathscr{H}$, for each subsequence (y_n) of (x_n)
(Cauchy-2) $\mathscr{H}$ is *reflexive*:
$(\forall u \in X)$: the constant sequence $(x_n = u; n \geq 0)$ fulfills $(x_n) \in \mathscr{H}$.

Each element of $\mathscr{H}$ will be referred to as a $\mathscr{H}$*-Cauchy* sequence in X; and the couple $(X; \mathscr{H})$ will be called a *Cauchy space*. Finally, the pair $(\mathscr{C}, \mathscr{H})$ will be referred to as a *conv-Cauchy structure* on X; and the triple $(X, \mathscr{C}, \mathscr{H})$, as a *conv-Cauchy* space. The natural conditions about the conv-Cauchy structure $(\mathscr{C}, \mathscr{H})$ to be considered here are

(CC-1) $(\mathscr{C}, \mathscr{H})$ is *regular*:
each $\mathscr{C}$-convergent sequence in X is $\mathscr{H}$-Cauchy
(CC-2) $(\mathscr{C}, \mathscr{H})$ is *complete*:
each $\mathscr{H}$-Cauchy sequence in X is $\mathscr{C}$-convergent.

(B) In the following, a basic example of conv-Cauchy structure is given.

By a *pseudometric* over X we shall mean any map $d : X \times X \to R_+$. Suppose that we fixed such an object; with, in addition,

(met-1) d is *triangular*: $d(x, z) \leq d(x, y) + d(y, z)$, for all $x, y, z \in X$
(met-2) d is *reflexive*: $d(x, x) = 0$, for each $x \in X$
(met-3) d is *sufficient*: $d(x, y) = 0$ implies $x = y$
(met-4) d is *symmetric*: $d(x, y) = d(y, x)$, $\forall x, y \in X$.

In this case, d is called a *metric* on X; and the couple (X, d) will be referred to as a *metric space*.

We introduce a d-convergence and a d-Cauchy structure on X as follows. Given the sequence (x_n) in X and the point $x \in X$, we say that (x_n), d*-converges* to x (written as: $x_n \xrightarrow{d} x$) provided $d(x_n, x) \to 0$ as $n \to \infty$; i.e.,

$$\forall \varepsilon > 0, \exists i = i(\varepsilon):\ i \leq n \Longrightarrow d(x_n, x) < \varepsilon.$$

This will be also referred to as: x is a d*-limit* of (x_n); and written: $x \in \lim_n(x_n)$; when such points x exist, we say that (x_n) is d*-convergent*. By this very definition, we have the hereditary and reflexive properties:

(conv-1) $(\xrightarrow{d})$ is *hereditary*: $x_n \xrightarrow{d} x$ implies $y_n \xrightarrow{d} x$,
for each subsequence $(y_n; n \geq 0)$ of $(x_n; n \geq 0)$
(conv-2) $(\xrightarrow{d})$ is *reflexive*:
$(\forall u \in X)$: the constant sequence $(x_n = u; n \geq 0)$ fulfills $x_n \xrightarrow{d} u$;

hence, $(\xrightarrow{d})$ is a convergence structure on X. As precise, the following condition about this structure is to be considered

(conv-3) $(\xrightarrow{d})$ is *separated* (referred to as: d is *separated*):
$\lim_n(x_n)$ is an asingleton, for each sequence (x_n) in X.

Note that this holds under the conditions imposed upon d; and then, $\{x\} = \lim_n(x_n)$ will be written as $x = \lim_n(x_n)$.

Further, call the sequence (x_n), *d-Cauchy* when $d(x_m, x_n) \to 0$ as $m, n \to \infty$ with $m < n$; i.e.,

$$\forall \varepsilon > 0, \exists j = j(\varepsilon):\ j \le m < n \Longrightarrow d(x_m, x_n) < \varepsilon;$$

the class of all such sequences will be denoted as $Cauchy(X, d)$. Clearly, we have the hereditary and reflexive properties

(Cauchy-1) $Cauchy(X, d)$ is *hereditary*:
(x_n) is d-Cauchy implies (y_n) is d-Cauchy,
for each subsequence $(y_n; n \ge 0)$ of $(x_n; n \ge 0)$
(Cauchy-2) $Cauchy(X, d)$ is *reflexive*:
$(\forall u \in X)$: the constant sequence $(x_n = u; n \ge 0)$ is d-Cauchy;

so that, $Cauchy(X, d)$ is a Cauchy structure on X.

Now—according to the general setting—call the couple $((\xrightarrow{d}), Cauchy(X, d))$, a *conv-Cauchy structure* induced by d. Remember that the following regularity conditions about this structure are to be considered

(CC-1) d is *regular*: each d-convergent sequence in X is d-Cauchy
(CC-2) d is *complete*: each d-Cauchy sequence in X is d-convergent;

note that the former of these holds in our setting.

Finally, let us say that $(x_n; n \ge 0)$ is *d-semi-Cauchy*, provided

$d(x_n, x_{n+1}) \to 0$ as $n \to \infty$; or, equivalently:
$d(x_n, x_{n+i}) \to 0$ as $n \to \infty$, for each $i \ge 1$.

In this case, for each $\gamma > 0$,

$\mathscr{S}((x_n); \gamma) := \{k \in N; n \in N(k, \le), i \in \{1, 2\} \Longrightarrow d(x_n, x_{n+i}) < \gamma\}$
is nonempty; hence, $n(\gamma) := \min \mathscr{S}((x_n); \gamma)$ exists;

we then say that $n(\gamma)$ is the *semi-Cauchy rank* attached to γ. Clearly,

$$\gamma \mapsto n(\gamma) \text{ is decreasing: } \gamma_1 \le \gamma_2 \Longrightarrow n(\gamma_1) \ge n(\gamma_2).$$

Finally, it is immediate that each d-Cauchy sequence appears as d-semi-Cauchy too; the reciprocal of this is not in general true.

Note that an extended setting of these concepts is possible, under the lines sketched by Hitzler [15, Ch 1, Sect 1.2]; we shall discuss these facts elsewhere.

(C) Concerning these developments, the following auxiliary statement is useful in the sequel.

Proposition 5 *The mapping* $(x, y) \mapsto d(x, y)$ *is* d*-Lipschitz, in the sense*

$$|d(x, y) - d(u, v)| \leq d(x, u) + d(y, v),\ \ (x, y), (u, v) \in X \times X.$$

As a consequence, this map is d*-continuous; i.e.,*

$$x_n \xrightarrow{d} x, y_n \xrightarrow{d} y \text{imply}\, d(x_n, y_n) \to d(x, y).$$

The proof is immediate, by the triangular and symmetric properties of $d(.,.)$; so, further details are not needed.

4 Meir-Keeler Admissible Functions

In the following, a basic class of real valued functions is introduced.

(A) Denote for simplicity $\mathscr{F}_0(R_+) = \{\varphi \in \mathscr{F}(R_+); \varphi(0) = 0\}$. Then, let us put

$\mathscr{F}_0(re)(R_+)$=the subclass of all $\varphi \in \mathscr{F}_0(R_+)$, endowed with
φ=*regressive*: $\varphi(t) < t$, for all $t \in R_+^0 :=]0, \infty[$.

Call $\varphi \in \mathscr{F}_0(re)(R_+)$, *Meir-Keeler admissible* [23] if

$\forall \gamma > 0, \exists \beta > 0, (\forall t)$: $(\gamma < t < \gamma + \beta \Longrightarrow \varphi(t) \leq \gamma)$.

In the following, some important examples of such objects are given.

(B) For any $\varphi \in \mathscr{F}_0(re)(R_+)$ and any $s \in R_+^0$, put

$\Lambda_+\varphi(s) = \inf\{\Phi(s+)(\varepsilon); \varepsilon > 0\}$,
where $\Phi(s+)(\varepsilon) = \sup \varphi(]s, s + \varepsilon[), \varepsilon > 0$.

From the regressive property of φ, these quantities are finite; precisely,

$$0 \leq \Lambda_+\varphi(s) \leq s,\ \ \forall s \in R_+^0.$$

The following consequence of this will be useful. Given the sequence $(r_n; n \geq 0)$ in R and the point $r \in R$, let us write

$r_n \to r+$, if $r_n \to r$ and $(r_n > r$, for all $n \geq 0)$.

Proposition 6 *Let* $\varphi \in \mathscr{F}(re)(R_+)$ *and* $s \in R_+^0$ *be arbitrary fixed. Then,*

(41-1) $\limsup_n(\varphi(t_n)) \leq \Lambda_+\varphi(s)$,
for each sequence (t_n) *in* R_+^0 *with* $t_n \to s+$
(41-2) there exists a sequence (r_n) *in* R_+^0 *with*

$r_n \to s+$ *and* $\varphi(r_n) \to \Lambda_+\varphi(s)$.

Proof Denote, for simplicity,

$\alpha = \Lambda_+\varphi(s)$; hence, $\alpha = \inf_{\varepsilon>0} \Phi(s+)(\varepsilon)$, and $0 \leq \alpha \leq s$,

i) Given $\varepsilon > 0$, there exists a rank $p(\varepsilon) \geq 0$ such that

$s < t_n < s + \varepsilon$, for all $n \geq p(\varepsilon)$;

hence (by definition)

$$\limsup_n(\varphi(t_n)) \leq \sup\{\varphi(t_n); n \geq p(\varepsilon)\} \leq \Phi(s+)(\varepsilon).$$

Passing to infimum over $\varepsilon > 0$, yields (see above)

$$\limsup_n(\varphi(t_n)) \leq \inf_{\varepsilon>0} \Phi(s+)(\varepsilon) = \alpha;$$

and the claim follows.

ii) Define $(\beta_n := \Phi(s+)(2^{-n}); n \geq 0)$; this is a descending sequence in R_+, with

$$(\beta_n \geq \alpha,\ \forall n) \text{ and } \inf_n \beta_n = \alpha; \text{ hence } \lim_n \beta_n = \alpha.$$

By these properties, there may be constructed a sequence $(\gamma_n; n \geq 0)$ in R, with

$$\gamma_n < \beta_n,\ \forall n;\ \lim_n \gamma_n = \lim_n \beta_n = \alpha.$$

(For example, we may take $(\gamma_n = \beta_n - 3^{-n}; n \geq 0)$; but this is not the only possible choice). Let $n \geq 0$ be arbitrary fixed. By the supremum definition, there exists $r_n \in]s, s + 2^{-n}[$ such that $\varphi(r_n) > \gamma_n$; moreover (again by definition), $\varphi(r_n) \leq \beta_n$. The obtained sequence $(r_n; n \geq 0)$ fulfills $r_n \to s+$ and $\varphi(r_n) \to \alpha$; wherefrom, the desired conclusion is clear.

Call $\varphi \in \mathscr{F}_0(re)(R_+)$, *Boyd-Wong admissible* [5] if

$\Lambda_+\varphi(s) < s$, for all $s > 0$.

Sufficient conditions for this property are being described in

Proposition 7 *Suppose that* $\varphi \in \mathscr{F}_0(re)(R_+)$ *fulfills one of the conditions*

(42-1) φ *is upper semicontinuous at the right on* R_+^0*:*
$\Lambda_+\varphi(s) \leq \varphi(s)$, $\forall s \in R_+^0$
(42-2) φ *is continuous at the right on* R_+^0.

Then, φ *is Boyd-Wong admissible.*

Proof

i) Evident, by the regressiveness of φ.
ii) From the right continuous property,

$\Lambda_+\varphi(s) = \varphi(s), \forall s \in R_+^0$;

so, this case reduces to the preceding one.

(C) Denote for simplicity

$\mathscr{F}(in)(R_+)$=the class of all increasing functions in $\mathscr{F}(R_+)$,
$\mathscr{F}_0(re, in)(R_+) = \mathscr{F}_0(re)(R_+) \cap \mathscr{F}(in)(R_+)$.

Call $\varphi \in \mathscr{F}_0(re, in)(R_+)$, *Matkowski admissible* [21] provided

$\varphi^n(t) \to 0$ as $n \to \infty$, for all $t > 0$.

[Here, for each $n \geq 0$, φ^n stands for the n-th iterate of φ]. As a matter of fact, the iterative condition we just imposed assures us that φ is regressive; but, this is not important for our developments.

As before, we need sufficient conditions under which our property holds. For each $\varphi \in \mathscr{F}_0(re, in)(R_+)$, denote

$\varphi(s + 0) := \lim_{t\to s+} \varphi(t)$, $s \in R_+^0$

(the *right limit* of φ at s). It is not hard to see that the following evaluation holds

$\varphi(s) \leq \varphi(s + 0) \leq s$, for all $s > 0$;

we do not give details. Finally, denote (over the same class of functions)

$M(\varphi) = \{s > 0; \varphi(s + 0) = s\}$.

Clearly, the extremal case of $M(\varphi) = \emptyset$ cannot be avoided; just take the function $\varphi \in \mathscr{F}_0(re, in)(R_+)$ as

φ is right continuous on R_+^0: $\varphi(s + 0) = \varphi(s)$, for each $s \in R_+^0$.

Concerning the other extremal case $M(\varphi) = R_+^0$, the following simple (negative) answer is available (see also Jachymski [17]):

Proposition 8 *For each function $\varphi \in \mathscr{F}_0(re, in)(R_+)$, we have*

$M(\varphi)$ is at most denumerable in R_+^0; hence, $M(\varphi) = R_+^0$ is false.

Proof By the increasing property of our function φ, the subset

$\Gamma = \{s \in R_+^0; \varphi$ is right discontinuous at $s\}$

appears as (at most) countable; cf. Natanson [28, Ch 8, Sect 1]; so, necessarily, $\Theta := R_+^0 \setminus \Gamma$ is nonempty in R_+^0, and

$\varphi(s + 0) = \varphi(s) < s$, for each $s \in \Theta$; whence $M(\varphi) \cap \Theta = \emptyset$.

This tells us that, necessarily, $M(\varphi) \subseteq \Gamma$; and conclusion follows.

We may now state an appropriate answer to the posed question.

Proposition 9 *Supppose that the function $\varphi \in \mathscr{F}_0(re, in)(R_+)$ fulfills*

φ is strongly regressive: $\varphi(s+0) < s$, for each $s > 0$ (i.e.: $M(\varphi) = \emptyset$).

Then, φ is Matkowski admissible.

Proof Given $s_0 > 0$, let $(s_n = \varphi^n(s_0); n \geq 0)$ be its iterative sequence. If

$s_k = 0$, for some $k \geq 0$

then, by the decreasing property of (s_n), we have $(s_n = 0$, for all $n \geq k)$; and conclusion follows. It remains now to discuss the case of

$s_n > 0$, for all $n \geq 0$.

By the regressive property of φ, (s_n) is strictly descending; hence, $s := \lim_n s_n$ exists, with $s_n > s$, for all n. Suppose by contradiction that $s > 0$. Combining with

$\varphi(s+0) = \lim_n \varphi(s_n) = \lim_n s_{n+1}$,

yields $\varphi(s+0) = s$; contradiction. Hence, $s > 0$; and we are done.

Remark 1 The reverse inclusion is not (in general) true. To verify this, let us consider the function $\varphi \in \mathscr{F}_0(re, in)(R_+)$, according to (for some $r > 0$):

$(\varphi(t) = 0$, if $t \leq r)$, $(\varphi(t) = r$, if $t > r)$.

Clearly, φ is Matkowski admissible; we do not give details. On the other hand,

$\varphi(r+0) = r$; whence, φ is not strongly regressive;

and this proves our claim.

(D) Now, it is natural to establish the connection between the introduced classes and the Meir-Keeler one. An appropriate answer to this is contained in

Proposition 10 *Under these conventions, the following inclusions are valid:*

(45-1) if $\varphi \in \mathscr{F}_0(re)(R_+)$ is Boyd-Wong admissible,
then φ is Meir-Keeler admissible
(45-2) the function $\varphi \in \mathscr{F}_0(re, in)(R_+)$ is Matkowski admissible,
if and only if it is Meir-Keeler admissible.

Proof

i) (cf. Meir and Keeler [23]). Suppose that $\varphi \in \mathscr{F}(re)(R_+)$ is Boyd-Wong admissible; and fix $\gamma > 0$. As $\Lambda + \varphi(\gamma) < \gamma$, there exists $\beta = \beta(\gamma) > 0$ such that $\Phi(\gamma+)(\beta) < \gamma$; wherefrom, $\gamma < t < \gamma + \beta$ implies $\varphi(t) < \gamma$; hence the claim.

ii-1) (cf. Jachymski [16]). Assume that $\varphi \in \mathscr{F}_0(re, in)(R_+)$ is Matkowski admissible; we have to establish that it is Meir-Keeler admissible. If the underlying property fails, then (for some $\gamma > 0$):

$\forall \beta > 0, \exists t \in]\gamma, \gamma + \beta[$, such that $\varphi(t) > \gamma$.

As φ is increasing, this yields (by the arbitrariness of β)

$(\varphi(t) > \gamma, \forall t > \gamma)$; whence, by induction: $(\varphi^n(t) > \gamma, \forall n, \forall t > \gamma)$.

Taking some $t > \gamma$ and passing to limit as $n \to \infty$, one gets $0 \geq \gamma$; contradiction. Hence, φ is Meir-Keeler admissible, as claimed.

ii-2) Assume that $\varphi \in \mathscr{F}_0(re, in)(R_+)$ is Meir-Keeler admissible; we have to establish that it is Matkowski admissible. Given $s_0 > 0$, let $(s_n = \varphi^n(s_0); n \geq 0)$ be the iterates sequence. If

$s_k = 0$, for some $k \geq 0$

then, by the decreasing property of (s_n), we have $(s_n = 0$, for all $n \geq k)$; and conclusion follows. It remains now to discuss the case of

$s_n > 0$, for all $n \geq 0$.

By the regressive property of φ, (s_n) is strictly descending; hence, $s := \lim_n s_n$ exists, with (in addition) $s_n > s$, for all n. Suppose by contradiction that $s > 0$; and let $r > 0$ be the number assured by the Meir-Keeler admissible property of φ. By definition, there exists a rank $n(r) \geq 0$, such that

$n \geq n(r)$ implies $s < s_n < s + r$.

This, by the underlying property, gives (for the same ranks)

$s < s_{n+1} = \varphi(s_n) \leq s$; contradiction.

Hence, $s = 0$; wherefrom φ is Matkowski admissible.

Remark 2 Concerning the reverse of our first inclusion, the answer is (in general) negative. In fact, let us again consider the function $\varphi \in \mathscr{F}_0(re, in)(R_+)$, we just introduced (for some $r > 0$):

$(\varphi(t) = 0$, if $t \leq r)$, $(\varphi(t) = r$, if $t > r)$.

Clearly,

$\varphi(r + 0) = r$; i.e.: $\Lambda_+\varphi(s) = r$;

which tells us that φ is not Boyd-Wong admissible. On the other hand, by definition (and the previous developments)

φ is Matkowski admissible; hence, Meir-Keeler admissible;

and this proves our claim.

(E) A bilateral version of the "right" type statements above may be done as follows. For any $\varphi \in \mathscr{F}_0(re)(R_+)$ and any $s \in R_+^0$, put

$\Lambda_\pm\varphi(s) = \inf\{\Phi(s\pm)(\varepsilon); 0 < \varepsilon < s\}$,
where $\Phi(s\pm)(\varepsilon) = \sup \varphi(]s - \varepsilon, s + \varepsilon[), 0 < \varepsilon < s.$

From the regressive property of φ, these quantities are finite; precisely,

$$0 \leq \Lambda_+\varphi(s) \leq \Lambda_\pm\varphi(s) \leq s, \quad \forall s \in R_+^0.$$

The following consequence of this will be useful.

Proposition 11 *Let $\varphi \in \mathscr{F}_0(re)(R_+)$ and $s \in R_+^0$ be arbitrary fixed. Then,*

(46-1) $\limsup_n(\varphi(t_n)) \leq \Lambda_\pm\varphi(s)$,
for each sequence (t_n) in R_+^0 with $t_n \to s$
(46-2) there exists a sequence (r_n) in R_+^0 with
$r_n \to s$ and $\varphi(r_n) \to \Lambda_\pm\varphi(s)$.

The proof mimics its "right" version; however, for completeness reasons, we provide an argument in what follows.

Proof Denote, for simplicity,

$\alpha = \Lambda_\pm\varphi(s)$; hence, $\alpha = \inf\{\Phi(s\pm)(\varepsilon); 0 < \varepsilon < s\}$ and $0 \leq \alpha \leq s$,

i) Given ε in $]0, s[$, there exists a rank $p(\varepsilon) \geq 0$ such that

$s - \varepsilon < t_n < s + \varepsilon$, for all $n \geq p(\varepsilon)$;

hence (by definition)

$$\limsup_n(\varphi(t_n)) \leq \sup\{\varphi(t_n); n \geq p(\varepsilon)\} \leq \Phi(s\pm)(\varepsilon).$$

Passing to infimum over ε in $]0, s[$, yields (see above)

$$\limsup_n(\varphi(t_n)) \leq \inf\{\Phi(s\pm)(\varepsilon); 0 < \varepsilon < s\} = \alpha;$$

and the claim follows.

ii) Define $(\beta_n := \Phi(s\pm)(s2^{-n-1}); n \geq 0)$; this is a descending sequence in R_+, with the properties

$$(\beta_n \geq \alpha, \ \forall n) \text{ and } \inf_n \beta_n = \alpha; \text{ hence } \lim_n \beta_n = \alpha.$$

By these properties, there may be constructed a sequence $(\gamma_n; n \geq 0)$ in R, with

$$\gamma_n < \beta_n, \ \forall n; \ \lim_n \gamma_n = \lim_n \beta_n = \alpha.$$

(For example, we may take ($\gamma_n = \beta_n - 3^{-n}$; $n \geq 0$); but this is not the only possible choice). Let $n \geq 0$ be arbitrary fixed. By the supremum definition, there exists $r_n \in]s - s2^{-n-1}, s + s2^{-n-1}[$ such that $\varphi(r_n) > \gamma_n$; moreover (again by definition), $\varphi(r_n) \leq \beta_n$. The obtained (in R^0_+) sequence (r_n; $n \geq 0$) fulfills $r_n \to s$ and $\varphi(r_n) \to \alpha$; wherefrom, all is clear.

Having these precise, call $\varphi \in \mathscr{F}_0(re)(R_+)$, *bilateral Boyd-Wong admissible*, if

$\Lambda_\pm \varphi(s) < s$, for all $s > 0$;

note that, by a previous relation, any such function is Boyd-Wong admissible. Sufficient conditions for this property are being described in

Proposition 12 *Suppose that $\varphi \in \mathscr{F}_0(re)(R_+)$, fulfills one of the conditions*

(47-1) φ is upper semicontinuous on R^0_+: $\Lambda_\pm \varphi(s) \leq \varphi(s)$, $\forall s \in R^0_+$
(47-2) φ is continuous on R^0_+.

Then, φ is bilateral Boyd-Wong admissible.

As before, the proof of this is a direct translation of its "right" counterpart; so, further details are not given.

(F) A useful completion of these is the following. Let $\varphi \in \mathscr{F}_0(re)(R_+)$ be a function; we call it *Geraghty admissible* [13], provided

(t_n; $n \geq 0$)= sequence in R^0_+ and $\varphi(t_n)/t_n \to 1$ imply $t_n \to 0$.

Technically speaking, this class of functions may be viewed as a particular case of the previous (bilateral) Boyd-Wong one. Precisely, we have

Proposition 13 *Let $\varphi \in \mathscr{F}_0(re)(R_+)$ be Geraghty admissible. Then,*

(48-1) φ is bilateral Boyd-Wong admissible
(48-2) φ is Boyd-Wong admissible.

Proof

i) Suppose that $\varphi \in \mathscr{F}_0(re)(R_+)$ is not bilateral Boyd-Wong admissible. From a previous relation, there exists some $s \in R^0_+$ with $\Lambda_\pm \varphi(s) = s$. By the auxiliary fact we just exposed, there exists a sequence (r_n; $n \geq 0$) in R^0_+ with

$r_n \to s$ and $\varphi(r_n) \to s$; whence $\varphi(r_n)/r_n \to 1$;

i.e.: φ is not Geraghty admissible. The obtained contradiction proves our claim.

ii) Evident, by the observations above.

Remark 3 Concerning the reverse inclusion, note that, for the (continuous) bilateral Boyd-Wong admissible function in $\mathscr{F}_0(re)(R_+)$

$\varphi(t) = t(1 - e^{-t})$, $t \geq 0$,

and the sequence ($t_n = n + 1$; $n \geq 0$) in R^0_+, we have

$\varphi(t_n)/t_n \to 1$; but, evidently, $t_n \to \infty$.

Hence, φ is not Geraghty admissible; so that the reciprocals of both these inclusions are not in general true.

5 Statement of the Problem

Let X be a nonempty set; and $d : X \times X \to R_+$ be a *metric* [i.e.: a triangular, reflexive, sufficient, symmetric pseudometric] on X; then, (X, d) will be referred to as a *metric space*. Further, let $(\leq)$ be a *quasi-order* [i.e.: a reflexive transitive relation] over X; then, $(X, d, \leq)$ will be referred to as a *quasi-ordered metric space*.

Call the subset Y of X, $(\leq)$-*asingleton* if $[y_1, y_2 \in Y, y_1 \leq y_2]$ imply $y_1 = y_2$; and $(\leq)$-*singleton* if, in addition, Y is nonempty. Clearly, the generic inclusions hold

$$(\forall Y \in \exp[X]) : \text{ asingleton } \Longrightarrow (\leq)\text{-asingleton, singleton } \Longrightarrow (\leq)\text{-singleton.}$$

An instance when the reciprocal inclusions hold too is described as follows.

Proposition 14 *Suppose that*

(dir) X is $(\leq)$-directed: $\forall z_1, z_2 \in X$, $\exists z_3 \in X$: $z_1, z_2 \leq z_3$.

Then,

(51-1) The generic inclusions are valid

$$(\forall Y \in \exp[X]) : (\leq)\text{-}\textit{asingleton} \Longrightarrow \textit{asingleton}, \ (\leq)\text{-}\textit{singleton} \Longrightarrow \textit{singleton.}$$

(51-2) Hence, in the precise context,

$$(\forall Y \in \exp[X]) : (\leq)\text{-}\textit{asingleton} \Longleftrightarrow \textit{asingleton}, \ (\leq)\text{-}\textit{singleton} \Longleftrightarrow \textit{singleton.}$$

Proof It will suffice verifying the first half of this conclusion. Let the subset Y of X be $(\leq)$-asingleton; and take a couple $y_1, y_2 \in Y$. By the directed property, there exists $y_3 \in X$, such that $y_1, y_2 \leq y_3$. This yields (by hypothesis)

$$y_1 = y_3, \ y_2 = y_3; \text{ whence } y_1 = y_2.$$

The proof is complete.

(A) Take some $T \in \mathscr{F}(X)$. Assume in the following that

(s-pro) T is *semi-progressive*: $X(T, \leq) := \{x \in X; x \leq Tx\} \neq \emptyset$
(incr) T is *increasing*: $x \leq y$ implies $Tx \leq Ty$.

Concerning the former of these conditions, the following aspect is to be noted. Let $(<)$ stand for the relation

$x < y$ iff $x \leq y$ and $x \neq y$;

clearly, $(<)$ is irreflexive but not in general transitive, as long as $(\leq)$ is not antisymmetric. Now, evidently,

(s-s-pro) T is *strongly semi-progressive*: $X(T, <) := \{x \in X; x < Tx\} \neq \emptyset$

is a particular case of the semi-progressive condition above; however, it is not obtainable from the underlying property, as simple examples show.

Now, as in the trivial quasi-order case, we are interested in establishing sufficient conditions for the determination of elements in $Fix(T)$. The basic directions for getting these fixed points are described in our list below, comparable with the one proposed by Turinici [47]:

pic-0) We say that T is *fix-*$(\leq)$*-asingleton*, when $\mathrm{Fix}(T)$ is $(\leq)$-asingleton; and *fix-*$(\leq)$*-singleton* when $\mathrm{Fix}(T)$ is $(\leq)$-singleton

pic-1) We say that $x \in X(T, \leq)$ is a *Picard point* (modulo $(d, \leq; T)$) if the iterative sequence $(T^n x; n \geq 0)$ is d-Cauchy; when this property holds for all $x \in X(T, \leq)$, then T is called a *Picard operator* (modulo $(d, \leq)$)

pic-2) We say that $x \in X(T, \leq)$ is a *strong Picard point* (modulo $(d, \leq; T)$) if the iterative sequence $(T^n x; n \geq 0)$ is d-convergent and $\lim_n (T^n x) \in \mathrm{Fix}(T)$; when this property holds for all $x \in X(T, \leq)$, then T is called a *strong Picard operator* (modulo $(d, \leq)$)

pic-3) We say that $x \in X(T, \leq)$ is a *Bellman Picard point* (modulo $(d, \leq; T)$) if the iterative sequence $(T^n x; n \geq 0)$ is d-convergent and $T^n x \leq \lim_n (T^n x) \in \mathrm{Fix}(T), \forall n \geq 0$; when this property holds for all $x \in X(T, \leq)$, then T is called a *Bellman Picard operator* (modulo $(d, \leq)$).

The sufficient (regularity) conditions for such properties are being founded on *strictly ascending strongly orbital full* concepts (in short: (asa-so-f)-concepts) and on *strictly ascending strongly orbital full* concepts (in short: (sa-so-f)-concepts). Call the sequence $(z_n; n \geq 0)$ in X,

strictly ascending, if $z_i < z_j$ whenever $i < j$;
strongly T*-orbital*, if $(z_n = T^n x; n \geq 0)$, for some $x \in X$;
full, when $n \mapsto z_n$ is injective ($i < j$ implies $x_i \neq x_j$);

the intersection of these notions yields the precise ones.

reg-1) Call X, *(sa-so-f,d)-complete* provided

(for each (sa-so-f)-sequence) d-Cauchy $\Longrightarrow$ d-convergent.

reg-2) We say that T is *(sa-so-f,d)-continuous*, if

$((z_n)$=(sa-so-f)-sequence and $z_n \xrightarrow{d} z)$ imply $Tz_n \xrightarrow{d} Tz$.

reg-3) Call $(\leq)$, *(sa-so-f,d)-selfclosed*, when

$((z_n)$=(sa-so-f)-sequence and $z_n \xrightarrow{d} z)$ imply $(z_n \leq z$, for each $n \geq 0)$.

When the strong orbital properties are ignored, these conventions may be written in terms of (almost strictly ascending full) and (strictly ascending full) sequences; we do not give details.

(B) As an essential completion of these facts, we have to discuss the contractive type conditions to be used. Let us introduce the mappings (for $x, y \in X$)

$M_1(x, y) = d(Tx, Ty)$, $M_2(x, y) = d(x, y)$, $M_3(x, y) = d(x, Tx)$,
$M_4(x, y) = d(y, Ty)$, $M_5(x, y) = d(x, Ty)$, $M_6(x, y) = d(Tx, y)$.

By taking elementary order/algebraic combinations between these, one gets a lot of functions to be used in our reasonings; the basic ones are

$A_1 = M_2$, $A_2 = (1/2)[M_3 + M_4]$,
$A_3 = \max\{M_3, M_4\}$, $A_4 = (1/2)[M_5 + M_6]$;
or, explicitly (for $x, y \in X$)
$A_1(x, y) = d(x, y)$, $A_2(x, y) = (1/2)[d(x, Tx) + d(y, Ty)]$,
$A_3(x, y) = \max\{d(x, Tx), d(y, Ty)\}$, $A_4(x, y) = (1/2)[d(x, Ty) + d(Tx, y)]$.

Let $P : X \times X \to R_+$ be a map. For an easy reference, we give the list of *orbital normality conditions* (to be fulfilled – or not – by $P(., .)$):

(o-nor-1) P is *orbitally small*:
for each d-semi-Cauchy (sa-so-f)-sequence $(x_n = T^n x_0; n \geq 0)$ in $X(T, \leq)$ and each (ε, δ) with $\varepsilon > \delta > 0$, there exists $\gamma \in]0, \delta/2[$ (and the attached semi-Cauchy rank $n(\gamma) \geq 0$), such that:
for each $j \geq 2$, $k \geq n(\gamma)$ with $d(x_m, x_{m+i}) < \varepsilon + \delta/2$ for $(m \geq k, i \in \{1, \dots, j\})$, we have:
$P(x_n, x_{n+j}) < \varepsilon + \delta$, whenever $(n \geq k, d(x_n, x_{n+j+1}) \geq \varepsilon + \delta/2)$
(o-nor-2) P is *orbitally singular asymptotic*:
for each (sa-so-f)-sequence $(x_n = T^n x_0; n \geq 0)$ in $X(T, \leq)$,
and each $z \in X$ with
$(x_n \xrightarrow{d} z, Tx_n \xrightarrow{d} z)$, $(x_n < z$ for almost all $n)$, and $d(z, Tz) > 0$,
we have $\liminf_n P(x_n, z) < d(z, Tz)$
(o-nor-3) P is *orbitally regular asymptotic*:
for each (sa-so-f)-sequence $(x_n = T^n x_0; n \geq 0)$ in $X(T, \leq)$,
and each $z \in X$ with
$(x_n \xrightarrow{d} z, Tx_n \xrightarrow{d} z)$, $(x_n < z$ for almost all $n)$, and $d(z, Tz) > 0$,
we have $P(x_n, z) \to d(z, Tz)$
(o-nor-4) P is *orbitally strongly regular asymptotic*:
for each (sa-so-f)-sequence $(x_n = T^n x_0; n \geq 0)$ in $X(T, \leq)$,
and each $z \in X$ with
$(x_n \xrightarrow{d} z, Tx_n \xrightarrow{d} z)$, $(x_n < z$ for almost all $n)$, and $d(z, Tz) > 0$,
we have $P(x_n, z) \to\to d(z, Tz)$.

Here, a property $\pi(n)$ involving $n \in N$ is holding for *almost all n*, provided

there exists $h \in N$ such that $\pi(n)$ holds for all $n \geq h$.

Likewise, given the sequence $(r_n; n \geq 0)$ in R and the point $r \in R$, we denoted

$r_n \to\to r$, if $r_n \to r$ and there exists a subsequence
$(s_n = r_{i(n)}; n \geq 0)$ of $(r_n; n \geq 0)$ such that $[s_n = r, \forall n \geq 0]$.

Further, starting from the same mapping $P : X \times X \to R_+$, we give the list of *global normality conditions* (to be fulfilled—or not—by $P(.,.)$):

(g-nor-1) P is $(\leq)$-*sufficient*:
$P(x, y) > 0$, for each $x, y \in X, x < y$
(g-nor-2) P is *telescopic bounded*:
for each $x \in X(T, <)$, we have $P(x, Tx) \leq A_3(x, Tx)$
(g-nor-3) P is *fix-bounded*:
$(x, y \in \text{Fix}(T), x < y) \Longrightarrow P(x, y) \leq M(x, y)$
(g-nor-4) P is *chain diametrally bounded*:
for each $x, y \in X$ taken so as (x, Tx, y, Ty) is $(<)$-chain,
we have $P(x, y) \leq M(x, y)$
(g-nor-5) P is *diametrally bounded*:
for each $x, y \in X$ we have $P(x, y) \leq M(x, y)$.

Here, for each (nonempty) subset Z in X, we put

$\text{diam}(Z) = \sup\{d(x, y); x, y \in Z\}$ (the *diameter* of Z);

and the element in $M \in \mathscr{F}(X \times X, R_+)$ introduced via

$M(x, y) = \text{diam}\{x, Tx, y, Ty\}, x, y \in X$,

will be referred to as the *diameter mapping* relative to T. Likewise, the 4-tuple $(z_1, z_2, z_3, z_4) \in X^4$ is called a $(<)$-*chain*, provided

$z_i < z_j$, whenever $i < j$.

Concerning the orbitally small concept above, the following practical criteria will be useful for us.

Proposition 15 *Under the above conventions,*

(52-1) If the mapping $P : X \times X \to R_+$ is chain diametrally bounded, then it is orbitally small
(52-2) If the maps $P_1, P_2 : X \times X \to R_+$ are orbitally small, then $P_3 := \max\{P_1, P_2\}$ is orbitally small.

Proof

i) Let the d-semi-Cauchy (sa-so-f)-sequence $(x_n = T^n x_0; n \geq 0)$ in $X(T, \leq)$ and the couple (ε, δ) with $\varepsilon > \delta > 0$ be given. Further, take some $\gamma \in]0, \delta/2[$; and let $n(\gamma)$ stand for the attached semi-Cauchy rank. We claim that for each $j \geq 2$ and $k \geq n(\gamma)$ with

$d(x_m, x_{m+i}) < \varepsilon + \delta/2$ for $(m \geq k, i \in \{1, \ldots, j\})$, the relations below hold

$P(x_n, x_{n+j}) < \varepsilon + \delta$, for each $n \geq k$;

and this will complete the argument. In fact, let $n \geq k$ be arbitrary fixed. By the very choice of our sequence,

$(x_n, x_{n+1}, x_{n+j}, x_{n+j+1})$ is a $(<)$-chain in X^4.

On the other hand, by the working hypothesis above, we have

$d(x_n, x_{n+j}), d(x_{n+1}, x_{n+j}), d(x_{n+1}, x_{n+j+1}) < \varepsilon + \delta/2$;

and, by the very definition of our index $n(\gamma)$, one gets (as $k \leq n(\gamma)$)

$d(x_n, x_{n+1}), d(x_{n+j}, x_{n+j+1}) < \gamma < \delta/2 < \varepsilon + \delta/2$.

Finally, taking the triangular inequality into account, one gets (by the choice of γ)

$d(x_n, x_{n+j+1}) \leq d(x_n, x_{n+1}) + d(x_{n+1}, x_{n+j+1}) < \gamma + \varepsilon + \delta/2 < \varepsilon + \delta$.

Putting these together, yields (by the chain diametrally bounded property)

$P(x_n, x_{n+j}) \leq M(x_n, x_{n+j}) < \varepsilon + \delta$;

and our claim follows.

ii) Given the couple (ε, δ) with $\varepsilon > \delta > 0$, let $\gamma_1 \in]0, \delta/2[$ (with the associated semi-Cauchy rank $n(\gamma_1)$) and $\gamma_2 \in]0, \delta/2[$ (with the associated semi-Cauchy rank $n(\gamma_2)$) be assured by the orbitally small property of P_1 and P_2, respectively. Then, let us put

$\gamma_3 = \min\{\gamma_1, \gamma_2\}$, $n(\gamma_3)$ = the associated semi-Cauchy rank (hence, $n(\gamma_3) \geq \max\{n(\gamma_1), n(\gamma_2)\}$);

we claim that the desired property of P_3 is fulfilled with respect to the obtained pair. In fact, let $j \geq 2$, $k \geq n(\gamma_3)$ be such that

$d(x_m, x_{m+i}) < \varepsilon + \delta/2$ for $(m \geq k, i \in \{1, \ldots, j\})$;

we have to establish that

$P_3(x_n, x_{n+j}) < \varepsilon + \delta$, whenever $(n \geq k, d(x_n, x_{n+j+1}) \geq \varepsilon + \delta/2)$.

To verify this, note that, by the imposed hypothesis

$d(x_m, x_{m+i}) < \varepsilon + \delta/2$ for $(m \geq k \geq n(\gamma_k), i \in \{1, \ldots, j\}, r \in \{1, 2\})$.

On the other hand, letting $n \geq k$ be as in the premise above, we have

$n \geq k \geq n(\gamma_k), d(x_n, x_{n+j+1}) \geq \varepsilon + \delta/2, r \in \{1, 2\}$.

Putting these together, gives (by the admitted properties of P_1 and P_2)

$P_k(x_n, x_{n+j}) < \varepsilon + \delta, r \in \{1, 2\}$; whence, $P_3(x_n, x_{n+j}) < \varepsilon + \delta$;

and the conclusion follows.

(C) Finally, given the mapping $P : X \times X \to R_+$, let us say that T is *Meir-Keeler* $(d, \leq; P)$*-contractive*, in case

(mk-1) $[x < y, P(x, y) > 0]$ imply $d(Tx, Ty) < P(x, y)$;
referred to as: T is *strictly nonexpansive* (modulo $(d, \leq; P)$)
(mk-2) for each $\varepsilon > 0$, there exists $\delta > 0$, such that:
$(x < y, \varepsilon < P(x, y) < \varepsilon + \delta) \Longrightarrow d(Tx, Ty) \leq \varepsilon$;
expressed as: T has the *Meir-Keeler property* (modulo $(d, \leq; P)$).

Note that, by the former of these, the Meir-Keeler property may be written as

(mk-3) for each $\varepsilon > 0$, there exists $\delta > 0$, such that:
$(x < y, 0 < P(x, y) < \varepsilon + \delta) \Longrightarrow d(Tx, Ty) \leq \varepsilon$.

In particular, when $P = A_1$, this convention is comparable with the standard one proposed by Meir and Keeler [23] and refined by Matkowski [22]; see also Ćirić [9]. Further aspects may be found in Jachymski [16] and Samet [37].

In the following, two basic examples of such contractions are given.

(I) Given $\varphi \in \mathscr{F}_0(R_+)$, call T, $(d, \leq; P; \varphi)$*-contractive* if

(contr-1) $d(Tx, Ty) \leq \varphi(P(x, y)), \forall x, y \in X, x < y, P(x, y) > 0$.

Proposition 16 *Assume that T is $(d, \leq; P; \varphi)$-contractive, where $\varphi \in \mathscr{F}_0(re)(R_+)$ is Meir-Keeler admissible. Then, T is Meir-Keeler $(d, \leq; P)$-contractive.*

Proof

i) Let $x, y \in X$ be such that $x < y$, $P(x, y) > 0$. By the contractive condition and φ=regressive

$d(Tx, Ty) \leq \varphi(P(x, y)) < P(x, y)$;

so, T is strictly nonexpansive (modulo $(d, \leq; P)$).

ii) Let $\varepsilon > 0$ be arbitrary fixed; and $\delta > 0$ be the number assured by the Meir-Keeler admissible property of φ. Further, let $x, y \in X$ be such that $x < y$, and $\varepsilon < P(x, y) < \varepsilon + \delta$. By the contractive condition and admissible property,

$$d(Tx, Ty) \leq \varphi(P(x, y)) \leq \varepsilon;$$

so that, T has the Meir-Keeler property (modulo $(d, \leq; P)$).

(II) Let us say that (ψ, φ) is an *admissible pair* of functions in $\mathscr{F}_0(R_+)$, if

(a-p) ψ is increasing, and φ is sufficient ($\varphi(t) > 0$ if $t > 0$).

The following sequential type conditions involving this couple are considered

(seq-1) φ is right sequentially positive:
$(\limsup_n \varphi(t_n) > 0$ when $t_n \to \varepsilon+)$, for each $\varepsilon > 0$

(seq-2) ψ is φ-bounded left oscillating:
$\varphi(\varepsilon) > \psi(\varepsilon) - \psi(\varepsilon - 0)$, for each $\varepsilon > 0$
(seq-3) ψ is φ-bounded bilateral oscillating:
($\limsup_n \varphi(t_n) > \psi(\varepsilon + 0) - \psi(\varepsilon - 0)$, when $t_n \to \varepsilon$), for each $\varepsilon > 0$.

Here, given the sequence $(r_n; n \geq 0)$ in R and the point $r \in R$, we write

$r_n \to r+$ ($r_n \to r-$) if $r_n \to r$ and ($r_n > r$ ($r_n < r$), for all $n \geq 0$).

Now, given the mapping $P : X \times X \to R_+$ and the couple (ψ, φ) of functions in $\mathscr{F}_0(R_+)$, let us say that T is *$(d, \leq; P; (\psi, \varphi))$-contractive*, provided

(contr-2) $\psi(d(Tx, Ty)) \leq \psi(P(x, y)) - \varphi(P(x, y))$,
$\forall x, y \in X, x < y, P(x, y) > 0$.

Proposition 17 *Suppose that T is $(d, \leq; P; (\psi, \varphi))$-contractive, for an admissible pair (ψ, φ) of functions in $\mathscr{F}_0(R_+)$, with φ = right sequentially positive. Then, necessarily, T is Meir-Keeler $(d, \leq; P)$-contractive, in (ZF-AC+DC).*

Proof

i) Let $x, y \in X$ be such that $x < y$, $P(x, y) > 0$. As φ is sufficient, we derive

$\varphi(P(x, y)) > 0$; wherefrom $\psi(d(Tx, Ty)) < \psi(P(x, y))$.

This, via [ψ=increasing], yields $d(Tx, Ty) < P(x, y)$; so that, T is strictly nonexpansive (modulo $(d, \leq; P)$).

ii) Assume by contradiction that T does not have the Meir-Keeler property (modulo $(d, \leq; P)$); i.e. (for some $\varepsilon > 0$)

$C(\delta) := \{(u, v) \in X \times X; u < v, \varepsilon < P(u, v) < \varepsilon + \delta,\ d(Tu, Tv) > \varepsilon\}$
is nonempty, for each $\delta > 0$.

Taking a zero converging sequence $(\delta_n; n \geq 0)$ in R^0_+, we get by the Denumerable Axiom of Choice (AC(N)) [deductible, as precise, in (ZF-AC+DC)], a couple of sequences $(x_n; n \geq 0)$ and $(y_n; n \geq 0)$ in X, so as

($\forall n$): (x_n, y_n) is an element of $C(\delta_n)$;

or, equivalently (by definition and preceding step)

$$(\forall n): \ \ x_n < y_n,\ \varepsilon < d(Tx_n, Ty_n) < P(x_n, y_n) < \varepsilon + \delta_n.$$

By the contractive condition, we get

$$\psi(d(Tx_n, Ty_n)) \leq \psi(P(x_n, y_n)) - \varphi(P(x_n, y_n)),\ \ \forall n;$$

or, equivalently,

$$(0 <)\ \varphi(P(x_n, y_n)) \leq \psi(P(x_n, y_n)) - \psi(d(Tx_n, Ty_n)),\ \ \forall n.$$

From the preceding relation, $P(x_n, y_n) \to \varepsilon+$, $d(Tx_n, Ty_n) \to \varepsilon+$; so, passing to lim sup as $n \to \infty$,

$$(0 \leq) \ \limsup_n \varphi(P(x_n, y_n)) \leq \psi(\varepsilon + 0) - \psi(\varepsilon + 0) = 0.$$

But, as φ is right sequentially positive, the obtained relations cannot hold simultaneously. Hence, the working hypothesis is not acceptable; and the Meir-Keeler property follows. Putting these together, ends the argument.

(III) A basic particular case of this last construction is to be described as follows. Take a triple of functions (ψ, λ, μ) over $\mathscr{F}_0(R_+)$; we call it *admissible*, provided

(admi) $\psi(.)$ is increasing and the function $\chi := \mu - \lambda + \psi$
is sufficient ($\chi(t) > 0$ when $t > 0$); whence, $\chi \in \mathscr{F}_0(R_+)$

The following sequential type conditions involving this triple are considered

(sequ-1) $\mu - \lambda$ is right-lsc (resp., lsc) on R^0_+
(sequ-2) ψ is left continuous on R^0_+ ($\psi(s - 0) = \psi(s), \forall s > 0$).
(sequ-3) ψ is continuous on R^0_+ ($\psi(s + 0) = \psi(s - 0), \forall s > 0$).

Here, a function $\gamma \in \mathscr{F}(R_+)$ is called *right-lsc* (resp., *lsc*) on R^0_+, provided

$\liminf_{t \to s+} \gamma(t) \geq \gamma(s)$, (resp., $\liminf_{t \to s+} \gamma(t) \geq \gamma(s)$), $\forall s \in R^0_+$.

Proposition 18 *Let the triple of functions (ψ, λ, μ) over $\mathscr{F}_0(R_+)$ be admissible (see above). Then,*

(55-1) if $\mu - \lambda$ is right-lsc on R^0_+, then χ is sequentially positive
(55-2) if (in addition to the right-lsc property), if ψ is left continuous on R^0_+, then ψ is χ-bounded left oscillating
(55-3) if (in addition to the right-lsc property), if $\mu - \lambda$ is lsc on R^0_+, and ψ is χ-bounded bilateral oscillating.

Proof

i) Suppose by contradiction that there exist $\varepsilon > 0$ and a sequence $(t_n; n \geq 0)$ in R^0_+ with $t_n \to \varepsilon+$, such that

$\limsup_n(\chi(t_n)) = 0$; hence, $\liminf_n(\chi(t_n)) = 0$.

By the properties of lim inf operator, one gets (as ψ is increasing and $\mu - \lambda = \chi - \psi$)

$\liminf_n(\mu(t_n) - \lambda(t_n)) \leq \liminf_n(\chi(t_n)) - \psi(\varepsilon + 0) = -\psi(\varepsilon + 0)$.

As $\mu - \lambda$ is right-lsc on R^0_+, we get

$\mu(\varepsilon) - \lambda(\varepsilon) \leq \liminf_n(\mu(t_n) - \lambda(t_n)) \ \leq -\psi(\varepsilon + 0) \leq -\psi(\varepsilon)$;
or, equivalently, $\chi(\varepsilon) \leq 0$;

in contradiction with the sufficiency of χ. As a consequence, our working assumption is not acceptable; and the claim follows.

ii) For each $\varepsilon > 0$, we have (as χ is sufficient and ψ is left continuous)

$$\chi(\varepsilon) > 0 = \psi(\varepsilon) - \psi(\varepsilon - 0);$$

and the assertion follows.

iii) For each $\varepsilon > 0$, we have (as χ is sufficient and ψ is bilaterally continuous)

$$\limsup_n(\chi(t_n)) \geq \liminf_n(\mu(t_n) - \lambda(t_n) + \psi(t_n)) \geq \mu(\varepsilon) - \lambda(\varepsilon) + \psi(\varepsilon)$$
$$= \chi(\varepsilon) > 0 = \psi(\varepsilon + 0) - \psi(\varepsilon - 0), \text{ when } t_n \to \varepsilon;$$

whence the claim.

Now, given the mapping $P : X \times X \to R_+$ and the triple (ψ, λ, μ) of functions in $\mathscr{F}_0(R_+)$, let us say that T is $(d, \leq; P; (\psi, \lambda, \mu))$*-contractive*, provided

(contr-3) $\psi(d(Tx, Ty)) \leq \lambda(P(x, y)) - \mu(P(x, y))$,
$\forall x, y \in X, x < y, P(x, y) > 0$.

Proposition 19 *Suppose that T is $(d, \leq; P; (\psi, \lambda, \mu))$-contractive, for an admissible triple (ψ, λ, μ) of functions in $\mathscr{F}_0(R_+)$, with $\mu - \lambda$ = right-lsc on R^0_+. Then, necessarily, T is Meir-Keeler $(d, \leq; P)$-contractive in (ZF-AC+DC).*

The proof is immediately obtainable from the preceding auxiliary fact; we do not give details.

6 Main Result

Let $(X, d, \leq)$ be a quasi-ordered metric space; and $T \in \mathscr{F}(X)$ be a selfmap of X; supposed to be semi-progressive and increasing. The general directions under which the problem of determining fixed points of T is to be solved were already made precise; moreover, the (sufficient) regularity conditions and metrical contractive properties of the same were settled.

The main (fixed point) result of this exposition (referred to as *Function Meir-Keeler theorem*; in short: (MK-f)) may be stated as below.

Theorem 2 *Assume (under the precise general conditions) that T is Meir-Keeler $(d, \leq; P)$-contractive, for some mapping $P : X \times X \to R_+$, with:*

(co-basic) P is $(\leq)$-sufficient, telescopic bounded, and orbitally small.

In addition, let X be (sa-so-f,d)-complete. Then

(61-a) T is a strong Picard operator (modulo $(d, \leq)$), provided the following extra condition holds

(exco-a1) T is (sa-so-f,d)-continuous

(61-b) T is a Bellman Picard operator (modulo $(d, \leq)$), whenever $(\leq)$ is (sa-so-f,d)-selfclosed and one of the following extra conditions holds

(exco-b1) P is orbitally singular asymptotic
(exco-b2) P is orbitally regular asymptotic, and T is
$(d, \leq; P; \varphi)$-contractive, where $\varphi \in \mathscr{F}_0(re)(R_+)$ is bilateral Boyd-Wong admissible (hence, Meir-Keeler admissible)
(exco-b3) P is orbitally strongly regular asymptotic, and T is $(d, \leq; P; \varphi)$-contractive, where $\varphi \in \mathscr{F}_0(re)(R_+)$ is Meir-Keeler admissible
(exco-b4) P is orbitally regular asymptotic, and T is
$(d, \leq; P; (\psi, \varphi))$-contractive, for an admissible couple (ψ, φ) of functions in $\mathscr{F}_0(R_+)$, such that φ is sequentially positive and ψ is φ-bounded
bilateral oscillating
(exco-b5) P is orbitally strongly regular asymptotic, and T is
$(d, \leq; P; (\psi, \varphi))$-contractive, for an admissible couple (ψ, φ) of functions in $\mathscr{F}_0(R_+)$, such that φ is sequentially positive and ψ is φ-bounded
left oscillating

(61-c) T is fix-$(\leq)$-asingleton (hence, fix-$(\leq)$-singleton) whenever (in addition to the above setting)

(exco-c1) P is fix-bounded (see above)

(61-d) T is fix-asingleton (hence, fix-singleton) whenever (in addition to the above setting) the fix-bounded condition we just listed is combined with

(exco-d1) X is $(\leq)$-directed.

Proof There are several steps to be passed.

Part 0 We start with the last two affirmations in the statement. Assume that

P is *fix-bounded*: $(x, y \in \text{Fix}(T), x < y) \Longrightarrow P(x, y) \leq M(x, y)$;

and let $z_1, z_2 \in \text{Fix}(T)$ be such that $z_1 \leq z_2$. If, by absurd,

$z_1 \neq z_2$; hence, $z_1 < z_2$,

we necessarily have

$P(z_1, z_2) > 0$ (as P is $(\leq)$-sufficient);

so that, by the strict nonexpansive property (modulo $(d, \leq; P)$)

$d(z_1, z_2) = d(Tz_1, Tz_2) < P(z_1, z_2)$.

On the other hand, clearly,

$P(z_1, z_2) \leq M(z_1, z_2) = d(z_1, z_2)$ (by the fix-bounded property).

The contradiction at which we arrived shows that our working assumption is not acceptable; and then, our first affirmation follows. Finally, the second affirmation is evident, by an auxiliary fact above.

Having these precise, we may now pass to the basic part of our developments. Take some $x_0 \in X(T, \leq)$; and put $(x_n = T^n x_0; n \geq 0)$; clearly, this is an ascending strongly orbital sequence. If $x_n = x_{n+1}$ for some $n \geq 0$, we are done; so, without loss, one may assume that, for each $n \geq 0$,

(asa-cond) $x_n \neq x_{n+1}$; hence, $x_n < x_{n+1}$, $\rho_n := d(x_n, x_{n+1}) > 0$.

Part 1 We firstly assert that, for each $n \geq 0$,

$(0 <)\ \rho_{n+1} < P(x_n, x_{n+1}) \leq \rho_n$, (hence, $(0 <)\rho_{n+1} < \rho_n$).

In fact, let $n \geq 0$ be arbitrary fixed. Clearly,

$P(x_n, x_{n+1}) > 0$ (since P is $(\leq)$-sufficient);

so that, by the strict nonexpansive property of T (modulo $(d, \leq; P)$),

$$\rho_{n+1} = d(Tx_n, Tx_{n+1}) < P(x_n, x_{n+1}).$$

On the other hand, as P is telescopic-bounded, we must have

$$P(x_n, x_{n+1}) \leq A_3(x_n, x_{n+1}) = \max\{\rho_n, \rho_{n+1}\}.$$

Combining with the preceding relation gives, for each $n \geq 0$,

$\rho_{n+1} < \max\{\rho_n, \rho_{n+1}\}$; wherefrom: $\rho_{n+1} < \rho_n$, $A_3(x_n, x_{n+1}) = \rho_n$;

and the claim follows.

Part 2 From the preceding part, one derives $(\rho_{n+1} < \rho_n, \forall n)$; so that, the sequence $(\rho_n; n \geq 0)$ is strictly descending. As a consequence, $\rho := \lim_n \rho_n$ exists as an element of R_+. Assume by contradiction that $\rho > 0$; and let $\sigma > 0$ be the number given by the Meir-Keeler property of T (modulo $(d, \leq; P)$). By definition, there exists a rank $n(\sigma)$ such that

$n \geq n(\sigma)$ implies $\rho < \rho_n < \rho + \sigma$.

On the other hand, taking a previous relation into account, we have

$(\forall n)$: $(0 <)\ \rho_{n+1} < P(x_n, x_{n+1}) \leq \rho_n$.

From the preceding relation involving (ρ_n), we then have

$n \geq n(\sigma)$ implies ($x_n < x_{n+1}$ and) $\rho < P(x_n, x_{n+1}) < \rho + \sigma$;

so that, by the Meir-Keeler property,

$(\forall n \geq n(\sigma))$: $\rho < \rho_{n+1} = d(Tx_n, Tx_{n+1}) \leq \rho$;

a contradiction. Hence, $\rho = 0$; so that,

$\rho_n := d(x_n, x_{n+1}) = d(x_n, Tx_n) \to 0$, as $n \to \infty$;

or, in other words: $(x_n; n \geq 0)$ is d-semi-Cauchy.

Part 3 Suppose that

there exist $i, j \in N$ such that $i < j, x_i = x_j$.

Denoting $p = j - i$, we thus have $p > 0$ and $x_i = x_{i+p}$; so that

$x_{i+1} = x_{i+p+1}$; whence, $\rho_i = \rho_{i+p}$;

in contradiction with the strictly descending property of $(\rho_n; n \geq 0)$. Hence, our working hypothesis cannot hold; wherefrom

$(x_n; n \geq 0)$ is a full sequence
$(i < j$ implies $x_i \neq x_j$; hence, $x_i < x_j, d(x_i, x_j) > 0)$.

Part 4 As a consequence of this, the iterative sequence $(x_n = T^n x_0; n \geq 0)$ in $X(T, \leq)$ is strictly ascending, strongly orbital and full. We now establish that $(x_n; n \geq 0)$ is d-Cauchy. Let $\varepsilon > 0$ be given; and $\delta > 0$ be assured by the Meir-Keeler property of T (modulo $(d, \leq; P)$); clearly, without loss, one may assume that $\delta < \varepsilon$. Further, given the couple (ε, δ) as before, let the number $\gamma \in]0, \delta/2[$ [and the associated semi-Cauchy rank $n(\gamma)$] be assured via P=orbitally small. We claim that, under these conditions,

$(\forall i \geq 1)$: $d(x_n, x_{n+i}) < \varepsilon + \delta/2$, for each $n \geq n(\gamma)$;

and, from this, the d-Cauchy property of (x_n) follows. To verify the assertion, an (ordinary) induction is being performed upon $i \geq 1$. The case $i \in \{1, 2\}$ is evident, via $\gamma < \delta/2$ and the very definition of our semi-Cauchy rank $n(\gamma)$. Suppose that the underlying relation holds for all $i \in \{1, \ldots, j\}$, where $j \geq 2$; we must establish its validity for $i = j + 1$:

$d(x_n, x_{n+j+1}) < \varepsilon + \delta/2$, for all $n \geq n(\gamma)$.

Suppose by contradiction that this does not hold:

$C(\varepsilon, \delta) := \{n \in N(n(\gamma), \leq); d(x_n, x_{n+j+1}) \geq \varepsilon + \delta/2\}$ is nonempty;

and let $n \in C(\varepsilon, \delta)$ be one of these ranks; for example, one may take $n = \min C(\varepsilon, \delta)$. By the choice of our data

$P(x_n, x_{n+j}) < \varepsilon + \delta$ (as P is orbitally small).

On the other hand, by the preceding step (and $P = (\leq)$-sufficient)

$x_n < x_{n+j}$ (whence, $P(x_n, x_{n+j}) > 0$).

Combining with the (variant of) Meir-Keeler property of T (modulo $(d, \leq; P)$), gives

$d(x_{n+1}, x_{n+j+1}) = d(Tx_n, Tx_{n+j}) \leq \varepsilon$;

so that, taking the triangular inequality into account,

$d(x_n, x_{n+j+1}) \leq d(x_n, x_{n+1}) + d(x_{n+1}, x_{n+j+1}) < \varepsilon + \gamma < \varepsilon + \delta/2$;

in contradiction with the choice of $n \in C(\varepsilon, \delta)$. Hence, the precise inductive relation holds; wherefrom, $(x_n; n \geq 0)$ is d-Cauchy, as claimed.

Part 5 As X is (sa-so-f,d)-complete, $x_n \xrightarrow{d} z$ for some (uniquely determined) $z \in X$. There are two cases to discuss.

Case 5a Suppose that T is (sa-so-f,d)-continuous. Then $y_n := Tx_n \xrightarrow{d} Tz$ as $n \to \infty$. On the other hand, $(y_n = x_{n+1}; n \geq 0)$ is a subsequence of $(x_n; n \geq 0)$; whence $y_n \xrightarrow{d} z$; and this yields (as d is separated), $z = Tz$.

Case 5b Suppose that $(\leq)$ is (sa-so-f,d)-almost-selfclosed. For the moment, it is clear that

$x_n \leq z$ for all $n \geq 0$.

We show that $b := d(z, Tz) > 0$ yields a contradiction.

From the d-semi-Cauchy and convergence properties one gets (taking a metrical property of $d(.,.)$ into account)

$d(x_n, z),\ d(Tx_n, z) \to 0, d(x_n, Tx_n) \to 0, d(x_n, Tz),\ d(Tx_n, Tz) \to b.$

On the other hand, by the full property of $(x_n; n \geq 0)$,

$E := \{n \in N; x_n = z\}$ is an asingleton;

so that, the following separation property holds:

(sepa) $\exists h = h(z) \geq 0$: $n \geq h \Longrightarrow x_n \neq z$ (hence, $x_n < z$).

There are several sub-cases to be analyzed.

Alter 1 Assume that P is orbitally singular asymptotic. As P is $(\leq)$-sufficient,

$P(x_n, z) > 0, \forall n \geq h.$

This tells us that the Meir-Keeler contractive condition applies to (x_n, z), $\forall n \geq h$; and yields (by the strict nonexpansive property)

$d(x_{n+1}, Tz) < P(x_n, z)$, for all $n \geq h$;

wherefrom (passing to lim inf as $n \to \infty$)

$b = \liminf_n d(x_{n+1}, Tz) \leq \liminf_n P(x_n, z).$

This, however, contradicts the orbital singular asymptotic property of P. Hence, necessarily, $b = 0$ [i.e.: $z = Tz$]; and conclusion follows.

Alter 2 Suppose that P is orbitally regular asymptotic, and T is $(d, \leq; P; \varphi)$-contractive, where $\varphi \in \mathscr{F}_0(re)(R_+)$ is bilateral Boyd-Wong admissible (hence, Meir-Keeler admissible). By the previous convergence relations concerning (x_n) we have (from the orbital regular asymptotic property)

$P(x_n, z) \to b$, as $n \to \infty$.

On the other hand, by the imposed contractive conditions,

$d(x_{n+1}, Tz) \leq \varphi(P(x_n, z))$, for all $n \geq h$.

Passing to lim sup as $n \to \infty$, one derives (by an auxiliary fact)

$b \leq \limsup_n \varphi(P(x_n, z)) \leq \Lambda_{\pm}\varphi(b) < b$; a contradiction.

Hence, necessarily, $b = 0$; i.e.: $z = Tz$.

Alter 3 Assume further that P is orbitally strongly regular asymptotic, and T is $(d, \leq; P; \varphi)$-contractive, where $\varphi \in \mathscr{F}_0(re)(R_+)$ is Meir-Keeler admissible. From the previous convergence relations concerning (x_n) we have (by the posed hypothesis upon P)

$P(x_n, z) \to\to b$, as $n \to \infty$.

According to the definition of this relation, there must be a subsequence $(u_n := x_{j(n)}; n \geq 0)$ of $(x_n; n \geq 0)$, such that (in addition)

$P(u_n, z) = b(> 0)$, for all $n \geq 0$.

Note that, as $\lim_n j(n) = \infty$, one may arrange for

$j(n) \geq h$ (hence, $u_n < z$), for all $n \geq 0$.

From the imposed contractive condition, we get

$d(Tu_n, Tz) \leq \varphi(P(u_n, z)) = \varphi(b)$, for all $n \geq 0$.

Passing to limit as $n \to \infty$ in the previous relation, it results that

$b \leq \varphi(b) < b$; a contradiction.

Hence, $b = 0$; i.e.: $z = Tz$; and conclusion follows.

Alter 4 Suppose that P is orbitally regular asymptotic, and T is $(d, \leq; P; (\psi, \varphi))$-contractive, for an admissible couple (ψ, φ) of functions in $\mathscr{F}_0(R_+)$, such that φ is sequentially positive and ψ is φ-bounded bilateral oscillating. By the previous convergence relations concerning (x_n) (see above), we have (by the orbital regular asymptotic property)

$P(x_n, z) \to b$, as $n \to \infty$.

On the other hand, by the imposed contractive conditions,

$\psi(d(x_{n+1}, Tz)) \leq \psi(P(x_n, z)) - \varphi(P(x_n, z))$, for all $n \geq h$;

or, equivalently,

$\varphi(P(x_n, z)) \leq \psi(P(x_n, z)) - \psi(d(x_{n+1}, Tz))$, for all $n \geq h$.

Passing to lim sup as $n \to \infty$ one derives

$\limsup_n \varphi(P(x_n, z)) \leq \psi(b + 0) - \psi(b - 0)$;

in contradiction with ψ being φ-bounded bilateral oscillating. Hence, necessarily, $b = 0$ (i.e.: $z = Tz$); and this establishes our claim.

Alter 5 Assume that the mapping P is orbitally strongly regular asymptotic and T is $(d, \leq; P; (\psi, \varphi))$-contractive, for an admissible couple (ψ, φ) of functions in $\mathscr{F}_0(R_+)$, such that φ is sequentially positive and ψ is φ-bounded left oscillating. From the previous convergence relations concerning (x_n) we have (by the posed hypothesis upon P)

$P(x_n, z) \to\to b$, as $n \to \infty$.

According to the definition of this relation, there must be a subsequence $(u_n := x_{j(n)}; n \geq 0)$ of $(x_n; n \geq 0)$, such that (in addition)

$P(u_n, z) = b(> 0)$, for all $n \geq 0$.

Note that, as $\lim_n j(n) = \infty$, one may arrange for

$j(n) \geq h$ (hence, $u_n < z$), for all $n \geq 0$.

From the imposed contractive conditions, one gets

$\psi(d(Tu_n, Tz)) \leq \psi(P(u_n, z)) - \varphi(P(u_n, z)) = \psi(b) - \varphi(b), \forall n \geq 0;$

or, equivalently,

$(0 <)\varphi(b) \leq \psi(b) - \psi(d(Tu_n, Tz))$, for all $n \geq 0$.

By the left part of this relation, we have (along with ψ=increasing)

$d(Tu_n, Tz) < b$, for all n; whence $d(Tu_n, Tz) \to b-$.

Passing to lim sup as $n \to \infty$ in the right part of our previous relation, yields

$\varphi(b) \leq \psi(b) - \psi(b - 0);$

in contradiction with ψ being φ-bounded left oscillating. Hence, $b = 0$ (i.e.: $z = Tz$); and conclusion follows. The proof is thereby complete.

Note that, further enlargements of these facts are possible, over quasi-metric spaces taken as in Roldán et al. [34]. On the other hand, this result admits multivalued type versions, under Nadler's model [26]; but, in this case, the setting of our problem is (ZF-AC+DC). Finally, non-sufficient versions of our main result are possible, under the lines in Choudhury and Kundu [8]. We shall discuss all these facts in a separate paper.

7 Particular Cases

Let $(X, d, \leq)$ be a quasi-ordered metric space. Further, let T be a selfmap of X; supposed to be semi-progressive and increasing. As precise, we have to determine appropriate conditions under which Fix(T) is nonempty. The specific directions under which this problem is to be solved were already listed. Sufficient conditions for getting such properties are being founded on the (almost) strictly ascending strongly orbital full concepts we just introduced. Finally, the specific contractive properties to be used have been described; and the main result incorporating all these is Function Meir-Keeler theorem (MK-f). It is our aim in the sequel to expose a certain particular case of it, with some technical relevance. To do this, remember that for each $x, y \in X$, we defined the (basic) maps

$$M_1(x, y) = d(Tx, Ty),\ M_2(x, y) = d(x, y),\ M_3(x, y) = d(x, Tx),$$
$$M_4(x, y) = d(y, Ty),\ M_5(x, y) = d(x, Ty),\ M_6(x, y) = d(Tx, y).$$

By taking elementary order/algebraic combinations between these, one gets a lot of functions to be used in our reasonings; the basic ones are

$A_1 = M_2$, $A_2 = (1/2)[M_3 + M_4]$,
$A_3 = \max\{M_3, M_4\}$, $A_4 = (1/2)[M_5 + M_6]$;
or, explicitly (for $x, y \in X$)
$A_1(x, y) = d(x, y)$, $A_2(x, y) = (1/2)[d(x, Tx) + d(y, Ty)]$,
$A_3(x, y) = \max\{d(x, Tx), d(y, Ty)\}$, $A_4(x, y) = (1/2)[d(x, Ty) + d(Tx, y)]$.

Then, by means of (further) intricate order/algebraic operations, we may define some other functions of this type; the following ones will be taken as concrete examples in our developments. Let us introduce the diagonal type subset of R_+^2

$$\Delta = \{(\xi, \eta) \in R_+ \times R_+^0; \xi \le \eta\}.$$

This set is composed of a "singular" and "regular" part, expressed as

$$\Delta_s = \{(\xi, \eta) \in \Delta; \xi < \eta\},$$
$$\Delta_r = \{(\xi, \eta) \in \Delta; \xi = \eta\} = \{(\zeta, \zeta); \zeta \in R_+^0\}.$$

For each $(\xi, \eta) \in \Delta$, let us introduce the map $B := B[\xi, \eta] : X \times X \to R_+$, as

$B = M_4(\xi + M_3)/(\eta + M_2)$; or, explicitly (for $x, y \in X$)
$B(x, y) = d(y, Ty)[\xi + d(x, Tx)]/[\eta + d(x, y)]$.

Further, let us define

B_s=one of the maps $B[\xi, \eta]$ with $(\xi, \eta) \in \Delta_s$,
B_r=one of the maps $B[\xi, \eta]$ with $(\xi, \eta) \in \Delta_r$; or, equivalently:
B_r=one of the maps $B[\zeta, \zeta]$ with $\zeta \in R_+^0$.

The reason of splitting these maps will become clear later. Finally, given $(\alpha, \beta) \in \Delta$, let us introduce the map $C := C[\alpha, \beta] : X \times X \to R_+$, according to

$C = M_6(\alpha + M_5)/(\beta + M_2)$; or, explicitly (for $x, y \in X$):
$C(x, y) = d(Tx, y)[\alpha + d(x, Ty)]/[\beta + d(x, y)]$.

Having these precise, fix in the following the couples $(\xi, \eta) \in \Delta_s$, $(\zeta, \zeta) \in \Delta_r$, $(\alpha, \beta) \in \Delta$; and (according to the previous conventions), denote

$$\mathscr{A} = \{A_1, A_2, A_3, A_4, B_s, B_r, C\},\ \mathscr{A}_1 = \{A_2, A_3, A_4, B_s, B_r, C\}.$$

For each (nonempty) subset $\Theta \in \exp(\mathscr{A})$, let $\max(\Theta) \in \mathscr{F}(X \times X, R_+)$ be the mapping defined as

$\max(\Theta)(x, y) = \max\{E(x, y); E \in \Theta\}, x, y \in X.$

Denote also

$\mathscr{A}_1^* = \{\Theta \in \exp(\mathscr{A}); A_1 \in \Theta\} = \{A_1\} \cup \exp[\mathscr{A}_1];$

clearly, there are card $\exp[\mathscr{A}_1] = 2^6 = 64$ subsets of this type. Technically speaking, the admissible maps $P : X \times X \to R_+$ to be considered are of the form

$P = \max(\Theta)$; where, $\Theta \in \mathscr{A}_1^*$.

So, it remains to establish of to what extent is this functional family compatible with the (orbital or global) normality conditions required by our main result.

I) First, as a direct consequence of this very construction, we have

Proposition 20 *All maps* $P = \max(\Theta)$ *where* $\Theta \in \mathscr{A}_1^*$ *are* ($\leq$)*-sufficient.*

Proof Evident, in view of $P \geq A_1$.

II) The next property to be checked is telescopic boundedness. A positive result in this direction is given below.

Proposition 21 *All maps* $P = \max(\Theta)$ *where* $\Theta \in \mathscr{A}_1^*$ *are telescopic bounded.*

Proof Given the arbitrary point $x \in X(T, <)$, we have

$$\begin{aligned}
&A_1(x, Tx) = d(x, Tx) \leq \max\{d(x, Tx), d(Tx, T^2x)\} = A_3(x, Tx),\\
&A_2(x, Tx) = (1/2)[d(x, Tx) + d(Tx, T^2x)] \leq\\
&\max\{d(x, Tx), d(Tx, T^2x)\} = A_3(x, Tx),\\
&A_4(x, Tx) = (1/2)d(x, T^2x)) \leq A_2(x, Tx) \leq A_3(x, Tx);\\
&B(x, Tx) = d(Tx, T^2x)[\xi + d(x, Tx)]/[\eta + d(x, Tx)] \leq\\
&d(Tx, T^2x) \leq A_3(x, Tx),\\
&C(x, Tx) = 0 \leq A_3(x, Tx);
\end{aligned}$$

and this, along with any map B_s or B_r having the form $B = B[\xi, \eta]$ where $(\xi, \eta) \in \Delta$, ends the argument.

III) Passing to the orbitally small property, we have

Proposition 22 *All maps* $P = \max(\Theta)$*, where* $\Theta \in \mathscr{A}_1^*$ *are orbitally small.*

Proof There argument consists of two steps.

Step 1 Let us first establish that all maps $Q \in \mathscr{A}$ have such a property. There are two cases to be discussed.

Case 1 $Q \in \{A_1, A_2, A_3, A_4\}$. By definition, we have for $1 \leq i \leq 4$,

$A_i(x, y) \leq M(x, y)$, for all $x, y \in X$;

whence, the maps $\{A_1, A_2, A_3, A_4\}$ are chain diametrally bounded; this, along with a previous auxiliary fact, assures us that the underlying maps are orbitally small.

Case 2 $Q \in \{B, C\}$, where

$B \in \{B_s, B_r\}$; i.e.: $B = B[\xi, \eta]$, for some $(\xi, \eta) \in \Delta$.

Let the d-semi-Cauchy (sa-so-f)-sequence $(x_n = T^n x_0; n \geq 0)$ in $X(T, \leq)$ be given, as well as the couple (ε, δ) with $\varepsilon > \delta > 0$. Further, let $\gamma \in]0, \delta/2[$ be arbitrary for the moment; and $n(\gamma)$ be the attached semi-Cauchy rank. Finally, let $j \geq 2$ and $k \geq n(\gamma)$ be such that be some index with

$d(x_m, x_{m+i}) < \varepsilon + \delta/2$ for $(m \geq k, i \in \{1, \ldots, j\})$.

Suppose now that $n \geq k$ is such that

$d(x_n, x_{n+j+1}) \geq \varepsilon + \delta/2$.

Denote, as usual, $(\rho_n = d(x_n, x_{n+1}); n \geq 0)$. By these hypotheses, we have

$d(x_n, x_{n+j+1}) \leq d(x_n, x_{n+j}) + \rho_{n+j} < \varepsilon + \delta/2 + \gamma$.

On the other hand, the triangular inequality (and our choice of $n \geq k$) give

$d(x_n, x_{n+j}) \geq d(x_n, x_{n+j+1}) - \rho_{n+j} \geq \varepsilon + \delta/2 - \gamma (> 0)$.

In this case, by definition,

$$\begin{aligned} &B(x_n, x_{n+j}) = \\ &\rho_{n+j}[\xi + \rho_n]/[\eta + d(x_n, x_{n+j})] \leq \\ &\rho_{n+j}[\xi + \rho_n]/[\eta + \varepsilon + \delta/2 - \gamma] < \gamma[\xi + \gamma]/[\eta + \varepsilon + \delta/2 - \gamma], \\ &C(x_n, x_{n+j}) = d(x_{n+1}, x_{n+j})[\alpha + d(x_n, x_{n+j+1})]/[\beta + d(x_n, x_{n+j})] < \\ &(\varepsilon + \delta/2)[\alpha + \varepsilon + \delta/2 + \gamma]/[\beta + \varepsilon + \delta/2 - \gamma]. \end{aligned}$$

Denote, for $0 < \gamma < \delta/2$,

$$\begin{aligned} &\Phi(\gamma) = \gamma[\xi + \gamma]/[\eta + \varepsilon + \delta/2 - \gamma], \\ &\Psi(\gamma) = (\varepsilon + \delta/2)[\alpha + \varepsilon + \delta/2 + \gamma]/[\beta + \varepsilon + \delta/2 - \gamma]. \end{aligned}$$

By the above evaluations, we have (for all such γ)

$B(x_n, x_{n+j}) < \Phi(\gamma), C(x_n, x_{n+j}) < \Psi(\gamma)$.

On the other hand,

$\lim_{\gamma \to 0+} \Phi(\gamma) = 0 < \varepsilon + \delta$,
$\lim_{\gamma \to 0+} \Psi(\gamma) = (\varepsilon + \delta/2)[\alpha + \varepsilon + \delta/2]/[\beta + \varepsilon + \delta/2] < \varepsilon + \delta$.

This tells us that, if $\gamma \in]0, \delta/2[$ is small enough, we have

$\Phi(\gamma) < \varepsilon + \delta$; hence, $B(x_n, x_{n+j}) < \varepsilon + \delta$,
$\Psi(\gamma) < \varepsilon + \delta$; hence, $C(x_n, x_{n+j}) < \varepsilon + \delta$.

Putting these together, one gets the desired assertion involving the class $\mathscr{A}$.

Step 2 The final conclusion relative to the maps $P = \max(\Theta)$, where $\Theta \in \mathscr{A}_1$ is now clear—by a previous auxiliary fact—via all elements in Θ being endowed with the orbitally small property.

IV) Concerning the orbital asymptotic properties, the situation is a little bit complicated. Precisely, the following synthetic answer is available.

Proposition 23 *Under the above conventions,*

(74-1) Each (admissible) map $P = \max(\Theta)$*, where* $\Theta \in \mathscr{A}_1^*$ *fulfills* $A_3, B_r \notin \Theta$ *is orbitally singular asymptotic*
(74-2) Each (admissible) map $P = \max(\Theta)$*, where* $\Theta \in \mathscr{A}_1^*$ *fulfills* $B_r \in \Theta$ *is orbitally regular asymptotic*
(74-3) Each (admissible) map $P = \max(\Theta)$*, where* $\Theta \in \mathscr{A}_1^*$ *fulfills* $A_3 \in \Theta$*,* $B_r \notin \Theta$ *is orbitally strongly regular asymptotic.*

Proof There are three steps to be passed.

Step 1 First, we have to discuss the orbital asymptotic properties of the maps $Q \in \{A_1, A_2, A_3, A_4\}$. Let the (sa-so-f)-sequence $(x_n = T^n x_0; n \geq 0)$ in $X(T, \leq)$ and the point $z \in X$ be such that

$$x_n \xrightarrow{d} z, Tx_n \xrightarrow{d} z, (x_n < z \text{ for almost all } n), \text{ and } b := d(z, Tz) > 0.$$

From these convergence properties one gets (taking a metrical property of $d(.,.)$ into account)

$$d(x_n, z),\ d(Tx_n, z) \to 0, d(x_n, Tx_n) \to 0, d(x_n, Tz),\ d(Tx_n, Tz) \to b.$$

This, by definition, gives (as $n \to \infty$)

$$A_1(x_n, z) \to 0, A_2(x_n, z) \to b/2, A_3(x_n, z) \to b, A_4(x_n, z) \to b/2;$$

whence, any map $Q \in \{A_1, A_2, A_4\}$ is orbitally singular asymptotic. Moreover, the same convergence properties of $(x_n; n \geq 0)$ tell us that, for a certain rank $n(z) \geq 0$, we must have for all $n \geq n(z)$,

$$d(x_n, z), d(Tx_n, z) < b/2, d(x_n, Tx_n) < b/2.$$

This, by the d-Lipschitz property of $d(.,.)$, gives for all $n \geq n(z)$,

$$|d(x_n, Tz) - b| \leq d(x_n, z) < b/2,$$
$$|d(Tx_n, Tz) - b| \leq d(Tx_n, z) < b/2;$$

wherefrom (for the same ranks)

$$b/2 < d(x_n, Tz), d(Tx_n, Tz) < 3b/2.$$

Combining these, yields, for all $n \geq n(z)$

$A_1(x_n, z) < b/2 < b$, $A_2(x_n, z) < 3b/4 < b$,
$A_3(x_n, z) = b$, $A_4(x_n, z) < b$;

which, in particular, tells us that

$Q = A_3$ is orbitally strongly regular asymptotic.

Step 2 Second, we discuss the orbital asymptotic properties of the maps $Q \in \{B_s, B_r, C\}$. Let the (sa-so-f)-sequence $(x_n = T^n x_0; n \geq 0)$ in $X(T, \leq)$, and the point $z \in X$ be such that

$x_n \xrightarrow{d} z$, $Tx_n \xrightarrow{d} z$, ($x_n < z$ for almost all n), and $b := d(z, Tz) > 0$.

By definition, we have (under the notation $(\rho_n := d(x_n, x_{n+1}); n \geq 0)$)

$B(x_n, z) = b[\xi + \rho_n]/[\eta + d(x_n, z)]$, $\forall n$; so, $\lim_n B(x_n, z) = b\xi/\eta$.

This, along with the above conventions, means

$\lim_n B_s(x_n, z) = b\xi/\eta < b$;
whence, $Q = B_s$ is orbitally singular asymptotic;
$\lim_n B_r(x_n, z) = b\zeta/\zeta = b$;
whence, $Q = B_r$ is orbitally regular asymptotic.

On the other hand,

$C(x_n, z) = d(x_{n+1}, z)/[\alpha + d(x_n, Tz)]/[\beta + d(x_n, z)]$, $\forall n$;
so, $\lim_n C(x_n, z) = 0 < b$;
wherefrom: $Q = C$ is orbitally singular asymptotic.

Step 3 By the above discussion, it is clear that our conclusion follows.

V) Finally, let us see what happens with the fix-bounded property.

Proposition 24 *All maps* $P = \max(\Theta)$ *where* $\Theta \in \mathscr{A}_1^*$, *are fix-bounded.*

Proof Let $x, y \in \text{Fix}(T)$ be such that $x < y$. Then (under the convention $B = B[\xi, \eta]$, where $(\xi, \eta) \in \Delta$)

$A_1(x, y) = A_4(x, y) = d(x, y) = M(x, y)$,
$A_2(x, y) = A_3(x, y) = 0 \leq d(x, y) = M(x, y)$,
$B(x, y) = 0 \leq d(x, y) = M(x, y)$,
$C(x, y) = d(x, y)[\alpha + d(x, y)]/[\beta + d(x, y)] \leq d(x, y) = M(x, y)$;

and, from this, we are done.

Now, by simply combining these with our main result, one gets the following *rational* type fixed point statement (referred to as *Rational Function Meir-Keeler theorem*; in short: (MK-f-ra)).

Theorem 3 *Assume that the selfmap T is (semi-progressive, increasing and) Meir-Keeler $(d, \leq; \max(\Theta))$-contractive, for some subset $\Theta \in \mathscr{A}_1^*$. In addition, let X be (sa-so-f,d)-complete. Then*

(71-a) T is a strong Picard operator (modulo $(d, \leq)$), provided the following extra condition holds

(exco-a1) T is (sa-so-f,d)-continuous

(71-b) T is a Bellman Picard operator (modulo $(d, \leq)$), provided $(\leq)$ is (sa-so-f,d)-selfclosed and one of the following extra conditions holds

(exco-b1) $\{A_3, B_r\}$ is disjoint from Θ
(exco-b2) $B_r \in \Theta$, and T is $(d, \leq; \max(\Theta); \varphi)$-contractive, where the function $\varphi \in \mathscr{F}_0(re)(R_+)$ is bilateral Boyd-Wong admissible (hence, Meir-Keeler admissible as well)
(exco-b3) $A_3 \in \Theta$, $B_r \notin \Theta$, and T is $(d, \leq; \max(\Theta); \varphi)$-contractive, where $\varphi \in \mathscr{F}_0(re)(R_+)$ is Meir-Keeler admissible
(exco-b4) $B_r \in \Theta$, and T is $(d, \leq; \max(\Theta); (\psi, \varphi))$-contractive, for an admissible couple (ψ, φ) of functions in $\mathscr{F}_0(R_+)$, such that φ is sequentially positive and ψ is φ-bounded bilateral oscillating
(exco-b5) $A_3 \in \Theta$, $B_r \notin \Theta$, and T is $(d, \leq; \max(\Theta); (\psi, \varphi))$-contractive, for an admissible couple (ψ, φ) of functions in $\mathscr{F}_0(R_+)$, such that φ is sequentially positive and ψ is φ-bounded left oscillating

(71-c) T is fix-$(\leq)$-asingleton (hence, fix-$(\leq)$-singleton); moreover, when (in addition) X is $(\leq)$-directed, then T is fix-asingleton (hence, fix-singleton).

Some particular cases of this result may be described as follows.

Case 1 Suppose that $(\leq)$ is the trivial quasi-order on X. Then, (MK-f-ra) our particular main result includes directly the basic statements in Boyd and Wong [5], Matkowski [21] and Leader [20].

Case 2 Suppose that $(\leq)$ is, in addition, antisymmetric; hence, a partial order on X. Then, (MK-f-ra) includes the related statements in Agarwal et al [1] when $\Theta = \{A_1, A_3, A_4\}$; and the ones in Cabrera et al [6], when $\Theta = \{A_1, B_r\}$. Further aspects may be given in Saluja et al [36].

Finally, as a particular case of Rational Function Meir-Keeler theorem (i.e.: (MK-f-ra)), we have the *linear* type fixed point statement to be used further (referred to as: *Linear Rational Meir-Keeler theorem*; in short: (MK-ra-lin)). Let the mapping $K \in \mathscr{F}(X \times X, R_+)$ be introduced as

$K = \min\{M_3(2 + M_4)/(2 + M_2), M_4(2 + M_3)/(2 + M_2)\}$;
or, explicitly (for $x, y \in X$)
$K(x, y) = \min\{K_1(x, y), K_2(x, y)\}$, where
$K_1(x, y) = d(x, Tx)(2 + d(y, Ty))/(2 + d(x, y))$,
$K_2(x, y) = d(y, Ty)(2 + d(x, Tx)/(2 + d(x, y))$.

Then, let us define the mapping $P : X \times X \to R_+$, according to

$P = \max\{A_1, A_2, A_4, K\}$; or, explicitly (for $x, y \in X$)
$P(x, y) = \max\{A_1(x, y), A_2(x, y), A_4(x, y), K(x, y)\}$.

Finally, given $\mu \geq 0$, let us say that T is $(d, \leq; P, \mu)$-contractive, provided

$d(Tx, Ty) \leq \mu P(x, y)$, for all $x, y \in X, x \leq y$.

Theorem 4 *Assume that the selfmap T is (semi-progressive, increasing as well as) $(d, \leq; P, \mu)$-contractive, for some $\mu \in [0, 1[$. In addition, let X be (sa-so-f,d)-complete. Then*

(72-a) T is a strong Picard operator (modulo $(d, \leq)$), provided T is (sa-so-f,d)-continuous
(72-b) T is a Bellman Picard operator (modulo $(d, \leq)$), provided $(\leq)$ is (sa-so-f,d)-selfclosed.
(72-c) T is fix-$(\leq)$-asingleton (hence, fix-$(\leq)$-singleton); moreover, when (in addition) X is $(\leq)$-directed, then T is fix-asingleton (hence, fix-singleton).

Proof Denote, for $x, y \in X$,

$B_r(x, y) = d(y, Ty)[2 + d(x, Tx)]/[2 + d(x, y)]$,
$Q(x, y) = \max\{A_1(x, y), A_2(x, y), A_4(x, y), B_r(x, y)\}$.

By the very definition of the mapping K, one has

$P(x, y) \leq Q(x, y)$, for all $x, y \in X$.

As a consequence of this, T is $(d, \leq; Q, \mu)$-contractive. In addition, (by the introduced conventions)

$Q = \max(\Theta)$, where $\Theta = \{A_1, A_2, A_4, B_r\}$ (so that, $B_r \in \Theta$).

Finally, the function $(\varphi(t) = \mu t; t \geq 0)$ is (regressive and) bilateral Boyd-Wong admissible; hence, Meir-Keeler admissible as well. Summing up, the Rational Function Meir-Keeler theorem (MK-f-ra) applies to our data; wherefrom, all is clear.

Finally, it is worth noting that, by the used techniques, our particular fixed point statement does not include the one in Chandok et al [7]. However, if one starts from a certain bi-dimensional refinement of its developments, this inclusion holds; we do not give details. Further aspects may be found in Harjani et al [14]; see also Yadava et al [49].

8 Coupled Fixed Points

In the following, an application of the obtained facts is given—under the lines in Samet et al [40]—to existence results involving coupled fixed points, taken as in Bhaskar and Lakshmikantham [4].

Let $(X, d, \leq)$ be a quasi-ordered metric space. Denote for simplicity $X^2 = X \times X$; and let the metric Δ over X^2 be introduced as

$\Delta(z, w) = d(x, u) + d(y, v)$, for $z = (x, y)$, $w = (u, v)$ in X^2.

Further, define a quasi-order and a conjugate map over X^2 according to:

$(x, y) \preceq (u, v)$ iff $x \leq u$, $y \geq v$; $z^* = (y, x)$, if $z = (x, y)$.

By this very convention,

(p-1) $z \mapsto z^*$ is *involutive*: $(z^*)^* = z$, $\forall z \in X^2$.

Some basic properties of the quasi-ordered metric space $(X^2, \Delta, \preceq)$ related to this conjugate map are

(p-2) for each ($z = (x, y)$, $w = (u, v)$ in X^2): $\Delta(z, w) = \Delta(z^*, w^*)$
(p-3) the conjugation map $z \mapsto z^*$ is Δ-continuous: $z_n \xrightarrow{\Delta} z$ implies $z_n^* \xrightarrow{\Delta} z^*$
(p-4) for each ($z = (x, y)$, $w = (u, v)$ in X^2): $z \preceq w$ if and only if $w^* \preceq z^*$.

Having these precise, let $F : X^2 \to X$ be a map; and

$(G : X \to X)$: $G(x) = F(x, x)$, $x \in X$

be the associated *diagonal operator*. Denote

$\text{Cfix}(F) := \{(a, b) \in X^2; a = F(a, b), b = F(b, a)\}$;

each element of it will be referred to as a *coupled fixed point* of F. The following useful properties involving such elements are available.

Proposition 25 *Under these conventions, we have*

(81-1) $c := (a, b) \in \text{Cfix}(F)$ *if and only if* $c^* := (b, a) \in \text{Cfix}(F)$
(81-2) $(a, a) \in \text{Cfix}(F)$, *if and only if* $a \in \text{Fix}(G)$
(81-3) if $\text{Cfix}(F)$ *is a singleton* $\{c = (a, b)\}$, *then* $a = b$; *hence,* $c = (a, a)$; *moreover, we have* $\text{Fix}(G) = \{a\}$.

Proof

i), ii) Evident.
iii) By the first part, $c^* = (b, a) \in \text{Cfix}(F)$; and then, $c = c^*$; wherefrom, $a = b$ and $\text{Cfix}(F) = \{(a, a)\}$; so that (by the second part), $a \in \text{Fix}(G)$. Suppose that $b \in \text{Fix}(G)$. Then, again by the second part, $(b, b) \in \text{Cfix}(F)$; so, by the above representation, $(a, a) = (b, b)$; wherefrom $a = b$. The proof is complete.

In what follows, we are primarily concerned with existence and uniqueness results for the coupled fixed points of F. But, as long as the singleton property of $\text{Cfix}(F)$ is available, we also get existence and uniqueness fixed point results for the associated diagonal operator G.

To reach this objective, the following basic construction will be considered. Given F as before, define the selfmap of X^2

$Tz = (Fz, Fz^*)$, for $z := (x, y) \in X^2$.

Clearly, the compatible type relation holds

$$Tz^* = (Tz)^*, \text{ for each } z \in X^2.$$

Moreover, it is easy to see that

$$\text{Cfix}(F) = \text{Fix}(T); \text{ i.e.: } (a, b) \in \text{Cfix}(F) \text{ iff } (a, b) \in \text{Fix}(T).$$

In other words; the coupled fixed points of $F : X \times X \to X$ are just the fixed points of $T : X^2 \to X^2$. Hence, the regularity conditions we are looking for are the ones appearing in (precise particular versions of) our main result, applied to the quasi-ordered metric space $(X^2, \Delta, \preceq)$ and the selfmap T. For technical reasons, it would be useful expressing these conditions in terms of our initial data $(X, d, \leq)$ and F. We have three groups of such requirements.

I) The first group consists of initial type conditions.

I-1) We say that F is $(\leq, \geq)$-*semi-progressive*, when

$X^2(F, \leq, \geq) := \{(a, b) \in X^2; a \leq F(a, b), b \geq F(b, a)\}$ is nonempty.

Note that, by this definition,

$(a \leq F(a, b), b \geq F(b, a))$ iff $(a, b) \preceq T(a, b)$;

so, this condition assures us that T is $(\preceq)$-semi-progressive.

I-2) Let us say that the mapping F is *mixed monotone*, provided

$(x, y) \preceq (u, v)$ implies $F(x, y) \leq F(u, v)$.

A simpler way of expressing this is as follows. Call F, *(1-increasing,2-decreasing)* if it is increasing in the first variable and decreasing in the second one:

$\forall (a, b) \in X^2$: $F(., b)$=increasing, $F(a, .)$=decreasing.

Proposition 26 *Under these conventions, we have*

(F is mixed monotone) iff (F is (1-increasing,2-decreasing)).

Proof

i) Assume that F is mixed monotone; and let $(a, b) \in X^2$ be arbitrary fixed. If $x_1 \leq x_2$, then, as $(x_1, b) \preceq (x_2, b)$, we must have, $F(x_1, b) \leq F(x_2, b)$. Likewise, taking $y_1, y_2 \in X$ with $y_1 \geq y_2$, then, as $(a, y_1) \preceq (a, y_2)$, one gets $F(a, y_1) \leq F(a, y_2)$.

ii) Assume that the function F is (1-increasing,2-decreasing); and let the points $z_1 := (x_1, y_1)$, $z_2 := (x_2, y_2)$ be such that

$z_1 \preceq z_2$; that is: $x_1 \leq x_2$, $y_1 \geq y_2$.

Then

$$F(x_1, y_1) \leq F(x_2, y_1) \leq F(x_2, y_2);$$

and this ends the argument.

Concerning the relationships with the corresponding properties of T, we have

Proposition 27 *Suppose that F is mixed monotone. Then, T is $(\preceq)$-increasing.*

Proof Let $z_1 := (x_1, y_1)$, $z_2 := (x_2, y_2)$ be such that

$z_1 \preceq z_2$; i.e.: $x_1 \leq x_2$, $y_1 \geq y_2$.

Then (by the mixed monotone property)

$$F(x_1, y_1) \leq F(x_2, y_2),\ \ F(y_1, x_1) \geq F(y_2, x_2);\ \text{ hence, } T(z_1) \preceq T(z_2);$$

and the claim follows.

II) The second group consists of global conditions relative to the structure $(X, d, \leq)$ and the mapping F.

II-1) Call the sequence $(x_n; n \geq 0)$ in X, *monotone* when it is either ascending or descending. In this case, let us say that X is *(monotone, d)-complete*, provided

each monotone d-Cauchy sequence in X is d-convergent.

It is not hard to see that, under such a condition,

X^2 is $(\preceq, \Delta)$-complete:
each $(\preceq)$-ascending Δ-Cauchy sequence.

II-2) Let us say that $(\leq, \geq)$ is *self-closed*, provided

both $(\leq)$ and $(\geq)$ are self-closed; i.e.: the d-limit of each ascending (descending) sequence is an upper (lower) bound of it.

Note that, in this case,

$(\preceq)$ is self-closed: the Δ-limit of each $(\preceq)$-ascending sequence in X^2
is an upper bound of it (modulo $(\preceq)$).

II-3) Remember that F is continuous (modulo d), when

$x_n \xrightarrow{d} x, y_n \xrightarrow{d} y$ imply $F(x_n, y_n) \xrightarrow{d} F(x, y)$.

It is not hard to see that, under such a condition, we have

T is continuous (modulo Δ): $z_n \xrightarrow{\Delta} z$ implies $Tz_n \xrightarrow{\Delta} Tz$.

II-4) Let us say that X is $(\leq, \geq)$-*directed*, when

for each $x, y \in X$, $\{x, y\}$ has upper and lower bounds (modulo $(\leq)$).

In this case, we claim that the quasi-ordered structure X^2 is $(\preceq)$-directed. In fact, given $z_1 = (x_1, y_1)$, $z_2 = (x_2, y_2)$ in X^2, an upper bound (modulo $(\preceq)$) of $\{z_1, z_2\}$ is $w = (u, v)$; where u is an upper bound of $\{x_1, x_2\}$ and v is a lower bound of $\{y_1, y_2\}$.

III) Finally, a third group of conditions concerns the contractive property. Denote, for $(x, y), (u, v) \in X^2$,

$H((x, y); (u, v)) = \min\{H_1((x, y); (u, v)), H_2((x, y); (u, v))\}$;

where, by definition,

$$H_1((x, y); (u, v)) =$$
$$d(x, F(x, y))[2 + d(u, F(u, v)) + d(v, F(v, u))]/[2 + d(x, u) + d(y, v)],$$
$$H_2((x, y); (u, v)) =$$
$$d(u, F(u, v))[2 + d(x, F(x, y)) + d(y, F(y, x))]/[2 + d(x, u) + d(y, v)].$$

Given $(\alpha, \beta, \gamma, \delta) \in R^4_+$, call F, *coupled* $(d, \leq, \geq; \alpha, \beta, \gamma, \delta)$-*contractive* provided

$$d(F(x, y), F(u, v)) \leq$$
$$(\alpha/2)[d(x, u) + d(y, v)] + \beta H((x, y); (u, v))+$$
$$(\gamma/2)[d(x, F(x, y)) + d(y, F(y, x)) + d(u, F(u, v)) + d(v, F(v, u))]+$$
$$(\delta/2)[d(x, F(u, v)) + d(y, F(v, u)) + d(u, F(x, y)) + d(v, F(y, x)],$$

for all $x, y, u, v \in X$ with $x \leq u, y \geq v$.

The following auxiliary statement will be useful for us. Let the system of maps $\{A_1, A_2, A_3, A_4\}$ in $\mathscr{F}(X^2 \times X^2, R_+)$ be introduced as in the standard metrical case; namely: for $z = (x, y) \in X^2$, $w = (u, v) \in X^2$,

$A_1(z, w) = \Delta(z, w)$, $A_2(z, w) = (1/2)[\Delta(z, Tz) + \Delta(w, Tw)]$,
$A_3(z, w) = \max\{\Delta(z, Tz), \Delta(w, Tw)\}$, $A_4(z, w) = (1/2)[\Delta(z, Tw)+\Delta(Tz, w)]$.

Further, let the maps

$K_1 : X^2 \times X^2 \to R_+, K_2 : X^2 \times X^2 \to R_+, K : X^2 \times X^2 \to R_+$

be introduced as: for $z = (x, y) \in X^2$, $w = (u, v) \in X^2$,

$K_1(z, w) = \Delta(z, Tz)[2 + \Delta(w, Tw)]/[2 + \Delta(z, w)]$,
$K_2(z, w) = \Delta(w, Tw)[2 + \Delta(z, Tz)]/[2 + \Delta(z, w)]$,
$K(z, w) = \min\{K_1(z, w), K_2(z, w)\}$.

Proposition 28 *Suppose that F is coupled $(d, \leq, \geq; \alpha, \beta, \gamma, \delta)$-contractive (see above). Then, T is $(\Delta, \preceq; L, \mu)$-contractive, where $\mu := \alpha + \beta + 2\gamma + 2\delta$; and for $z = (x, y) \in X^2$, $w = (u, v) \in X^2$,*

$L(z, w) = \max\{A_1(z, w), A_2(z, w), A_4(z, w), K(z, w))\}$.

Proof By the introduced definition (and properties of conjugate operator) one has, for $z = (x, y) \in X^2$, $w = (u, v) \in X^2$,

$H_1((x, y); (u, v)) = d(x, F(x, y))[2 + \Delta(w, Tw)]/[2 + \Delta(z, w)]$,
$H_2((x, y); (u, v)) = d(u, F(u, v))[2 + \Delta(z, Tz)]/[2 + \Delta(z, w)]$;
$H_1((y, x); (v, u)) = d(y, F(y, x))[2 + \Delta(w, Tw)]/[2 + \Delta(z, w)]$,
$H_2((y, x); (v, u)) = d(v, F(v, u))[2 + \Delta(z, Tz)]/[2 + \Delta(z, w)]$.

Combining with the immediate relation

$\min\{t_1, s_1\} + \min\{t_2, s_2\} \leq \min\{t_1 + t_2, s_1 + s_2\}, t_1, s_1, t_2, s_2 \in R$,

we derive

$$H((x, y); (u, v)) + H((y, x); (v, u)) \leq \min\{K_1((x, y); (u, v)), K_2((x, y); (u, v))\};$$

where

$$K_1((x, y); (u, v)) =$$
$$(d(x, F(x, y)) + d(y, F(y, x)))[2 + \Delta(w, Tw)]/[2 + \Delta(z, w)] =$$
$$\Delta(z, Tz)[2 + \Delta(w, Tw)]/[2 + \Delta(z, w)] = K_1(z, w),$$
$$K_2((x, y); (u, v)) =$$
$$(d(u, F(u, v) + d(v, F(v, u)))[2 + \Delta(z, Tz)]/[2 + \Delta(z, w)] =$$
$$\Delta(w, Tw)[2 + \Delta(z, Tz)]/[2 + \Delta(z, w)] = K_2(z, w).$$

On the other hand, from the contractive condition we get

$$d(F(x, y), F(u, v)) + d(F(y, x), F(v, u)) \leq$$
$$\alpha\Delta(z, w) + \beta[H((x, y); (u, v)) + H((y, x); (v, u))]+$$
$$2\gamma(1/2)[\Delta(z, Tz) + \Delta(w, Tw)] + 2\delta(1/2)[\Delta(z, Tw) + \Delta(Tz, w)],$$

for all $x, y, u, v \in X$ with $x \leq u$, $y \geq v$.

Taking the above evaluation into account, yields

$\Delta(Tz, Tw) \leq \mu L(z, w)$, for all $z, w \in X^2$, $z \preceq w$;

and we are done.

Now, by simply combining this with the Linear Rational Meir-Keeler theorem (i.e.: (MK-ra-lin)), one gets the following coupled fixed point theorem involving these data.

Theorem 5 *Assume that F is coupled $(d, \leq, \geq; \alpha, \beta, \gamma, \delta)$-contractive, for some quadruple $(\alpha, \beta, \gamma, \delta)$ in R^4_+ with $\mu := \alpha + \beta + 2\gamma + 2\delta < 1$. In addition, let (X, d) be monotone complete, X be (monotone, d)-complete and $(\leq, \geq)$-directed, F be $(\leq, \geq)$-semi-progressive, mixed monotone, and one of the extra conditions below is holding*

(ext-1) *F is d-continuous*
(ext-2) *$(\leq, \geq)$ is self-closed.*

Then

(81-a) *F has a unique coupled fixed point, (a, a) with $a \in X$;*
(81-b) *the associated diagonal operator $[G(x) = F(x, x)$, $x \in X]$ admits this $a \in X$ as its unique fixed point (in X)*
(81-c) *for each $(x_0, y_0) \in X^2$ with $[x_0 \leq F(x_0, y_0)$, $y_0 \geq F(y_0, x_0)]$, the iterative process $(x_{n+1} = F(x_n, y_n), y_{n+1} = F(y_n, x_n); n \geq 0)$, Δ-converges towards (a, a); whence, $x_n \xrightarrow{d} a$, $y_n \xrightarrow{d} a$.*

Proof By the quoted fixed point result, T is $(\Delta, \preceq; L, \mu)$-contractive. This, along with the above remarks, gives us all desired conclusions.

The obtained coupled fixed point result extends a related one in Nashine and Kadelburg [27]; which, in turn, refines the statement in Samet and Yazidi [39]. Further aspects may be found in Pathak et al [30].

References

1. R.P. Agarwal, M.A. El-Gebeily, D. O'Regan, Generalized contractions in partially ordered metric spaces. Appl. Anal. **87**, 109–116 (2008)
2. S. Banach, Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. Fund. Math. **3**, 133–181 (1922)
3. P. Bernays, A system of axiomatic set theory: Part III. Infinity and enumerability analysis. J. Symb. Log. **7**, 65–89 (1942)
4. T.G. Bhaskar, V. Lakshmikantham, Fixed point theorems in partially ordered metric spaces and applications. Nonlinear Anal. **65**, 1379–1393 (2006)
5. D.W. Boyd, J.S.W. Wong, On nonlinear contractions. Proc. Am. Math. Soc. **20**, 458–464 (1969)
6. I. Cabrera, J. Harjani, K. Sadarangani, A fixed point theorem for contractions of rational type in partially ordered metric spaces. Ann. Univ. Ferrara **59**, 251–258 (2013)

7. S. Chandok, B.S. Choudhury, N. Metiya, Fixed point results in ordered metric spaces for rational type expressions with auxiliary functions. J. Egypt. Math. Soc. **23**, 95–101 (2015)
8. B.S. Choudhury, A. Kundu, A Kannan-like contraction in partially ordered spaces. Demonstr. Math. **46**, 327–334 (2013)
9. L.B. Cirić, A new fixed-point theorem for contractive mappings. Publ. Inst. Math. **30**(44), 25–27 (1981)
10. P.J. Cohen, *Set Theory and the Continuum Hypothesis* (Benjamin, New York, 1966)
11. P. Collaco, J.C.E Silva, A complete comparison of 25 contractive definitions. Nonlinear Anal. **30**, 441–476 (1997)
12. P.N. Dutta, B.S. Choudhury, A generalisation of contraction principle in metric spaces. Fixed Point Theory Appl. **2008**, 406368 (2008)
13. M.A. Geraghty, On contractive mappings. Proc. Am. Math. Soc. **40**, 604–608 (1973)
14. J. Harjani, B. Lopez, K. Sadarangani, A fixed point theorem for mappings satisfying a contractive condition of rational type on a partially ordered metric space. Abstr. Appl. Anal. **2010**, 190701 (2010)
15. P. Hitzler, Generalized metrics and topology in logic programming semantics. PhD Thesis, National University of Ireland, University College Cork, 2001
16. J. Jachymski, Common fixed point theorems for some families of mappings. Indian J. Pure Appl. Math. **25**, 925–937 (1994)
17. J. Jachymski, Equivalence of some contractivity properties over metrical structures. Proc. Am. Math. Soc. **125**, 2327–2335 (1997)
18. J. Jachymski, The contraction principle for mappings on a metric space with a graph. Proc. Am. Math. Soc. **136**, 1359–1373 (2008)
19. S. Kasahara, On some generalizations of the Banach contraction theorem. Publ. Res. Inst. Math. Sci. Kyoto Univ. **12**, 427–437 (1976)
20. S. Leader, Fixed points for general contractions in metric spaces. Math. Jpn. **24**, 17–24 (1979)
21. J. Matkowski, Integrable solutions of functional equations. *Dissertationes Mathematicae*, vol. 127 (Polish Scientific Publishers, Warsaw, 1975)
22. J. Matkowski, Fixed point theorems for contractive mappings in metric spaces. Časopis Pest. Mat. **105**, 341–344 (1980)
23. A. Meir, E. Keeler, A theorem on contraction mappings. J. Math. Anal. Appl. **28**, 326–329 (1969)
24. G.H. Moore, *Zermelo's Axiom of Choice: Its Origin, Development and Influence* (Springer, New York, 1982)
25. Y. Moskhovakis, *Notes on Set Theory* (Springer, New York, 2006)
26. S.B. Nadler Jr., Multi-valued contraction mappings. Pac. J. Math. **30**, 475–488 (1969)
27. H.K. Nashine, Z. Kadelburg, Partially ordered metric spaces, rational contractive expressions, and coupled fixed points. Nonlinear Funct. Anal. Appl. **17**, 471–489 (2012)
28. I.P. Natanson, *Theory of Functions of a Real Variable*, vol. I (Frederick Ungar Publishing, New York, 1964)
29. J.J. Nieto, R. Rodriguez-Lopez, Contractive mapping theorems in partially ordered sets and applications to ordinary differential equations. Order **22**, 223–239 (2005)
30. R.P. Pathak, R. Tiwari, R. Bhardwaj, Fixed point theorems through rational expressions in altering distance functions. Math. Theory Model. **4**, 78–83 (2014)
31. A.C.M. Ran, M.C. Reurings, A fixed point theorem in partially ordered sets and some applications to matrix equations. Proc. Am. Math. Soc. **132**, 1435–1443 (2004)
32. S. Reich, Fixed points of contractive functions. Boll. Un. Mat. Ital. **5**, 26–42 (1972)
33. B.E. Rhoades, A comparison of various definitions of contractive mappings. Trans. Am. Math. Soc. **226**, 257–290 (1977)
34. A.-F. Roldán, E. Karapinar, C. Roldán, J. Martinez-Moreno, Coincidence point theorems on metric spaces via simulation functions. J. Comput. Appl. Math. **275**, 345–355 (2015)
35. I.A. Rus, *Generalized Contractions and Applications* (Cluj University Press, Cluj-Napoca, 2001)

36. A.S. Saluja, R.A. Rashwan, D. Magarde, P.K. Jhade, Some result in ordered metric spaces for rational type expressions. Facta Univ. Niš (Ser. Math. Inform.) **31**, 125–138 (2016)
37. B. Samet, Coupled fixed point theorems for a generalized Meir-Keeler contraction in partially ordered metric spaces. Nonlinear Anal. **72**, 4508–4517 (2010)
38. B. Samet, M. Turinici, Fixed point theorems on a metric space endowed with an arbitrary binary relation and applications. Commun. Math. Anal. **13**, 82–97 (2012)
39. B. Samet, H. Yazidi, Coupled fixed point theorems for contraction involving rational expressions in partially ordered metric spaces (2010). arXiv:1005-3142-v1
40. B. Samet, E. Karapinar, H. Aydi, V. Ćojbašić Rajić, Discussion on some coupled fixed point theorems. Fixed Point Theory Appl. **2013**, 50 (2013)
41. E. Schechter, *Handbook of Analysis and its Foundation* (Academic Press, New York, 1997)
42. A. Tarski, Axiomatic and algebraic aspects of two theorems on sums of cardinals. Fund. Math. **35**, 79–104 (1948)
43. M. Turinici, Fixed points of implicit contraction mappings. An. Şt. Univ. "Al. I. Cuza" Iaşi (S I-a, Mat) **22**, 177–180 (1976)
44. M. Turinici, Fixed points for monotone iteratively local contractions. Dem. Math. **19**, 171–180 (1986)
45. M. Turinici, Abstract comparison principles and multivariable Gronwall-Bellman inequalities. J. Math. Anal. Appl. **117**, 100–127 (1986)
46. M. Turinici, Function pseudometric VP and applications. Bul. Inst. Polit. Iaşi (Sect.: Mat., Mec. Teor., Fiz.) **53**(57), 393–411 (2007)
47. M. Turinici, Ran-Reurings theorems in ordered metric spaces. J. Indian Math. Soc. **78**, 207–214 (2011)
48. E.S. Wolk, On the principle of dependent choices and some forms of Zorn's lemma. Can. Math. Bull. **26**, 365–367 (1983)
49. R.N. Yadava, R. Shrivastava, S.S. Yadav, Rational type contraction mapping in T-orbitally complete metric space. Math. Theory Model. **4**, 115–133 (2014)

A Multiple Hilbert-Type Integral Inequality in the Whole Space

Bicheng Yang

1 Introduction

If $p > 1, \frac{1}{p} + \frac{1}{q} = 1, f(\geq 0) \in L^p(\mathbf{R}_+), g(\geq 0) \in L^q(\mathbf{R}_+), ||f||_p, ||g||_q > 0$, then we have the following equivalent Hardy-Hilbert's integral inequalities (cf. [1]):

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{x+y} dxdy < \frac{\pi}{\sin(\pi/p)} ||f||_p ||g||_q, \tag{1}$$

$$\left[\int_0^\infty \left(\int_0^\infty \frac{f(x)}{x+y} dx\right)^p dy\right]^{\frac{1}{p}} < \frac{\pi}{\sin(\pi/p)} ||f||_p, \tag{2}$$

where, the constant factor $\frac{\pi}{\sin(\pi/p)}$ is the best possible. Inequality (1)–(2) are important in analysis and its applications (cf. [2]).

In 2002, [3] considered the property of Hilbert's integral operator and gave an improvement of (1) (for $p = q = 2$). In 2004, by introducing another pair of conjugate exponents (r, s) $(r > 1, \frac{1}{r} + \frac{1}{s} = 1)$ and an independent parameter $\lambda > 0$, Yang [4] gave a best extensions of (1) as follows:

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{x^\lambda + y^\lambda} dxdy < \frac{\pi}{\lambda \sin(\pi/r)} ||f||_{p,\phi} ||g||_{q,\psi}, \tag{3}$$

B. Yang (✉)
Department of Mathematics, Guangdong University of Education, Guangzhou, Guangdong, People's Republic of China
e-mail: bcyang@gdei.edu.cn

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization
and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_27

where, $\phi(x) = x^{p(1-\frac{\lambda}{r})-1}$, $\psi(x) = x^{q(1-\frac{\lambda}{s})-1}$, $||f||_{p,\phi} = (\int_0^\infty \phi(x) f^p(x)dx)^{\frac{1}{p}} > 0$, and $||g||_{q,\psi} > 0$. In 2007, Yang [5] gave the following inequality with the non-homogeneous kernel and the best possible constant factor $B(\frac{\lambda}{2}, \frac{\lambda}{2})(\lambda > 0;\ B(u, v)$ is the beta function):

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{(1+xy)^\lambda} dxdy$$
$$< B(\frac{\lambda}{2}, \frac{\lambda}{2}) \left(\int_0^\infty x^{1-\lambda} f^2(x)dx \int_0^\infty x^{1-\lambda} g^2(x)dx \right)^{\frac{1}{2}}. \tag{4}$$

In recent years, [6] gave another extension of (4) to the general kernel $k_\lambda(1, xy)$ $(\lambda > 0)$ with one pair of conjugate exponents (p, q). Some other kind of Hilbert-type inequalities and operators are provided by Milovanovic and Rassias [7], Huang [8], Krnić and Pečarić [9], Milovanovic and Rassias [10].

Definition 1 If $n \in \mathbf{N} = \{1, 2, \cdots\}$,

$$\mathbf{R}_+^n := \{(x_1, \cdots, x_n) | x_i \in \mathbf{R}_+ = (0, \infty)\ (i = 1, \cdots, n)\},$$

$\lambda \in \mathbf{R} = (-\infty, \infty)$, $k_\lambda(x_1, \cdots, x_n)$ is a measurable function in $\mathbf{R}_+^n$, such that for any $u > 0$ and $(x_1, \cdots, x_n) \in \mathbf{R}_+^n$,

$$k_\lambda(ux_1, \cdots, ux_n) = u^{-\lambda} k_\lambda(x_1, \cdots, x_n),$$

then we call $k_\lambda(x_1, \cdots, x_n)$ the homogeneous function of degree $-\lambda$ in $\mathbf{R}_+^n$.

In 2009, [6] obtained the following multiple Hilbert-type integral inequality: Suppose that $n \in \mathbf{N} \backslash \{1\}$, $p_i > 1$, $\sum_{i=1}^n \frac{1}{p_i} = 1$, $\lambda > 0$, $k_\lambda(x_1, \cdots, x_n)\ (\geq 0)$ is a homogeneous function of degree $-\lambda$ in $\mathbf{R}_+^n$, such that for any $(r_1, \cdots, r_n)$ $(r_i > 1)$, satisfying $\sum_{i=1}^n \frac{1}{r_i} = 1$, and

$$k_\lambda = \int_{\mathbf{R}_+^{n-1}} k_\lambda(u_1, \cdots, u_{n-1}, 1) \prod_{j=1}^{n-1} u_j^{\frac{\lambda}{r_j}-1} du_1 \cdots du_{n-1} \in \mathbf{R}_+.$$

If $\phi_i(x) = x^{p_i(1-\frac{\lambda}{r_i})-1}$, $f_i\ (\geq 0) \in L_{\phi_i}^{p_i}(0, \infty)$, $||f||_{p_{i,}\phi_i} > 0$ $(i = 1, \cdots, n)$, then we have the following inequality:

$$\int_{\mathbf{R}_+^n} k_\lambda(x_1, \cdots, x_n) \prod_{i=1}^n f_i(x_i) dx_1 \cdots dx_n < k_\lambda \prod_{i=1}^n ||f_i||_{p_i, \phi_i}, \tag{5}$$

where, the constant factor k_λ is the best possible.

For $n = 2, k_\lambda(x, y) = \frac{1}{x^\lambda+y^\lambda}$, inequality (5) reduces to (3); for $\lambda = n - 1, r_i = \frac{(n-1)p_i}{p_i-1}$ $(i = 1, \cdots, n)$, (5) reduces to the following multiple Hardy-Hilbert-type integral inequality (cf. [1]):

$$\int_{\mathbf{R}_+^n} k_{n-1}(x_1, \cdots, x_n) \prod_{i=1}^{n} f_i(x_i) dx_1 \cdots dx_n < k_{n-1} \prod_{i=1}^{n} ||f_i||_{p_i}. \tag{6}$$

Recently, [11] also studied the corresponding multiple Hardy-Hilbert-type integral operator. Inequality (5) are some extensions of the results [12–16].

In this paper, by introducing some interval variables and using the weight functions and the way of real analysis, a multiple Hilbert-type integral inequality in the whole space with a best possible constant factor is given, which is an extension of (5). The equivalent forms, the operator expressions with the norm, the equivalent reverses, a few particular cases and some examples with the particular kernels are also considered.

2 Some Lemmas

In the following, we make appointment that $n \in \mathbf{N}\backslash\{1\}, \alpha_i \in (0, \pi)$,

$$u_i(x) := |x| + x \cos\alpha_i \; (x \in \mathbf{R} = (-\infty \, \infty)),$$

$\delta_i \in \{-1, 1\},\ p_i \in \mathbf{R}\backslash\{0, 1\}, \lambda_i \in \mathbf{R}$ $(i = 1, \cdots, n)$, $\sum_{i=1}^{n-1} \frac{1}{p_i} = 1 - \frac{1}{p_n} = \frac{1}{q_n}$, $\sum_{i=1}^{n-1} \lambda_i = \lambda_n = \frac{\lambda}{2}$, $k_\lambda(x_1, \cdots, x_n)$ (≥ 0) is a homogeneous function of degree $-\lambda$ in $\mathbf{R}_+^n$.

Lemma 1 (cf. [15]) *For $u_i > 0$, we have*

$$A := \prod_{i=1}^{n} \left[u_i^{(\delta_i\lambda_i-1)(1-p_i)} \prod_{j=1(j\neq i)}^{n} u_j^{\delta_j\lambda_j-1} \right]^{\frac{1}{p_i}} = 1. \tag{7}$$

Lemma 2 *If we define*

$$H(i) := \int_{\mathbf{R}_+^{n-1}} k_\lambda(u_1, \cdots, u_{i-1}, 1, u_{i+1}, \cdots, u_n) \times \prod_{j=1(j\neq i)}^{n} u_j^{\lambda_j-1} du_1 \cdots du_{i-1} du_{i+1} \cdots du_n \; (i = 1, \cdots, n),$$

satisfying

$$k_\lambda := H(n) = \int_{\mathbf{R}_+^{n-1}} k_\lambda(u_1, \cdots, u_{n-1}, 1) \prod_{j=1}^{n-1} u_j^{\lambda_j - 1} du_1 \cdots du_{n-1} \in \mathbf{R}, \qquad (8)$$

then, each $H(i) = H(n) = k_\lambda$*, and for* $i = 1, \cdots, n$*, we have*

$$\omega_i(x_i) := u_i^{\delta_i \lambda_i}(x_i) \int_{\mathbf{R}^{n-1}} k_\lambda(u_1^{\delta_1}(x_1) u_n^{\delta_n}(x_n), \cdots, u_{n-1}^{\delta_{n-1}}(x_{n-1}) u_n^{\delta_n}(x_n), 1)$$

$$\times \prod_{j=1(j \neq i)}^{n} u_j^{\delta_j \lambda_j - 1}(x_j) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n$$

$$= 2^{n-1} k_\lambda \prod_{j=1(j \neq i)}^{n} \csc^2 \alpha_j. \qquad (9)$$

Proof Setting $u_j = u_n v_j \ (j \neq i, n)$ in the integral of $H(i)$, we find

$$H(i) = \int_{\mathbf{R}_+^{n-1}} k_\lambda(v_1, \cdots, v_{i-1}, u_n^{-1}, v_{i+1}, \cdots, v_{n-1}, 1) \prod_{j=1(j \neq i)}^{n-1} v_j^{\lambda_j - 1}$$

$$\times u_n^{-1-\lambda_i} dv_1 \cdots dv_{i-1} dv_{i+1} \cdots dv_{n-1} du_n.$$

Setting $v_i = u_n^{-1}$ in the above integral, we obtain $H(i) = H(n) = k_\lambda$.

Since $u_j(x_j) = x_j(sgn(x_j) + \cos\alpha_j)$, setting $v_j = u_j(x_j)$ $(j \neq i)$ in the integral of (9), we find

$$dv_j = du_j(x_j) = \begin{cases} (1 + \cos\alpha_j) dx_j, x_j > 0, \\ (-1 + \cos\alpha_j) dx_j, x_j < 0, \end{cases}$$

and

$$\omega_i(x_i) = u_i^{\delta_i \lambda_i}(x_i)$$

$$\times \int_{\mathbf{R}_+^{n-1}} k_\lambda(v_1^{\delta_1} v_n^{\delta_n}, \cdots, v_{i-1}^{\delta_{i-1}} v_n^{\delta_n}, u_i^{\delta_i}(x_i) v_n^{\delta_n}, v_{i+1}^{\delta_{i+1}} v_n^{\delta_n}, \cdots, v_{n-1}^{\delta_{n-1}} v_n^{\delta_n}, 1)$$

$$\times \prod_{j=1(j \neq i)}^{n} v_j^{\delta_j \lambda_j - 1} \left(\frac{1}{1 - \cos\alpha_j} + \frac{1}{1 + \cos\alpha_j} \right) dv_1 \cdots dv_{i-1} dv_{i+1} \cdots dv_n$$

$$= 2^{n-1} W_i(x_i) \prod_{j=1(j \neq i)}^{n} \csc^2 \alpha_j, \qquad (10)$$

where,

$$\begin{aligned}
&W_i(x_i) := u_i^{\delta_i\lambda_i}(x_i)\\
&\times \int_{\mathbf{R}_+^{n-1}} k_\lambda(v_1^{\delta_1} v_n^{\delta_n}, \cdots, v_{i-1}^{\delta_{i-1}} v_n^{\delta_n}, u_i^{\delta_i}(x_i) v_n^{\delta_n}, v_{i+1}^{\delta_{i+1}} v_n^{\delta_n}, \cdots, v_{n-1}^{\delta_{n-1}} v_n^{\delta_n}, 1)\\
&\times \prod_{j=1 (j\neq i)}^{n} v_j^{\delta_j\lambda_j-1} dv_1 \cdots dv_{i-1} dv_{i+1} \cdots dv_n.
\end{aligned}$$

Since $\lambda - \lambda_n = \lambda_n$, we find

$$\begin{aligned}
&\varpi_i(x_i) = u_i^{\delta_i\lambda_i}(x_i)\\
&\times \int_{\mathbf{R}_+^{n-1}} k_\lambda(v_1^{\delta_1}, \cdots, v_{i-1}^{\delta_{i-1}}, u_i^{\delta_i}(x_i), v_{i+1}^{\delta_{i+1}}, \cdots, v_{n-1}^{\delta_{n-1}}, v_n^{-\delta_n}) v_n^{-\delta_n\lambda_n-1}\\
&\times \prod_{j=1 (j\neq i)}^{n-1} v_j^{\delta_j\lambda_j-1} dv_1 \cdots dv_{i-1} dv_{i+1} \cdots dv_{n-1} dv_n.
\end{aligned}$$

Setting $y_n = v_n^{-1}$ in the above expression, we obtain

$$\begin{aligned}
&W_i(x_i) = u_i^{\delta_i\lambda_i}(x_i)\\
&\times \int_{\mathbf{R}_+^{n-1}} k_\lambda(v_1^{\delta_1}, \cdots, v_{i-1}^{\delta_{i-1}}, u_i^{\delta_i}(x_i), v_{i+1}^{\delta_{i+1}}, \cdots, v_{n-1}^{\delta_{n-1}}, y_n^{\delta_n}) y_n^{\delta_n\lambda_n+1}\\
&\times \prod_{j=1 (j\neq i)}^{n-1} v_j^{\delta_j\lambda_j-1} dv_1 \cdots dv_{i-1} dv_{i+1} \cdots dv_{n-1} (y_n^{-2}) dy_n\\
&= u_i^{\delta_i\lambda_i}(x_i) \int_{\mathbf{R}_+^{n-1}} k_\lambda(v_1^{\delta_1}, \cdots, v_{i-1}^{\delta_{i-1}}, u_i^{\delta_i}(x_i), v_{i+1}^{\delta_{i+1}}, \cdots, v_n^{\delta_n})\\
&\quad \times \prod_{j=1 (j\neq i)}^{n} v_j^{\delta_j\lambda_j-1} dv_1 \cdots dv_{i-1} dv_{i+1} \cdots dv_n.
\end{aligned}$$

Setting $u_j = u_i^{-\delta_i}(x_i) v_j^{\delta_j}$ $(j \neq i)$ in the above integral, we find

$$\begin{aligned}
&W_i(x_i) = u_i^{\delta_i\lambda_i}(x_i)\\
&\times \int_{\mathbf{R}_+^{n-1}} k_\lambda(u_i^{\delta_i}(x_i) u_1, \cdots, u_i^{\delta_i}(x_i) u_{i-1}, u_i^{\delta_i}(x_i), u_i^{\delta_i}(x_i) u_{i+1}, \cdots, u_i^{\delta_i}(x_i) u_n)
\end{aligned}$$

$$\times \prod_{j=1(j\neq i)}^{n} (u_i^{\delta_i/\delta_j}(x_i)u_j^{1/\delta_j})^{\delta_j\lambda_j-1}u_i^{\delta_i/\delta_j}(x_i)u_j^{(1/\delta_j)-1}du_1\cdots du_{i-1}du_{i+1}\cdots du_n$$
$$= H(i) = H(n) = k_\lambda.$$

Hence, by (10), expression (9) follows.

The lemma is proved.

Lemma 3 (cf. [15]) *The expression*

$$k(\widetilde{\lambda}_1, \cdots, \widetilde{\lambda}_{n-1}) := \int_{\mathbf{R}_+^{n-1}} k_\lambda(u_1, \cdots, u_{n-1}, 1)\prod_{j=1}^{n-1} u_j^{\widetilde{\lambda}_j-1}du_1\cdots du_{n-1}$$

is finite in a neighborhood of $(\lambda_1, \cdots, \lambda_{n-1})$ *if any only if* $k(\widetilde{\lambda}_1, \cdots, \widetilde{\lambda}_{n-1})$ *is continuous at* $(\lambda_1, \cdots, \lambda_{n-1})$.

Lemma 4 *We define the sets*

$$E_i := \{x \in \mathbf{R}; u_i^{\delta_i}(x_i) \geq 1\}\, (i = 1, \cdots, n).$$

If there exists a $\eta > 0$, *such that for* $\max_{1\leq i\leq n-1}\{|\eta_i|\} < \eta$, $k(\lambda_1 + \eta_1, \cdots, \lambda_{n-1} + \eta_{n-1}) \in \mathbf{R}$, $p_i \in \mathbf{R}\backslash\{0, 1\}(i = 1, \cdots, n)$, $0 < \varepsilon < \eta\min_{1\leq i\leq n}\{|p_i|\}$, *then*

$$\begin{aligned} I_\varepsilon \ : &= \varepsilon\int_{E_{n-1}}\cdots\int_{E_1}\left[\int_{\mathbf{R}\backslash E_n} u_n^{\delta_n(\lambda_n+\frac{\varepsilon}{p_n})-1}(x_n)\right. \\ &\left.\times k_\lambda(u_1^{\delta_1}(x_1)u_n^{\delta_n}(x_n), \cdots, u_n^{\delta_{n-1}}(x_{n-1})u_n^{\delta_n}(x_n), 1)dx_n\right] \\ &\times\prod_{j=1}^{n-1} u_j^{\delta_j(\lambda_j-\frac{\varepsilon}{p_j})-1}(x_j)dx_1\cdots dx_{n-1} = 2^n k_\lambda\prod_{j=1}^{n}\csc^2\alpha_j + o(1)\,(\varepsilon\to 0^+). \end{aligned} \tag{11}$$

Proof Setting $y_n = u_n^{-1}(x_n)$ in the integral of (11), we find

$$\begin{aligned} I_\varepsilon &= 2\varepsilon\csc^2\alpha_n\int_{E_{n-1}}\cdots\int_{E_1} \\ &\times\left[\int_{E_n} y_n^{-\delta_n(\lambda_n+\frac{\varepsilon}{p_n})-1}k_\lambda(u_1^{\delta_1}(x_1)y_n^{-\delta_n}, \cdots, u_{n-1}^{\delta_{n-1}}(x_{n-1})y_n^{-\delta_n}, 1)dy_n\right] \\ &\times\prod_{j=1}^{n-1} u_j^{\delta_j(\lambda_j-\frac{\varepsilon}{p_j})-1}(x_j)dx_1\cdots dx_{n-1}. \end{aligned}$$

Setting $u_j = u_j^{\delta_j}(x_j)y_n^{-\delta_n}$ $(j = 1, \cdots, n-1)$ in the above integral, since $\lambda - \lambda_n = \lambda_n$, by (9), we find

$$\begin{aligned} I_\varepsilon &= \varepsilon 2^n \prod_{j=1}^{n} \csc^2 \alpha_j \int_{E_n} y_n^{-1-\delta_n \varepsilon} \left[\int_{y_n^{-\delta_n}}^{\infty} \cdots \int_{y_n^{-\delta_n}}^{\infty} k_\lambda(u_1, \cdots, u_{n-1}, 1) \right. \\ &\quad \left. \times \prod_{j=1}^{n-1} u_j^{(\lambda_j - \frac{\varepsilon}{p_j})-1} du_1 \cdots du_{n-1} \right] dy_n \\ &= \varepsilon 2^n \prod_{j=1}^{n} \csc^2 \alpha_j \int_1^\infty x_n^{-1-\varepsilon} \left[\int_{x_n^{-1}}^{\infty} \cdots \int_{x_n^{-1}}^{\infty} k_\lambda(u_1, \cdots, u_{n-1}, 1) \right. \\ &\quad \left. \times \prod_{j=1}^{n-1} u_j^{(\lambda_j - \frac{\varepsilon}{p_j})-1} du_1 \cdots du_{n-1} \right] dx_n \, (x_n = y_n^{\delta_n}). \end{aligned} \tag{12}$$

Setting $D_j := \{(u_1, \cdots, u_{n-1}) | u_j \in (0, x_n^{-1}), u_k \in (0, \infty) \, (k \neq j)\}$ and

$$A_j(x_n) := \int \cdots \int_{D_j} k_\lambda(u_1, \cdots, u_{n-1}, 1) \prod_{j=1}^{n-1} u_j^{\lambda_j - \frac{\varepsilon}{p_j} - 1} du_1 \cdots du_{n-1},$$

then by (12), it follows that

$$\begin{aligned} I_\varepsilon &\geq 2^n \prod_{j=1}^{n} \csc^2 \alpha_j \int_{\mathbf{R}_+^{n-1}} k_\lambda(u_1, \cdots, u_{n-1}, 1) \prod_{j=1}^{n-1} u_j^{(\lambda_j - \frac{\varepsilon}{p_j})-1} du_1 \cdots du_{n-1} \\ &\quad - \varepsilon 2^n \prod_{j=1}^{n} \csc^2 \alpha_j \sum_{j=1}^{n-1} \int_1^\infty x_n^{-1} A_j(x_n) dx_n. \end{aligned} \tag{13}$$

Without loss of generality, in the following, we estimate the case that $j = n$, namely,

$$\int_1^\infty x_n^{-1} A_{n-1}(x_n) dx_n = O(1).$$

In fact, setting $\alpha > 0$, such that $|\frac{\varepsilon}{p_{n-1}} + \alpha| < \eta$, since

$$-u_{n-1}^{\alpha} \ln u_{n-1} \to 0 \, (u_{n-1} \to 0^+),$$

there exists a constant $M > 0$, such that

$$-u_{n-1}^{\alpha} \ln u_{n-1} \le M \, (u_{n-1} \in (0, 1]),$$

and then by Fubini theorem (cf. [17]), it follows that

$$\begin{aligned}
0 &\le \int_1^{\infty} x_n^{-1} A_{n-1}(x_n) dx_n = \int_1^{\infty} x_n^{-1} \Biggl(\int_{\mathbf{R}_+^{n-2}} \int_0^{x_n^{-1}} k_\lambda(u_1, \cdots, u_{n-1}, 1) \\
&\quad \times \prod_{j=1}^{n-1} u_j^{\lambda_j - \frac{\varepsilon}{p_j} - 1} du_{n-1} du_1 \cdots du_{n-2} \Biggr) dx_n \\
&= \int_0^1 \int_{\mathbf{R}_+^{n-2}} k_\lambda(u_1, \cdots, u_{n-1}, 1) \prod_{j=1}^{n-1} u_j^{\lambda_j - \frac{\varepsilon}{p_j} - 1} \left(\int_1^{u_{n-1}^{-1}} \frac{dx_n}{x_n} \right) du_1 \cdots du_{n-1} \\
&= \int_0^1 \int_{\mathbf{R}_+^{n-2}} k_\lambda(u_1, \cdots, u_{n-1}, 1) \prod_{j=1}^{n-1} u_j^{\lambda_j - \frac{\varepsilon}{p_j} - 1} (-\ln u_{n-1}) du_1 \cdots du_{n-1} \\
&\le M \int_0^1 \int_{\mathbf{R}_+^{n-2}} k_\lambda(u_1, \cdots, u_{n-1}, 1) \\
&\quad \times \prod_{j=1}^{n-2} u_j^{\lambda_j - \frac{\varepsilon}{p_j} - 1} u_{n-1}^{\lambda_{n-1} - (\frac{\varepsilon}{p_{n-1}} + \alpha) - 1} du_1 \cdots du_{n-1} \\
&\le M \int_{\mathbf{R}_+^{n-1}} k_\lambda(u_1, \cdots, u_{n-1}, 1) \prod_{j=1}^{n-2} u_j^{\lambda_j - \frac{\varepsilon}{p_j} - 1} u_{n-1}^{\lambda_{n-1} - (\frac{\varepsilon}{p_{n-1}} + \alpha) - 1} du_1 \cdots du_{n-1} \\
&= M \cdot k \left(\lambda_1 - \frac{\varepsilon}{p_1}, \cdots, \lambda_{n-2} - \frac{\varepsilon}{p_{n-2}}, \lambda_{n-1} - (\frac{\varepsilon}{p_{n-1}} + \alpha) \right) < \infty.
\end{aligned}$$

Hence by (13), we have

$$\begin{aligned}
I_\varepsilon &\ge 2^n \prod_{j=1}^{n} \csc^2 \alpha_j \\
&\quad \times \int_{\mathbf{R}_+^{n-1}} k_\lambda(u_1, \cdots, u_{n-1}, 1) \prod_{j=1}^{n-1} u_j^{\lambda_j - \frac{\varepsilon}{p_j} - 1} du_1 \cdots du_{n-1} - o_1(1) \\
&= 2^n k_\lambda \prod_{j=1}^{n} \csc^2 \alpha_j - o(1) \, (\varepsilon \to 0^+).
\end{aligned}$$

Since by Lemma 3, we obtain

$$I_\varepsilon \leq \varepsilon 2^n \prod_{j=1}^{n} \csc^2 \alpha_j \int_1^\infty x_n^{-1-\varepsilon} \left(\int_0^\infty \cdots \int_0^\infty k_\lambda(u_1, \cdots, u_{n-1}, 1) \right.$$

$$\left. \times \prod_{j=1}^{n-1} u_j^{\lambda_j - \frac{\varepsilon}{p_j} - 1} du_1 \cdots du_{n-1} \right) dx_n$$

$$= 2^n \prod_{j=1}^{n} \csc^2 \alpha_j \int_0^\infty \cdots \int_0^\infty k_\lambda(u_1, \cdots, u_{n-1}, 1) \prod_{j=1}^{n-1} u_j^{\lambda_j - \frac{\varepsilon}{p_j} - 1} du_1 \cdots du_{n-1}$$

$$= 2^n k(\lambda_1 - \frac{\varepsilon}{p_1}, \cdots, \lambda_{n-1} - \frac{\varepsilon}{p_{n-1}}) \prod_{j=1}^{n} \csc^2 \alpha_j$$

$$= 2^n k_\lambda \prod_{j=1}^{n} \csc^2 \alpha_j + o_2(1)(\varepsilon \to 0^+),$$

then we have (11).

The lemma is proved.

Lemma 5 *Suppose that*

$$k_\lambda = \int_{\mathbf{R}_+^{n-1}} k_\lambda(u_1, \cdots, u_{n-1}, 1) \prod_{j=1}^{n-1} u_j^{\lambda_j - 1} du_1 \cdots du_{n-1} \in \mathbf{R}.$$

If f_i (≥ 0) *are measurable functions in* $\mathbf{R}$ $(i = 1, \cdots, n-1)$, *putting*

$$\widetilde{k}(x_1, \cdots, x_n) := k_\lambda(u_1^{\delta_1}(x_1) u_n^{\delta_n}(x_n), \cdots, u_{n-1}^{\delta_{n-1}}(x_{n-1}) u_n^{\delta_n}(x_n), 1),$$

then

(i) for $p_i > 1$ $(i = 1, \cdots, n)$, *we have*

$$J := \left\{ \int_{-\infty}^{\infty} u_n^{\delta_n \lambda_n q_n - 1}(x_n) \right.$$

$$\left. \times \left[\int_{\mathbf{R}^{n-1}} \widetilde{k}(x_1, \cdots, x_n) \prod_{i=1}^{n-1} f_i(x_i) dx_1 \cdots dx_{n-1} \right]^{q_n} dx_n \right\}^{\frac{1}{q_n}}$$

$$\leq 2^{n-1} k_\lambda \prod_{j=1}^{n} \csc^{2(1-\frac{1}{p_j})} \alpha_j \prod_{i=1}^{n-1} \left[\int_{-\infty}^{\infty} u_i^{p_i(1-\delta_i \lambda_i)-1}(x_i) f^{p_i}(x_i) dx_i \right]^{\frac{1}{p_i}};$$

(14)

(ii) for $0 < p_1 < 1$, $p_i < 0$ $(i = 2, \cdots, n)$, *we have the reverse of (14).*

Proof

(i) For $p_i > 1$ $(i = 1, \cdots, n)$, by Hölder's inequality (cf. [18]) and (7), we have

$$\left(\int_{\mathbf{R}^{n-1}} \widetilde{k}(x_1, \cdots, x_n) \prod_{i=1}^{n-1} f_i(x_i) dx_1 \cdots dx_{n-1}\right)^{q_n}$$

$$= \left\{\int_{\mathbf{R}^{n-1}} \widetilde{k}(x_1, \cdots, x_n) \prod_{i=1}^{n-1} \left[u_i^{(\delta_i\lambda_i-1)(1-p_i)}(x_i) \prod_{j=1(j\neq i)}^{n} u_j^{\delta_j\lambda_j-1}(x_j)\right]^{\frac{1}{p_i}} f_i(x_i)\right.$$

$$\left.\times \left[u_n^{(\delta_n\lambda_n-1)(1-p_n)}(x_n) \prod_{j=1}^{n-1} u_j^{\delta_j\lambda_j-1}(x_j)\right]^{\frac{1}{p_n}} dx_1 \cdots dx_{n-1}\right\}^{q_n}$$

$$\leq \left\{\int_{\mathbf{R}^{n-1}} \widetilde{k}(x_1, \cdots, x_n) \prod_{i=1}^{n-1} \left[u_i^{(\delta_i\lambda_i-1)(1-p_i)}(x_i) \prod_{j=1(j\neq i)}^{n} u_j^{\delta_j\lambda_j-1}(x_j)\right]^{\frac{q_n}{p_i}}\right.$$

$$\left.\times f_i^{q_n}(x_i) dx_1 \cdots dx_{n-1}\right\}$$

$$\times \left\{\int_{\mathbf{R}_+^{n-1}} \widetilde{k}(x_1, \cdots, x_n) u^{(\delta_n\lambda_n-1)(1-p_n)}(x_n) \prod_{j=1}^{n-1} u^{\delta_j\lambda_j-1}(x_j) dx_1 \cdots dx_{n-1}\right\}^{q_n-1}$$

$$= (2^{n-1}k_\lambda \prod_{j=1}^{n-1} \csc^2\alpha_j)^{q_n-1} u^{1-\delta_n q_n \lambda_n}(x_n) \int_{\mathbf{R}_+^{n-1}} \widetilde{k}(x_1, \cdots, x_n)$$

$$\times \times \prod_{i=1}^{n-1} \left[u^{(\delta_i\lambda_i-1)(1-p_i)}(x_i) \prod_{j=1(j\neq i)}^{n} u^{\delta_j\lambda_j-1}(x_j)\right]^{\frac{q_n}{p_i}} f_i^{q_n}(x_i) dx_1 \cdots dx_{n-1}. \quad (15)$$

Then it follows that

$$J \leq (2^{n-1}k_\lambda \prod_{j=1}^{n-1} \csc^2\alpha_j)^{\frac{1}{p_n}} \left\{\int_{-\infty}^{\infty} \int_{\mathbf{R}^{n-1}} \widetilde{k}(x_1, \cdots, x_n)\right.$$

$$\left.\times \prod_{i=1}^{n-1} \left[u_i^{(\delta_i\lambda_i-1)(1-p_i)}(x_i) \prod_{j=1(j\neq i)}^{n} u_j^{\delta_j\lambda_j-1}(x_j)\right]^{\frac{q_n}{p_i}} f_i^{q_n}(x_i) dx_1 \cdots dx_{n-1} dx_n\right\}^{\frac{1}{q_n}}$$

$$= (2^{n-1}k_\lambda \prod_{j=1}^{n-1} \csc^2\alpha_j)^{\frac{1}{p_n}} \left\{\int_{\mathbf{R}^{n-1}} \left(\int_{-\infty}^{\infty} \widetilde{k}(x_1, \cdots, x_n) u_n^{\delta_n\lambda_n-1}(x_n) dx_n\right)\right.$$

$$\times\prod_{i=1}^{n-1}\left[u_i^{(\delta_i\lambda_i-1)(1-p_i)}(x_i)\prod_{j=1(j\neq i)}^{n-1}u_j^{\delta_j\lambda_j-1}(x_j)\right]^{\frac{q_n}{p_i}}f_i^{q_n}(x_i)dx_1\cdots dx_{n-1}\Bigg\}^{\frac{1}{q_n}}. \tag{16}$$

For $n\geq 3$, in view of $\sum_{i-1}^{n-1}\frac{q_n}{p_i}=1$, by H$\ddot{o}$lder's inequality again, it follows that

$$J\leq(2^{n-1}k_\lambda\prod_{j=1}^{n-1}\csc^2\alpha_j)^{\frac{1}{p_n}}\left\{\prod_{i=1}^{n-1}\left[\int_{\mathbf{R}^{n-1}}(\int_{-\infty}^{\infty}\widetilde{k}(x_1,\cdots,x_n)u_n^{\delta_n\lambda_n-1}(x_n)dx_n)\right.\right.$$

$$\left.\left.\times u_i^{(\delta_i\lambda_i-1)(1-p_i)}(x_i)\prod_{j=1(j\neq i)}^{n-1}u_j^{\delta_j\lambda_j-1}(x_j)f_i^{p_i}(x_i)dx_1\cdots dx_{n-1}\right]^{\frac{q_n}{p_i}}\right\}^{\frac{1}{q_n}}$$

$$\leq(2^{n-1}k_\lambda\prod_{j=1}^{n-1}\csc^2\alpha_j)^{\frac{1}{p_n}}\prod_{i=1}^{n-1}\left\{\int_{-\infty}^{\infty}\left[\int_{\mathbf{R}^{n-1}}\widetilde{k}(x_1,\cdots,x_n)u_i^{\delta_i\lambda_i}(x_i)\right.\right.$$

$$\left.\times\prod_{j=1(j\neq i)}^{n}u_j^{\delta_j\lambda_j-1}(x_j)dx_1\cdots dx_{i-1}dx_{i+1}\cdots dx_n\right]$$

$$\left.u_i^{p_i(1-\delta_i\lambda_i)-1}(x_i)f_i^{p_i}(x_i)dx_i\right\}^{\frac{1}{p_i}}$$

$$=(2^{n-1}k_\lambda\prod_{j=1}^{n-1}\csc^2\alpha_j)^{\frac{1}{p_n}}\prod_{i=1}^{n-1}\left[\int_{-\infty}^{\infty}\omega_i(x_i)u_i^{p_i(1-\delta_i\lambda_i)-1}(x_i)f_i^{p_i}(x_i)dx_i\right]^{\frac{1}{p_i}}.$$

Then by (9), we have (14) (Note: for $n=2$, we do not use H$\ddot{o}$lder's inequality again in the above).

(ii) For $0<p_1<1$, $p_i<0$ $(i=2,\cdots,n)$, by the reverse H$\ddot{o}$lder's inequality (cf. [18]) and the same way, we obtain the reverse of (14).

The lemma is proved.

3 Main Results and Operator Expressions

Setting the functions

$$\phi_i(x) := u_i^{p_i(1-\delta_i\lambda_i)-1}(x) \ (x \in \mathbf{R}; i = 1, \cdots, n),$$

we find $\phi_n^{q_n-1}(x) = u_n^{\delta_n q_n \lambda_n - 1}(x)$. If $p_i > 1$ $(i = 1, \cdots, n)$, we define the following real function spaces:

$$L_{\phi_i}^{p_i}(\mathbf{R}) := \left\{ f_i;\ ||f_i||_{p_i,\phi_i} = \left(\int_{-\infty}^{\infty} \phi_i(x)|f_i(x)|^{p_i} dx \right)^{\frac{1}{p_i}} < \infty \right\} \ (i = 1, \cdots, n),$$

$$\prod_{i=1}^{n-1} L_{\phi_i}^{p_i}(\mathbf{R}) := \left\{ f = (f_1, \cdots, f_{n-1});\ f_i \in L_{\phi_i}^{p_i}(\mathbf{R}), i = 1, \cdots, n-1 \right\},$$

and a multiple Hilbert-type integral operator $T : \prod_{i=1}^{n-1} L_{\phi_i}^{p_i}(\mathbf{R}) \to L_{\phi_n^{q_n-1}}^{q_n}(\mathbf{R})$ as follows: For $f = (f_1, \cdots, f_{n-1}) \in \prod_{i=1}^{n-1} L_{\phi_i}^{p_i}(\mathbf{R})$, there exists a unified expression Tf, satisfying for $x_n \in \mathbf{R}$,

$$(Tf)(x_n) := \int_{\mathbf{R}^{n-1}} \widetilde{k}(x_1, \cdots, x_n) \prod_{i=1}^{n-1} f_i(x_i) dx_1 \cdots dx_{n-1}. \tag{17}$$

Then by (14), it follows that $Tf \in L_{\phi_n^{q_n-1}}^{q_n}(\mathbf{R})$. T is bounded satisfying

$$||Tf||_{q_n,\phi_n^{q_n-1}} \leq 2^{n-1} k_\lambda \prod_{j=1}^{n} \csc^{2(1-\frac{1}{p_j})} \alpha_j \prod_{i=1}^{n-1} ||f_i||_{p_i,\phi_i},$$

and then $||T|| \leq 2^{n-1} k_\lambda \prod_{j=1}^{n} \csc^{2(1-\frac{1}{p_j})} \alpha_j$, where,

$$||T|| := \sup_{f(\neq\theta) \in \prod_{i=1}^{n-1} L_{\phi_i}^{p_i}(\mathbf{R})} \frac{||Tf||_{q_n,\phi_n^{q_n-1}}}{\prod_{i=1}^{n-1} ||f_i||_{p_i,\phi_i}}. \tag{18}$$

Define the formal inner product of $T(f_1, \cdots, f_{n-1})$ and f_n as follows:

$$(T(f_1, \cdots, f_{n-1}), f_n) := \int_{\mathbf{R}^n} \widetilde{k}(x_1, \cdots, x_n) \prod_{i=1}^{n} f_i(x_i) dx_1 \cdots dx_n. \tag{19}$$

Theorem 1 *Suppose that*

$$k_\lambda = \int_{\mathbf{R}_+^{n-1}} k_\lambda(u_1, \cdots, u_{n-1}, 1) \prod_{j=1}^{n-1} u_j^{\lambda_j - 1} du_1 \cdots du_{n-1} \in \mathbf{R}_+. \tag{20}$$

If $f_i (\geq 0) \in L_{\phi_i}^{p_i}(\mathbf{R})$, $||f||_{p_i,\phi_i} > 0$ $(i = 1, \cdots, n)$, *then*

(i) for $p_i > 1$ $(i = 1, \cdots, n)$, *we have the following equivalent inequalities:*

$$||T(f_1, \cdots, f_{n-1})||_{q_n, \phi_n^{q_{n-1}}} < 2^{n-1} k_\lambda \prod_{j=1}^{n} \csc^{2(1-\frac{1}{p_j})} \alpha_j \prod_{i=1}^{n-1} ||f_i||_{p_i,\phi_i}, \tag{21}$$

$$(T(f_1, \cdots, f_{n-1}), f_n) < 2^{n-1} k_\lambda \prod_{j=1}^{n} \csc^{2(1-\frac{1}{p_j})} \alpha_j \prod_{i=1}^{n} ||f_i||_{p_i,\phi_i}, \tag{22}$$

where, the constant factor $2^{n-1} k_\lambda \prod_{j=1}^{n} \csc^{2(1-\frac{1}{p_j})} \alpha_j$ *is the best possible, namely*

$$||T|| = 2^{n-1} k_\lambda \prod_{j=1}^{n} \csc^{2(1-\frac{1}{p_j})} \alpha_j;$$

(ii) for $0 < p_1 < 1, p_i < 0$ $(i = 2, \cdots, n)$, *using the formal symbols in the case of (i), we have the equivalent reverses of (21) and (22) with the same best constant factor.*

Proof

(i) For all $p_i > 1$, if (15) takes the form of equality, then for $n \geq 3$ in (21), there exist C_i and C_k $(i \neq k)$, such that they are not all zero and

$$C_i u_i^{(\delta_i \lambda_i - 1)(1-p_i)}(x_i) \prod_{j=1 (j \neq i)}^{n-1} u_j^{\delta_j \lambda_j - 1}(x_j) f_j^{p_j}(x_j)$$
$$= C_k u_k^{(\delta_k \lambda_k - 1)(1-p_k)}(x_k) \prod_{j=1 (j \neq k)}^{n-1} u_j^{\delta_j \lambda_j - 1}(x_j) f_j^{p_j}(x_j) \ a.e. \text{ in } \mathbf{R}^n,$$

namely,

$$C_i u_i^{p_i(1-\delta_i \lambda_i)}(x_i) f_i^{p_i}(x_i) = C_k u_k^{p_k(1-\delta_k \lambda_k)}(x_k) f_k^{p_k}(x_k) = C \ \ a.e. \, in \, \mathbf{R}_+^n.$$

Assuming that $C_i > 0$, then

$$u_i^{p_i(1-\delta_i\lambda_i)-1}(x_i)f_i^{p_i}(x_i) = \frac{C}{C_i u_i(x_i)},$$

which contradicts the fact that $0 < ||f||_{p_{i,\phi_i}} < \infty$ (Note: for $n = 2$, we consider (15) for $f_k^{p_i}(x_k) = 1$ in the above). Hence we have (21).

By H$\ddot{o}$lder's inequality, it follows that

$$(Tf, f_n) = \int_{-\infty}^{\infty} \left(u_n^{\delta_n\lambda_n - \frac{1}{q_n}}(x_n) \int_{\mathbf{R}^{n-1}} \widetilde{k}(x_1, \cdots, x_n) \prod_{i=1}^{n-1} f_i(x_i) dx_1 \cdots dx_{n-1} \right)$$

$$\times \left(u_n^{\frac{1}{q_n} - \delta_n\lambda_n}(x_n) f_n(x_n) \right) dx_n \leq ||T(f_1, \cdots, f_{n-1})||_{q_n, \phi_n^{q_n-1}} ||f_n||_{p_n,\phi_n}, \tag{23}$$

and then by (21), we have (22). Assuming that (22) is valid, setting

$$f_n(x_n) := u_n^{\delta_n q_n \lambda_n - 1}(x_n) \left(\int_{\mathbf{R}^{n-1}} \widetilde{k}(x_1, \cdots, x_n) \prod_{i=1}^{n-1} f_i(x_i) dx_1 \cdots dx_{n-1} \right)^{q_n-1},$$

then it follows that

$$J = \left[\int_{-\infty}^{\infty} u_n^{p_n(1-\delta_n\lambda_n)-1}(x_n) f_n^{p_n}(x_n) dx_n \right]^{\frac{1}{q_n}}.$$

By (14), it follows that $J < \infty$. If $J = 0$, then (21) is trivially valid. Assuming that $0 < J < \infty$, by (22), it follows that

$$\int_{-\infty}^{\infty} u_n^{p_n(1-\delta_n\lambda_n)-1}(x_n) f_n^{p_n}(x_n) dx_n$$

$$= J^{q_n} = (Tf, f_n) < 2^{n-1} k_\lambda \prod_{j=1}^{n} \csc^{2(1-\frac{1}{p_j})} \alpha_j \prod_{i=1}^{n} ||f_i||_{p_i,\phi_i},$$

$$\left[\int_{-\infty}^{\infty} u_n^{p_n(1-\delta_n\lambda_n)-1}(x_n) f_n^{p_n}(x_n) dx_n \right]^{\frac{1}{q_n}}$$

$$= J < 2^{n-1} k_\lambda \prod_{j=1}^{n} \csc^{2(1-\frac{1}{p_j})} \alpha_j \prod_{i=1}^{n-1} ||f_i||_{p_i,\phi_i},$$

and then (21) is valid, which is equivalent to (22).

For $E_i := \{x \in \mathbf{R}; {u_i}^{\delta_i}(x) \in [1,\infty)\}$ $(i = 1, \cdots, n)$, $\varepsilon > 0$, we set $\widetilde{f}_i(x_i)$ as follows:

$$\widetilde{f}_i(x_i) = 0, x_i \in \mathbf{R}\backslash E_i;$$
$$\widetilde{f}_i(x_i) = u_i^{\delta_i(\lambda_i - \frac{\varepsilon}{p_i})-1}(x_i), x_i \in E_i \ (i = 1, \cdots, n-1),$$
$$\widetilde{f}_n(x_n) = u_n^{\delta_n(\lambda_n + \frac{\varepsilon}{p_n})-1}(x_n), x \in \mathbf{R}\backslash E_n; \ \widetilde{f}_n(x_n) = 0, x_n \in E_n.$$

We find

$$||\widetilde{f}_i||_{p_i,\phi_i} = \left[\int_{E_i} u_i^{p_i(1-\delta_i\lambda_i)-1}(x_i) u_i^{\delta_i(p_i\lambda_i - \varepsilon)-p_i}(x_i)dx_i\right]^{\frac{1}{p_i}}$$
$$= \left(\int_{E_i} u_i^{-\delta_i\varepsilon-1}(x_i)dx_i\right)^{\frac{1}{p_i}}.$$

If $\delta_i = 1$ $(i = 1, \cdots, n-1)$, setting $y = u_i(x_i)$, then we have $dx_i = 2\csc^2\alpha_i dy$ and

$$||\widetilde{f}_i||_{p_i,\phi_i} = \left(2\csc^2\alpha_i \int_1^\infty y^{-\varepsilon-1}dy\right)^{\frac{1}{p_i}} = \left(\frac{2}{\varepsilon}\csc^2\alpha_i\right)^{\frac{1}{p_i}};$$

if $\delta_i = -1$, still setting $y = u_i(x_i)$, then we have

$$||\widetilde{f}_i||_{p_i,\phi_i} = \left(2\csc^2\alpha_i \int_0^1 y^{\varepsilon-1}dy\right)^{\frac{1}{p_i}} = \left(\frac{2}{\varepsilon}\csc^2\alpha_i\right)^{\frac{1}{p_i}}.$$

In the same way, we find

$$||\widetilde{f}_n||_{p_n,\phi_n} = \left(\frac{2}{\varepsilon}\csc^2\alpha_n\right)^{\frac{1}{p_n}}.$$

If there exists a positive constant $k \le k_\lambda$, such that (22) is still valid when replacing k_λ by k, then in particular, by Lemma 4, we have

$$2^n k_\lambda \prod_{j=1}^n \csc^2\alpha_j + o(1) = I_\varepsilon = \varepsilon(T(\widetilde{f}_1, \cdots, \widetilde{f}_{n-1}), \widetilde{f}_n)$$
$$< \varepsilon 2^{n-1} k \prod_{j=1}^n \csc^{2-\frac{2}{p_j}}\alpha_j \prod_{i=1}^n ||\widetilde{f}_i||_{p_i,\phi_i}$$

$$= \varepsilon 2^{n-1} k \prod_{j=1}^{n} \csc^{2-\frac{2}{p_j}} \alpha_j \prod_{i=1}^{n} (\frac{2}{\varepsilon} \csc^2 \alpha_i)^{\frac{1}{p_i}} = 2^n k \prod_{j=1}^{n} \csc^2 \alpha_j,$$

and then $k_\lambda \leq k$ $(\varepsilon \to 0^+)$. Hence $k = k_\lambda$ is the best value of (22).

The constant factor k_λ in (21) is still the best possible. Otherwise we would reach a contradiction by (23) that the constant factor in (22) is not the best possible.

(ii) For $0 < p_1 < 1, p_i < 0$ $(i = 2, \cdots, n)$, by using the reverse H$\ddot{o}$lder's inequality and in the same way, we have the equivalent reverses of (21) and (22) with the same best constant factor.

The theorem is proved.

Remark 3.1

(i) For $\delta_i = 1, \alpha_i = \frac{\pi}{2}$ $(i = 1, \cdots, n)$ in (21) and (22), we have the following equivalent inequalities with the non-homogeneous kernel and best possible constant factor $2^{n-1}k_\lambda$

$$\int_{\mathbf{R}^n} k_\lambda(|x_1 x_n|, \cdots, |x_{n-1}x_n|, 1) \prod_{i=1}^{n} f_i(x_i) dx_1 \cdots dx_n$$
$$< 2^{n-1} k_\lambda \prod_{i=1}^{n} \left[\int_{-\infty}^{\infty} |x_i|^{p_i(1-\lambda_i)-1} f^{p_i}(x_i) dx_i \right]^{\frac{1}{p_i}}, \tag{24}$$

$$\left\{ \int_{-\infty}^{\infty} |x_n|^{\lambda_n q_n - 1} \left[\int_{\mathbf{R}^{n-1}} k_\lambda(|x_1 x_n|, \cdots, |x_{n-1}x_n|, 1) \prod_{i=1}^{n-1} f_i(x_i) dx_1 \cdots dx_{n-1} \right]^{q_n} \right.$$
$$\left. dx_n \right\}^{\frac{1}{q_n}}$$
$$< 2^{n-1} k_\lambda \prod_{i=1}^{n-1} \left[\int_{-\infty}^{\infty} |x_i|^{p_i(1-\lambda_i)-1} f^{p_i}(x_i) dx_i \right]^{\frac{1}{p_i}}. \tag{25}$$

In particular, for $f_i(-x_i) = f_i(x_i)$ $(x_i > 0; i = 1, \cdots, n)$ in (24) and (25), we have (cf. [15]):

$$\int_{\mathbf{R}_+^n} k_\lambda(x_1 x_n, \cdots, x_{n-1}x_n, 1) \prod_{i=1}^{n} f_i(x_i) dx_1 \cdots dx_n$$
$$< k_\lambda \prod_{i=1}^{n} \left[\int_{0}^{\infty} {x_i}^{p_i(1-\lambda_i)-1} f^{p_i}(x_i) dx_i \right]^{\frac{1}{p_i}}, \tag{26}$$

$$\left\{\int_0^\infty x_n{}^{\lambda_n q_n-1}\left[\int_{\mathbf{R}_+^{n-1}} k_\lambda(x_1x_n,\cdots,x_{n-1}x_n,1)\prod_{i=1}^{n-1} f_i(x_i)dx_1\cdots dx_{n-1}\right]^{q_n} dx_n\right\}^{\frac{1}{q_n}}$$

$$< k_\lambda \prod_{i=1}^{n-1}\left[\int_0^\infty x_i{}^{p_i(1-\lambda_i)-1} f^{p_i}(x_i)dx_i\right]^{\frac{1}{p_i}}. \tag{27}$$

(ii) For $\delta_i = 1, \alpha_i = \frac{\pi}{2}$ $(i = 1, \cdots, n-1), \alpha_n = \frac{\pi}{2}, \delta_n = -1$ in (21) and (22), replacing $|x_n|^\lambda f_n(x_n)$ by $f_n(x_n)$, we have the following equivalent inequalities with the homogeneous kernel and a best possible constant factor k_λ :

$$\int_{\mathbf{R}^n} k_\lambda(|x_1|,\cdots,|x_n|)\prod_{i=1}^{n} f_i(x_i)dx_1\cdots dx_n$$

$$< 2^{n-1}k_\lambda \prod_{i=1}^{n}\left[\int_{-\infty}^\infty |x_i|^{p_i(1-\lambda_i)-1} f^{p_i}(x_i)dx_i\right]^{\frac{1}{p_i}}, \tag{28}$$

$$\left\{\int_{-\infty}^\infty |x_n|^{\lambda_n q_n-1}\left[\int_{\mathbf{R}^{n-1}} k_\lambda(|x_1|,\cdots,|x_n|)\prod_{i=1}^{n-1} f_i(x_i)dx_1\cdots dx_{n-1}\right]^{q_n} dx_n\right\}^{\frac{1}{q_n}}$$

$$< 2^{n-1}k_\lambda \prod_{i=1}^{n-1}\left[\int_{-\infty}^\infty |x_i|^{p_i(1-\lambda_i)-1} f^{p_i}(x_i)dx_i\right]^{\frac{1}{p_i}}. \tag{29}$$

For $\lambda_i = \frac{\lambda}{r_i} > 0$, $f_i(-x_i) = f_i(x_i)(x_i > 0)$ $(i = 1, \cdots, n)$, inequality (28) reduces to (5) (for $r_n = 2$).

(iii) For $n = 2$ in (24), we have

$$\int_{-\infty}^\infty\int_{-\infty}^\infty k_\lambda(|xy|,1)f(x)g(y)dxdy$$

$$< 2k_\lambda\left[\int_0^\infty x^{p(1-\frac{\lambda}{2})-1} f^p(x)dx\right]^{\frac{1}{p}}\left[\int_0^\infty x^{q(1-\frac{\lambda}{2})-1} g^q(x)dx\right]^{\frac{1}{q}}, \tag{30}$$

where, $k_\lambda = \int_0^\infty k_\lambda(u,1)u^{\frac{\lambda}{2}-1}du > 0\,(\lambda \in \mathbf{R})$ is the best possible. Inequality (30) is an extension of (4) (for $k_\lambda(xy,1) = \frac{1}{(xy+1)^\lambda}$, $p = q = 2$).

4 Some Examples

Example 1 For $\lambda > 0, \lambda_i = \frac{\lambda}{r_i}$ $(i = 1, \cdots, n), r_n = 2, \sum_{i=1}^{n} \frac{1}{r_i} = 1$,

$$k_\lambda(x_1, \cdots, x_n) = \frac{1}{(\sum_{i=1}^{n} x_i)^\lambda},$$

by mathematical induction, we can show that (cf. [6])

$$k_\lambda = \int_{\mathbf{R}_+^{n-1}} \frac{\prod_{j=1}^{n-1} u_j^{\frac{\lambda}{r_j}-1}}{(\sum_{i=1}^{n-1} u_i + 1)^\lambda} du_1 \cdots du_{n-1} = \frac{1}{\Gamma(\lambda)} \prod_{i=1}^{n} \Gamma(\frac{\lambda}{r_i}). \tag{31}$$

By (22), we have

$$\int_{\mathbf{R}^n} \frac{1}{(\sum_{i=1}^{n-1} u_i^{\delta_i}(x_i) u_n^{\delta_n}(x_n) + 1)^\lambda} \prod_{i=1}^{n} f_i(x_i) dx_1 \cdots dx_n$$
$$< \frac{2^{n-1}}{\Gamma(\lambda)} \prod_{j=1}^{n} \Gamma(\lambda_j) \csc^{2(1-\frac{1}{p_j})} \alpha_j \prod_{i=1}^{n} ||f_i||_{p_i,\phi_i}. \tag{32}$$

Example 2 For $\lambda > 0, \lambda_i = \frac{\lambda}{r_i}$ $(i = 1, \cdots, n), r_n = 2, \sum_{i=1}^{n} \frac{1}{r_i} = 1$,

$$k_\lambda(x_1, \cdots, x_n) = \frac{1}{\sum_{i=1}^{n} x_i^\lambda},$$

we can show that

$$k_\lambda = \int_{\mathbf{R}_+^{n-1}} \frac{\prod_{j=1}^{n-1} u_j^{\frac{\lambda}{r_j}-1}}{\sum_{i=1}^{n-1} u_i^\lambda + 1} du_1 \cdots du_{n-1} = \frac{1}{\lambda^{n-1}} \prod_{i=1}^{n} \Gamma(\frac{1}{r_i}). \tag{33}$$

In fact, setting $v_i = u_i^\lambda$ $(i = 1, \cdots, n-1)$ in the above integral, we find $u_i = v_i^{\frac{1}{\lambda}}, du_i = \frac{1}{\lambda} v_i^{\frac{1}{\lambda}-1} dv_i$ and

$$k_\lambda = \frac{1}{\lambda^{n-1}} \int_{\mathbf{R}_+^{n-1}} \frac{\prod_{j=1}^{n-1} v_j^{\frac{1}{r_j}-1}}{\sum_{i=1}^{n-1} v_i + 1} dv_1 \cdots dv_{n-1}.$$

In view of (31), for $\lambda = 1$, we have (32).

By (22), we have

$$\int_{\mathbf{R}^n} \frac{1}{\sum_{i=1}^{n-1} u_i^{\lambda\delta_i}(x_i) u_n^{\lambda\delta_n}(x_n) + 1} \prod_{i=1}^{n} f_i(x_i) dx_1 \cdots dx_n$$

$$< (\frac{2}{\lambda})^{n-1} \prod_{j=1}^{n} \Gamma(\lambda_j) \csc^{2(1-\frac{1}{p_j})} \alpha_j \prod_{i=1}^{n} ||f_i||_{p_i,\phi_i}. \tag{34}$$

Example 3 For $\lambda > 0, \lambda_i = \frac{\lambda}{r_i}$ $(i = 1, \cdots, n), r_n = 2, \sum_{i=1}^{n} \frac{1}{r_i} = 1$,

$$k_\lambda(x_1, \cdots, x_n) = \frac{1}{(\max_{1 \le i \le n}\{x_i\})^\lambda},$$

we can show that (cf. [6])

$$k_\lambda = \int_{R_+^{n-1}} \frac{1}{(\max_{1 \le i \le n-1}\{u_i\}, 1)^\lambda} \prod_{j=1}^{n-1} u_j^{\frac{\lambda}{r_j}-1} du_1 \cdots du_{n-1}$$

$$= \frac{1}{\lambda^{n-1}} \prod_{i=1}^{n} r_i. \tag{35}$$

By (22), we have

$$\int_{\mathbf{R}^n} \frac{1}{(\max_{1 \le i \le n-1}\{u_i^{\delta_i}(x_i) u_n^{\delta_n}(x_n), 1\})^\lambda} \prod_{i=1}^{n} f_i(x_i) dx_1 \cdots dx_n$$

$$< (\frac{2}{\lambda})^{n-1} \prod_{j=1}^{n} \Gamma(\frac{\lambda}{\lambda_j}) \csc^{2(1-\frac{1}{p_j})} \alpha_j \prod_{i=1}^{n} ||f_i||_{p_i,\phi_i}. \tag{36}$$

Acknowledgements This work is supported by the National Natural Science Foundation (No. 61772140), and Appropriative Researching Fund for Professors and Doctors, Guangdong University of Education (No. 2015ARF25). I am grateful for their help.

References

1. G.H. Hardy, J.E. Littlewood, G. Pólya, *Inequalities* (Cambridge University Press, Cambridge, 1934)
2. D.S. Mitrinović, J. Pečarić, A.M. Fink, *Inequalities Involving Functions and Their Integrals and Derivatives* (Kluwer Academic Publishers, Boston, 1991)
3. K.W. Zhang, A bilinear inequality. J. Math. Anal. Appl. **271**, 288–296 (2002)
4. B.C. Yang, On a extension of Hilbert's integral inequality with some parameters. Aust. J. Math. Anal. Appl. **1**(1), Article 11, 1–8 (2004)

5. B.C. Yang, A new Hilbert's type integral inequality. Soochow J. Math. **33**(4), 849–859 (2007)
6. B.C. Yang, *The Norm of Operator and Hilbert-Type Inequalities* (Science Press, Beijing, 2009)
7. G.V. Milovanovic, M.T. Rassias, Some properties of a hypergeometric function which appear in an approximation problem. J. Glob. Optim. **57**, 1173–1192 (2013)
8. Q.L. Huang, A new extension of Hardy-Hilbert-type inequality. J. Inequal. Appl. **2015**, 397 (2015)
9. M. Krnić, J. Pec̆arić, General Hilbert's and Hardy's inequalities. Math. Inequal. Appl. **8** (1), 29–51 (2005)
10. G.V. Milovanovic, M.T. Rassias (eds.), *Analytic Number Theory, Approximation Theory and Special Functions* (Springer, New York, 2014)
11. A. Benyi, C.T. Oh, Best constant for certain multilinear integral operator. J. Inequal. Appl. **2006**, Article ID 28582, 1–12 (2006)
12. H. Hong, All-side generalization about Hardy-Hilbert integral inequalities. Acta Math. Sinica **44**(4), 619–625 (2001)
13. L.P. He, J. Yu, M.Z. Gao, An extension of Hilbert's integral inequality. J. Shaoguan Univ. (Nat. Sci.) **23**(3), 25–30 (2002)
14. B. He, A multiple Hilbert-type discrete inequality with a new kernel and best possible constant factor. J. Math. Anal. Appl. **431**, 990–902 (2015)
15. Q.L. Huang, B.C. Yang, A multiple Hilbert-type inequality with a non-homogeneous kernel. J. Inequal. Appl. **2013**, 73 (2013)
16. I. Perić, P. Vuković, Multiple Hilbert's type inequalities with a homogeneous kernel. Banach J. Math. Anal. **5**(2), 33–43 (2011)
17. J.C. Kuang, *Real and Functional Analysis* (Continuation) 2nd vol. (Higher Education Press, Beijing, 2015)
18. J.C. Kuang, *Applied Inequalities* (Shangdong Science Technic Press, Jinan, 2004)

Generalizations of Metric Spaces: From the Fixed-Point Theory to the Fixed-Circle Theory

Nihal Yılmaz Özgür and Nihal Taş

1 Introduction

Fixed-point theory has been extensively studied on metric spaces since the time of Stefan Banach (see [5, 9] for more details). Applications of this theory to the applied areas such as differential equations, integral equations etc. are well-known (see [8, 11, 18, 20–22, 24, 45] (Jha, Banach contraction principle and some generalizations. Unpublished M. Phil. Thesis, Kathmandu University, Nepal (1999))). Recently some generalized metric spaces have been introduced and studied as the generalizations of metric spaces (see [1, 2], [4, 12, 15, 26, 30, 42–44, 46, 47, 49]). For example, S-metric spaces and S_b-metric spaces have been presented for this purpose. Using the obtained generalizations, the known classical fixed-point results have been generalized (see [29, 31, 32, 41, 43, 44, 48] for more details). Furthermore some applications have been obtained to the other areas (see [14, 25, 35–37] for more details).

Now we recall the definitions of an S-metric space and an S_b-metric space as follows:

Definition 1 ([43]) Let X be a nonempty set and $\mathscr{S} : X \times X \times X \to [0, \infty)$ be a function satisfying the following conditions for all $x, y, z, a \in X$:

$(S1)$ $\quad \mathscr{S}(x, y, z) = 0$ if and only if $x = y = z$,
$(S2)$ $\quad \mathscr{S}(x, y, z) \leq \mathscr{S}(x, x, a) + \mathscr{S}(y, y, a) + \mathscr{S}(z, z, a)$.

Then $\mathscr{S}$ is called an S-metric on X and the pair $(X, \mathscr{S})$ is called an S-metric space.

N. Y. Özgür (✉) · N. Taş
Balıkesir University, Department of Mathematics, Balıkesir, Turkey
e-mail: nihal@balikesir.edu.tr; nihaltas@balikesir.edu.tr

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_28

Definition 2 ([44]) Let X be a nonempty set and $b \geq 1$ be a given real number. A function $\mathscr{S}_b : X \times X \times X \to [0, \infty)$ is said to be an S_b-metric if and only if for all $x, y, z, a \in X$ the following conditions are satisfied:

(S_b1) $\mathscr{S}_b(x, y, z) = 0$ if and only if $x = y = z$,
(S_b2) $\mathscr{S}_b(x, y, z) \leq b[\mathscr{S}_b(x, x, a) + \mathscr{S}_b(y, y, a) + \mathscr{S}_b(z, z, a)]$.

The pair $(X, \mathscr{S}_b)$ is called an S_b-metric space.

An S_b-metric space is also a generalization of an S-metric space. Indeed, every S-metric is an S_b-metric with $b = 1$. But the converse of this statement is not always true as seen in the following example.

Example 1 ([48]) Let $X = \mathbb{R}$ and the S-metric be defined by

$$\mathscr{S}(x, y, z) = \frac{1}{4}(|x - y| + |y - z| + |x - z|),$$

for all $x, y, z \in \mathbb{R}$. Using this S-metric we define

$$\mathscr{S}_b(x, y, z) = \mathscr{S}(x, y, z)^2 = \frac{1}{16}(|x - y| + |y - z| + |x - z|)^2.$$

Then the function $\mathscr{S}_b$ is an S_b-metric with $b = 4$, but it is not an S-metric.

In the following diagram it can be seen the relationships among the metric, S-metric and S_b-metric spaces.

$$\begin{tikzcd}\boxed{\text{metric spaces}} \arrow[r] & \boxed{S\text{-metric spaces}} \arrow[r] & \boxed{S_b\text{-metric spaces}}\end{tikzcd}$$

It has been extensively studied the existence of fixed points of functions which satisfy certain conditions with different aspects. At first we recall the Banach's contraction principle as follows:

Theorem 1 ([9]) *Let (X, d) be a complete metric space and a self-mapping $T : X \to X$ be a contraction, that is, there exists an $h \in [0, 1)$ such that*

$$d(Tx, Ty) \leq hd(x, y),$$

for any $x, y \in X$. Then there exists a unique fixed point $x_0 \in X$ of T.

Many authors have been studied new fixed-point theorems on a complete metric space. For example, Caristi gave the following fixed-point theorem.

Theorem 2 ([7]) *Let (X, d) be a complete metric space and $T : X \to X$ be a mapping. If there exists a lower semicontinuous function φ mapping X into the nonnegative real numbers satisfying*

$$d(x, Tx) \leq \varphi(x) - \varphi(Tx), \tag{1}$$

for all $x \in X$ then T has a fixed point.

The following theorem was given on a compact metric space by Edelstein [13]. Also this result was proved independently by Nemytskii [27].

Theorem 3 ([13, 27]) *Let T be a mapping from a compact metric space (X, d) into itself satisfying*

$$d(Tx, Ty) < d(x, y),$$

for all $x, y \in X$ with $x \neq y$. Then T has a unique fixed point.

It has been also introduced and studied new contractive conditions to obtain new fixed-point theorems. For example, the following contractive condition was given in L. B. Ćirić's result [10]:

There exists a constant h, $0 < h < 1$, such that, for each $x, y \in X$,

$$d(Tx, Ty) \leq h \max\{d(x, y), d(x, Tx), d(y, Ty), d(x, Ty), d(y, Tx)\}.$$

As an another example, Rhoades defined the following condition (which is called the Rhoades' condition) [40]:

$$d(Tx, Ty) < \max\{d(x, y), d(x, Tx), d(y, Ty), d(x, Ty), d(y, Tx)\},$$

for all $x, y \in X, x \neq y$.

Some known fixed-point theorems have been extended on some generalized metric spaces such as an S-metric space. For example, Banach's contraction principle has been extended using some different methods (see [29, 43]). The classical Ćirić's fixed-point result has been generalized in [41]. Also the classical Nemytskii-Edelstein fixed-point theorem has been obtained on a compact S-metric space [43]. The present authors introduced the Rhoades' contractive condition and some generalizations of it on S-metric spaces (see [31, 32]). Using these generalized contractive conditions, they proved new fixed-point results as the generalizations of the Ćirić's fixed-point result and the Nemytskii-Edelstein fixed-point theorem.

In some special metric spaces, mappings with fixed points have been used in neural networks as activation functions. For example, Möbius transformations have been used for this purpose. A Möbius transformation M is a rational function of the form

$$Mz = \frac{az + b}{cz + d}, \tag{2}$$

where a, b, c, d are complex numbers satisfying $ad - bc \neq 0$. A Möbius transformation has at most two fixed points (see [19] for more details about Möbius transformations). In [23], Mandic identified the activation function of a neuron and a single-pole all-pass digital filter section as Möbius transformations. On the other hand, there are some examples of functions which fix a circle. For example, let $\mathbb{C}$ be the metric space with the usual metric

$$d(z, w) = |z - w|,$$

for all $z, w \in \mathbb{C}$ and the mapping T be defined as

$$Tz = \frac{1}{\overline{z}},$$

for all $z \in \mathbb{C} \setminus \{0\}$, where $\overline{z}$ is the complex conjugate of the complex number z. The mapping T fixes the unit circle $C_{0,1}$. In [28], using the self-mapping T, Özdemir, İskender and Özgür obtained new types of activation functions which fix a circle for a complex valued neural network (CVNN). The usage of these types activation functions leads us to guarantee the existence of the fixed points of a complex valued Hopfield neural network (CVHNN).

Therefore it is important to study the mappings with a fixed circle and the notion of a fixed circle. It will be an interesting problem to investigate some fixed-circle theorems on some spaces such as metric or normed spaces. More recently, some fixed-circle theorems have been presented as a different direction for the generalizations of the known fixed-point theorems. Existence and uniqueness conditions for fixed-circles of self-mappings have been investigated on a metric and an S-metric space (see [33, 34, 38]).

Let $X = \mathbb{C}$ and the mapping $\mathscr{S} : X \times X \times X \to [0, \infty)$ be defined as

$$\mathscr{S}(z_1, z_2, z_3) = |z_1 - z_3| + |z_1 + z_3 - 2z_2|, \tag{3}$$

for all $z_1, z_2, z_3 \in \mathbb{C}$. Then $(\mathbb{C}, \mathscr{S})$ is an S-metric space. Let us consider the circle $C^S_{0,3}$ and define the self-mapping $T : \mathbb{C} \to \mathbb{C}$ by

$$Tz = \begin{cases} \frac{9}{4\overline{z}} & ; z \neq 0 \\ 0 & ; z = 0 \end{cases},$$

for all $z \in \mathbb{C}$. Then $C^S_{0,3}$ is the fixed circle of T. If we consider the self-mapping $T : \mathbb{C} \to \mathbb{C}$ by

$$Tz = \begin{cases} \frac{9}{4z} & ; z \neq 0 \\ 0 & ; z = 0 \end{cases},$$

for all $z \in \mathbb{C}$, then clearly T does not fix the circle $C^S_{0,3}$ but T maps the circle $C^S_{0,3}$ onto itself. Especially T fixes the points $z_1 = \frac{3}{2}$ and $z_2 = -\frac{3}{2}$ only. Therefore it is important to study new existence and uniqueness conditions for fixed-circles of self-mappings.

In this study, we give a survey about the fixed-point theory on some generalized metric spaces and obtain new fixed-point (resp. fixed-circle) results on an S_b-metric space. In Sect. 2, we recall some basic facts and results on S-metric and S_b-metric spaces. In Sect. 3, we present new contractive conditions as the generalizations of the Rhoades' conditions using the theory of an S_b-metric space and study some fixed-point theorems for self-mappings satisfying the Rhoades' conditions. In Sect. 4, we obtain new generalizations of the Nemytskii-Edelstein fixed-point theorem and the Ćirić's fixed-point result. In Sect. 5, we give a brief survey about the known fixed-circle results on a metric (resp. an S-metric) space. We prove new fixed-circle theorems on metric and S_b-metric spaces with a geometric viewpoint. Especially we give some existence and uniqueness conditions for fixed-circle results on an S_b-metric space. Our results can be also considered as the new generalizations of the known fixed-point results on a metric and an S-metric space.

2 Some Generalized Metric Spaces

At first, we consider some basic properties about S-metric (resp. S_b-metric) spaces. We give necessary facts and theorems on these spaces which will be used in the next sections.

Definition 3 ([43]) Let $(X, \mathscr{S})$ be an S-metric space and $A \subset X$.

1. A sequence $\{x_n\}$ in X converges to x if and only if $\mathscr{S}(x_n, x_n, x) \to 0$ as $n \to \infty$. That is, for each $\varepsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that $\mathscr{S}(x_n, x_n, x) < \varepsilon$ for all $n \geq n_0$. We denote this by $\lim\limits_{n\to\infty} x_n = x$ or $\lim\limits_{n\to\infty} \mathscr{S}(x_n, x_n, x) = 0$.
2. A sequence $\{x_n\}$ in X is called a Cauchy sequence if $\mathscr{S}(x_n, x_n, x_m) \to 0$ as $n, m \to \infty$. That is, for each $\varepsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that $\mathscr{S}(x_n, x_n, x_m) < \varepsilon$ for all $n, m \geq n_0$.
3. The S-metric space $(X, \mathscr{S})$ is called complete if every Cauchy sequence in X is convergent.

Lemma 1 ([43]) *Let $(X, \mathscr{S})$ be an S-metric space. Then we have*

$$\mathscr{S}(x, x, y) = \mathscr{S}(y, y, x).$$

The relation between a metric and an S-metric is given in [17] as follows:

Lemma 2 ([17]) *Let (X, d) be a metric space. Then the following properties are satisfied:*

1. *$\mathscr{S}_d(x, y, z) = d(x, z) + d(y, z)$ for all $x, y, z \in X$ is an S-metric on X.*
2. *$x_n \to x$ in (X, d) if and only if $x_n \to x$ in $(X, \mathscr{S}_d)$.*
3. *$\{x_n\}$ is Cauchy in (X, d) if and only if $\{x_n\}$ is Cauchy in $(X, \mathscr{S}_d)$.*
4. *(X, d) is complete if and only if $(X, \mathscr{S}_d)$ is complete.*

The metric $\mathscr{S}_d$ defined in Lemma 2 (1) is called the S-metric generated by d [31].

Note that there exist some examples of S-metrics satisfying $\mathscr{S} \neq \mathscr{S}_d$ for all metrics d (see [17, 31]). We recall the following example.

Example 2 ([31]) Let $X = \mathbb{R}$ and define the function

$$\mathscr{S}(x, y, z) = |x - z| + |x + z - 2y|,$$

for all $x, y, z \in \mathbb{R}$. Then $(X, \mathscr{S})$ is an S-metric space. Now we prove that there is no any metric d such that $\mathscr{S} = \mathscr{S}_d$. Conversely, suppose that there exists a metric d such that

$$\mathscr{S}(x, y, z) = d(x, z) + d(y, z),$$

for all $x, y, z \in \mathbb{R}$. Then we obtain

$$\mathscr{S}(x, x, z) = 2d(x, z) = 2|x - z| \text{ and so } d(x, z) = |x - z|$$

and

$$\mathscr{S}(y, y, z) = 2d(y, z) = 2|y - z| \text{ and so } d(y, z) = |y - z|,$$

for all $x, y, z \in \mathbb{R}$. Hence we have

$$|x - z| + |x + z - 2y| = |x - z| + |y - z|,$$

which is a contradiction. Therefore $\mathscr{S} \neq \mathscr{S}_d$.

Let $(X, \mathscr{S})$ be any S-metric space. In [16], it was shown that every S-metric on X defines a metric d_S on X as follows:

$$d_S(x, y) = \mathscr{S}(x, x, y) + \mathscr{S}(y, y, x), \tag{4}$$

for all $x, y \in X$. However, in [31] it was noted that the function $d_S(x, y)$ defined in (4) does not always define a metric. It can be easily checked that the triangle inequality is not satisfied for all elements of X everywhen for the function d_S [31]. More precisely, it was proved that the function d_S defined in (4) is a b-metric on X, but it is not always a metric since every b-metric need not to be a metric [39]. In the case that d_S is a metric, d_S is called as the metric generated by S [31]. In the following we give an example of an S-metric which does not generate a metric.

Example 3 ([31]) Let $X = \{1, 2, 3\}$ and the function $\mathscr{S} : X \times X \times X \to [0, \infty)$ be defined as:

$$\begin{aligned}
&\mathscr{S}(1, 1, 2) = \mathscr{S}(2, 2, 1) = 5,\\
&\mathscr{S}(2, 2, 3) = \mathscr{S}(3, 3, 2) = \mathscr{S}(1, 1, 3) = \mathscr{S}(3, 3, 1) = 2,\\
&\mathscr{S}(x, y, z) = 0 \text{ if } x = y = z,\\
&\mathscr{S}(x, y, z) = 1 \text{ otherwise},
\end{aligned}$$

for all $x, y, z \in X$. Then the function $\mathscr{S}$ is an S-metric which is not generated by any metric and the pair $(X, \mathscr{S})$ is an S-metric space. But the function d_S defined in (4) is not a metric on X. Indeed, for $x = 1, y = 2, z = 3$ we get

$$d_S(1, 2) = 10 \nleq d_S(1, 3) + d_S(3, 2) = 8.$$

Proposition 1 ([39]) *Let $X \neq \emptyset$ be any set. If an S-metric is generated by any metric then this S-metric generates a metric. Especially we have $d_S(x, y) = 4d(x, y)$.*

The converse of the above proposition is not always true as seen in the following example.

Example 4 Let $X = \mathbb{R}$ and define the function $\mathscr{S} : X \times X \times X \to [0, \infty)$ by

$$\mathscr{S}(x, y, z) = \left|x^5 - z^5\right| + \left|x^5 + z^5 - 2y^5\right|,$$

for all $x, y, z \in \mathbb{R}$. Then the function $\mathscr{S}$ is an S-metric which is not generated by any metric and $(X, \mathscr{S})$ is an S-metric space. Now we show that this S-metric is not generated by any metric d, that is, $\mathscr{S} \neq \mathscr{S}_d$. On the contrary, we assume that there exists a metric d such that

$$\mathscr{S}(x, y, z) = \mathscr{S}_d(x, y, z) = d(x, z) + d(y, z),$$

for all $x, y, z \in \mathbb{R}$. Therefore we have

$$\mathscr{S}(x, x, z) = 2d(x, z) \text{ and so } d(x, z) = \left|x^5 - z^5\right|$$

and

$$\mathscr{S}(y, y, z) = 2d(y, z) \text{ and so } d(y, z) = \left|y^5 - z^5\right|,$$

for all $x, y, z \in \mathbb{R}$. Then we get

$$\left|x^5 - z^5\right| + \left|x^5 + z^5 - 2y^5\right| = \left|x^5 - z^5\right| + \left|y^5 - z^5\right|,$$

which is a contradiction for $x = 1, y = 2, z = 0 \in \mathbb{R}$. Consequently, $\mathscr{S} \neq \mathscr{S}_d$, that is, the S-metric is not generated by any metric d. However, this S-metric generates a metric d_S such that

$$d_S(x, y) = \mathscr{S}(x, x, y) + \mathscr{S}(y, y, x) = 2\mathscr{S}(x, x, y) = 2\left|x^5 - z^5\right|,$$

for all $x, y \in \mathbb{R}$.

Remark 1

1) For an S-metric which is not generated by any metric, d_S can be or can not be a metric on X (see Examples 3 and 4).
2) Let $X \neq \emptyset$, $\mathscr{S}_1$ be an S-metric on X which is not generated by any metric d and $\mathscr{S}_2$ be an S-metric on X which is generated by any metric d. Then d_{S_1} and d_{S_2} may be the same (see [39] for more details).

Definition 4 ([43]) Let $(X, \mathscr{S})$ be an S-metric space and $A \subset X$. A subset A of X is said to be S-bounded if there exists $r > 0$ such that

$$\mathscr{S}(x, x, y) < r,$$

for all $x, y \in A$.

Definition 5 ([17, 32]) Let $(X, \mathscr{S})$ be an S-metric space, $x \in X$ and $A \subset X$. The diameter of A is defined by

$$\delta(A) = \sup\{\mathscr{S}(x, x, y) : x, y \in A\}.$$

If A is S-bounded, then we will write $\delta(A) < \infty$.

Now we recall some basic facts on an S_b-metric space.

Definition 6 ([44]) Let $(X, \mathscr{S}_b)$ be an S_b-metric space.

1. A sequence $\{x_n\}$ in X converges to x if and only if $\mathscr{S}_b(x_n, x_n, x) \to 0$ as $n \to \infty$, that is, for each $\varepsilon > 0$ there exists $n_0 \in \mathbb{N}$ such that $\mathscr{S}_b(x_n, x_n, x) < \varepsilon$ for all $n \geq n_0$. It is denoted by
$$\lim_{n \to \infty} x_n = x.$$
2. A sequence $\{x_n\}$ in X is called a Cauchy sequence if for each $\varepsilon > 0$ there exists $n_0 \in \mathbb{N}$ such that $\mathscr{S}_b(x_n, x_n, x_m) < \varepsilon$ for each $n, m \geq n_0$.
3. An S_b-metric space $(X, \mathscr{S}_b)$ is said to be complete if every Cauchy sequence in X is convergent.

Definition 7 ([48]) Let $(X, \mathscr{S}_b)$ be an S_b-metric space, $x \in X$ and $A \subset X$.

(a) A subset A of X is said to be S_b-bounded if there exists $r > 0$ such that

$$\mathscr{S}_b(x, x, y) < r,$$

for all $x, y \in A$.

(b) The diameter of A is defined by

$$\delta(A) = \sup\{\mathscr{S}_b(x, x, y) : x, y \in A\}.$$

If A is S_b-bounded, then we will write $\delta(A) < \infty$.

Lemma 3 ([44]) *Let $(X, \mathscr{S}_b)$ be an S_b-metric space with $b \geq 1$, then we have*

$$\mathscr{S}_b(x, x, y) \leq b\mathscr{S}_b(y, y, x)$$

and

$$\mathscr{S}_b(y, y, x) \leq b\mathscr{S}_b(x, x, y).$$

Definition 8 ([48]) Let $(X, \mathscr{S}_b)$ be an S_b-metric space and $b > 1$. An S_b-metric $\mathscr{S}_b$ is called symmetric if

$$\mathscr{S}_b(x, x, y) = \mathscr{S}_b(y, y, x), \tag{5}$$

for all $x, y \in X$.

For $b = 1$, the symmetry condition (5) is satisfied by Lemma 1.

Lemma 4 ([48]) *Let $(X, \mathscr{S}_b)$ be an S_b-metric space. If the sequence $\{x_n\}$ in X converges to x then x is unique.*

In the following lemmas, we see the relationships between a b-metric and an S_b-metric.

Lemma 5 ([48]) *Let $(X, \mathscr{S}_b)$ be an S_b-metric space, $\mathscr{S}_b$ be a symmetric S_b-metric with $b \geq 1$ and the function $d : X \times X \to [0, \infty)$ be defined by*

$$d(x, y) = \mathscr{S}_b(x, x, y),$$

for all $x, y \in X$. Then d is a b-metric on X.

Lemma 6 ([48]) *Let (X, d) be a b-metric space with $b \geq 1$ and the function $\mathscr{S}_b : X \times X \times X \to [0, \infty)$ be defined by*

$$\mathscr{S}_b(x, y, z) = d(x, z) + d(y, z),$$

for all $x, y, z \in X$. Then $\mathscr{S}_b$ is an S_b-metric on X.

Now we recall the following definitions, propositions, corollaries and theorems on S-metric spaces. At first, we consider the Rhoades' contractive condition and its generalizations on an S-metric space.

Definition 9 ([32]) Let $(X, \mathscr{S})$ be an S-metric space and T be a self-mapping of X. We define

$$(\mathbf{S25}) \quad \mathscr{S}(Tx, Tx, Ty) < max\{\mathscr{S}(x, x, y), \mathscr{S}(Tx, Tx, x), \mathscr{S}(Ty, Ty, y), \mathscr{S}(Ty, Ty, x), \mathscr{S}(Tx, Tx, y)\},$$

for each $x, y \in X, x \neq y$.

Definition 10 ([31]) Let $(X, \mathscr{S})$ be an S-metric space and T be a self-mapping of X. We define the contractive conditions (**S50**), (**S75**), (**S100**) and (**S125**) as follows:

(**S50**) There exists a positive integer p such that

$$\mathscr{S}(T^p x, T^p x, T^p y) < \max\{\mathscr{S}(x, x, y), \mathscr{S}(T^p x, T^p x, x), \mathscr{S}(T^p y, T^p y, y), \mathscr{S}(T^p y, T^p y, x), \mathscr{S}(T^p x, T^p x, y)\},$$

for any $x, y \in X, x \neq y$.

(**S75**) There exist positive integers p, q such that

$$\mathscr{S}(T^px, T^px, T^qy) < \max\{\mathscr{S}(x, x, y), \mathscr{S}(T^px, T^px, x), \mathscr{S}(T^qy, T^qy, y), \mathscr{S}(T^qy, T^qy, x), \mathscr{S}(T^px, T^px, y)\},$$

for any $x, y \in X, x \neq y$.

(**S100**) For any given $x \in X$ there exists a positive integer $p(x)$ such that

$$\mathscr{S}(T^{p(x)}x, T^{p(x)}x, T^{p(x)}y) < \max\{\mathscr{S}(x, x, y), \mathscr{S}(T^{p(x)}x, T^{p(x)}x, x), \mathscr{S}(T^{p(x)}y, T^{p(x)}y, y), \mathscr{S}(T^{p(x)}y, T^{p(x)}y, x), \mathscr{S}(T^{p(x)}x, T^{p(x)}x, y)\},$$

for any $y \in X, x \neq y$.

(**S125**) For any given $x, y \in X, x \neq y$ there exists a positive integer $p(x, y)$ such that

$$\mathscr{S}(T^{p(x,y)}x, T^{p(x,y)}x, T^{p(x,y)}y) < \max\{\mathscr{S}(x, x, y), \mathscr{S}(T^{p(x,y)}x, T^{p(x,y)}x, x), \mathscr{S}(T^{p(x,y)}y, T^{p(x,y)}y, y),\mathscr{S}(T^{p(x,y)}y,T^{p(x,y)}y, x), \mathscr{S}(T^{p(x,y)}x, T^{p(x,y)}x, y)\}.$$

We recall some properties of the Rhoades' contractive conditions and fixed-point results on an S-metric space.

Proposition 2 ([31]) *Let $(X, \mathscr{S})$ be an S-metric space and T be a self-mapping of X. We obtain the following implications* :

$$(\mathbf{S25}) \Longrightarrow (\mathbf{S50}) \Longrightarrow (\mathbf{S75}) \text{ and } (\mathbf{S50}) \Longrightarrow (\mathbf{S100}) \Longrightarrow (\mathbf{S125}).$$

Theorem 4 ([31]) *Let $(X, \mathscr{S})$ be an S-metric space, T be a self-mapping of X which satisfies the inequality* (**S125**). *If T has a fixed point then it is unique.*

Corollary 1 ([31]) *Let $(X, \mathscr{S})$ be an S-metric space, T be a self-mapping of X and the inequality* (**S25**) *(resp. $T \in$* (**S50**), *$T \in$* (**S100**)) *be satisfied. If T has a fixed point then it is unique.*

Corollary 2 ([31]) *Let $(X, \mathscr{S})$ be an S-metric space, T be a self-mapping of X and the inequality* (**S75**) *be satisfied. If T has a fixed point then it is unique.*

Definition 11 ([32]) Let $(X, \mathscr{S})$ be an S-metric space and T be a self-mapping of X. T is called a C_S-mapping on X if for each $x \in X$ and each positive integer $n \geq 2$ satisfying

$$T^ix \neq T^jx, 0 \leq i < j \leq n-1, \tag{6}$$

we have

$$\mathscr{S}(T^nx, T^nx, T^ix) < \max_{1\leq j\leq n} \{\mathscr{S}(T^jx, T^jx, x)\}, i = 1, 2, \ldots, n-1.$$

Definition 12 ([32]) Let $(X, \mathscr{S})$ be an S-metric space and T be a self-mapping of X. T is called an L_S-mapping on X if for each $x \in X$ and each positive integer $n \geq 2$ with the condition (6) we have

$$\mathscr{S}(T^n x, T^n x, T^i x) < \max_{0\leq p<q\leq n} \{\mathscr{S}(T^p x, T^p x, T^q x)\}, i = 1, 2, \ldots, n-1.$$

Theorem 5 ([32]) *Let $(X, \mathscr{S})$ be an S-metric space and T be a self-mapping of X. If T satisfies the condition* (**S25**)*, then T is a C_S-mapping.*

Proposition 3 ([32]) *Let $(X, \mathscr{S})$ be an S-metric space. Then the notions of a C_S-mapping and of an L_S-mapping are equivalent.*

Theorem 6 ([32]) *Let T be a C_S-mapping from an S-metric space $(X, \mathscr{S})$ into itself. Then T has a fixed point in X if and only if there exist integers p and q, $p > q \geq 0$ and $x \in X$ satisfying*

$$T^p x = T^q x. \tag{7}$$

If the condition (7) is satisfied, then $T^q x$ is a fixed point of T.

Corollary 3 ([32]) *Let $(X, \mathscr{S})$ be an S-metric space and T be a self-mapping of X satisfying the condition* (**S25**)*. Then T has a fixed point in X if and only if there exist integers p and q, $p > q \geq 0$ and $x \in X$ satisfying the condition (7). Then $T^q x$ is the fixed point of T.*

Theorem 7 ([32]) *Let T be an L_S-mapping from an S-metric space $(X, \mathscr{S})$ into itself. Then T has a fixed point in X if and only if there exist integers p and q, $p > q \geq 0$ and $x \in X$ satisfying the condition (7). Then $T^q x$ is a fixed point of T.*

Theorem 8 ([31]) *Let $(X, \mathscr{S})$ be an S-metric space, $x \in X$, T be a self-mapping of X and the inequality* (**S125**) *be satisfied. Assume that x is a periodic point of T with periodic index m. Then T has a fixed point x in $\{T^n x\}(n \geq 0)$ if and only if for any $T^{n_1} x$, $T^{n_2} x \in \{T^n x\}(n \geq 0)$, $T^{n_1} x \neq T^{n_2} x$, there exist $T^{n_3} x$, $T^{n_4} x \in \{T^n x\}$ such that*

$$T^{p(T^{n_3} x, T^{n_4} x)}(T^{n_3} x) = T^{n_1} x \text{ and } T^{p(T^{n_3} x, T^{n_4} x)}(T^{n_4} x) = T^{n_2} x.$$

Then the point x is the unique fixed point of T in X.

Corollary 4 ([31]) *Let $(X, \mathscr{S})$ be an S-metric space, T be a self-mapping of X, the inequality* (**S100**) *be satisfied and $x \in X$ be a periodic point of T. Then the following conditions are equivalent:*

1. *T has a unique fixed point in $\{T^n x\}(n \geq 0)$,*
2. *There exists $T^{n_0} x \in \{T^n x\}(n \geq 0)$ such that*

$$T^{p(T^{n_0} x)}(T^{n_0} x) = T^{n_1} x,$$

for any $T^{n_1} x \in \{T^n x\}(n \geq 0)$, where $p(T^{n_0} x)$ is a positive integer.

Then the point x is the unique fixed point of T in X.

Corollary 5 ([31]) *Let $(X, \mathscr{S})$ be an S-metric space, T be a self-mapping of X, the inequality* (**S75**) *be satisfied and $x \in X$ be a periodic point of T. Then x is the unique fixed point of T if there exist $T^{n_3}x$, $T^{n_4}x \in \{T^n x\}(n \geq 0)$, $T^{n_3}x \neq T^{n_4}x$ such that*

$$T^p(T^{n_3}x) = T^{n_1}x \text{ and } T^q(T^{n_4}x) = T^{n_2}x,$$

for any $T^{n_1}x$, $T^{n_2}x \in \{T^n x\}(n \geq 0)$, $T^{n_1}x \neq T^{n_2}x$, where p and q are positive integers.

Corollary 6 ([31]) *Let $(X, \mathscr{S})$ be an S-metric space, T be a self-mapping of X and the inequality* (**S50**) *be satisfied. Then the following conditions are equivalent:*

1. *T has a fixed point in X,*
2. *There exists a periodic point $x \in X$ of T.*

Then the point x is the unique fixed point of T in X.

Definition 13 ([32]) Let $(X, \mathscr{S})$ be an S-metric space, $x, y \in X$, T be a self-mapping of X, $U_x = \{T^n x : n \in \mathbb{N}\}$, $diam\{U_x\} < \infty$ and $diam\{U_y\} < \infty$. We define

$$(\mathbf{S25a}) \quad \mathscr{S}(Tx, Tx, Ty) < diam\{U_x \cup U_y\},$$

for each $x, y \in X$ with $x \neq y$.

Proposition 4 ([32]) *Let $(X, \mathscr{S})$ be an S-metric space and T be a self-mapping of X. If T satisfies the condition* (**S25**)*, then T satisfies the condition* (**S25a**).

Theorem 9 ([32]) *Let T be a continuous self-mapping from a compact S-metric space $(X, \mathscr{S})$ into itself and T satisfies the condition* (**S25a**)*. Then T has a unique fixed point.*

Corollary 7 ([32]) *Let T be a continuous self-mapping from a compact S-metric space $(X, \mathscr{S})$ into itself and T satisfies the condition* (**S25**)*. Then T has a unique fixed point.*

3 New Generalizations of Rhoades' Contractive Conditions

In this section we consider S_b-metric spaces and present new contractive conditions such as ($\mathbf{S}_b\mathbf{25}$), ($\mathbf{S}_b\mathbf{50}$), ($\mathbf{S}_b\mathbf{75}$), ($\mathbf{S}_b\mathbf{100}$) and ($\mathbf{S}_b\mathbf{125}$) as the generalizations of the Rhoades' contractive conditions mentioned in the previous sections. We investigate some properties of these new contractive conditions and give some illustrative examples.

At first we generalize the Rhoades' conditions given in Sect. 2.

Let $(X, \mathscr{S}_b)$ be an S_b-metric space with $b \geq 1$ and T be a self-mapping of X.

($\mathbf{S}_b\mathbf{25}$) For any $x, y \in X$ with $x \neq y$

$$\mathcal{S}_b(Tx, Tx, Ty) < max\{\mathcal{S}_b(x, x, y), \mathcal{S}_b(Tx, Tx, x), \mathcal{S}_b(Ty, Ty, y),\\ \mathcal{S}_b(Ty, Ty, x), \mathcal{S}_b(Tx, Tx, y)\}.$$

(**S**$_b$**50**) There exists a positive integer p such that

$$\mathcal{S}_b(T^px, T^px, T^py) < \max\{\mathcal{S}_b(x, x, y), \mathcal{S}_b(T^px, T^px, x), \mathcal{S}_b(T^py, T^py, y),\\ \mathcal{S}_b(T^py, T^py, x), \mathcal{S}_b(T^px, T^px, y)\},$$

for any $x, y \in X$ with $x \neq y$.

(**S**$_b$**75**) There exist positive integers p, q such that

$$\mathcal{S}_b(T^px, T^px, T^qy) < \max\{\mathcal{S}_b(x, x, y), \mathcal{S}_b(T^px, T^px, x), \mathcal{S}_b(T^qy, T^qy, y),\\ \mathcal{S}_b(T^qy, T^qy, x), \mathcal{S}_b(T^px, T^px, y)\},$$

for any $x, y \in X$ with $x \neq y$.

(**S**$_b$**100**) For any given $x \in X$ there exists a positive integer $p(x)$ such that

$$\mathcal{S}_b(T^{p(x)}x, T^{p(x)}x, T^{p(x)}y) < \max\{\mathcal{S}_b(x, x, y), \mathcal{S}_b(T^{p(x)}x, T^{p(x)}x, x),\\ \mathcal{S}_b(T^{p(x)}y, T^{p(x)}y, y), \mathcal{S}_b(T^{p(x)}y, T^{p(x)}y, x),\\ \mathcal{S}_b(T^{p(x)}x, T^{p(x)}x, y)\},$$

for any $y \in X$ with $x \neq y$.

(**S**$_b$**125**) For any given $x, y \in X$ with $x \neq y$ there exists a positive integer $p(x, y)$ such that

$$\mathcal{S}_b(T^{p(x,y)}x, T^{p(x,y)}x, T^{p(x,y)}y) < \max\{\mathcal{S}_b(x, x, y), \mathcal{S}_b(T^{p(x,y)}x, T^{p(x,y)}x, x),\\ \mathcal{S}_b(T^{p(x,y)}y, T^{p(x,y)}y, y), \mathcal{S}_b(T^{p(x,y)}y,\\ T^{p(x,y)}y, x),\\ \mathcal{S}_b(T^{p(x,y)}x, T^{p(x,y)}x, y)\}.$$

For $b = 1$, we note that the condition (**S**$_b$**25**) (resp. (**S**$_b$**50**), (**S**$_b$**75**), (**S**$_b$**100**) and (**S**$_b$125)) coincides with the condition (**S25**) (resp. (**S50**), (**S75**), (**S100**) and (**S125**)).

Proposition 5 *Let* $(X, \mathcal{S}_b)$ *be an* S_b*-metric space with* $b \geq 1$ *and* T *be a self-mapping of* X*. We obtain the following implications:*

(**S**$_b$**25**) $\Longrightarrow$ (**S**$_b$**50**) $\Longrightarrow$ (**S**$_b$**75**) *and* (**S**$_b$**50**) $\Longrightarrow$ (**S**$_b$**100**) $\Longrightarrow$ (**S**$_b$**125**).

Proof Let T satisfies the condition (**S**$_b$**25**). Then we have

$$\mathscr{S}_b(Tx, Tx, Ty) < max\{\mathscr{S}_b(x, x, y), \mathscr{S}_b(Tx, Tx, x), \mathscr{S}_b(Ty, Ty, y),$$
$$\mathscr{S}_b(Ty, Ty, x), \mathscr{S}_b(Tx, Tx, y)\},$$

for any $x, y \in X$ with $x \neq y$. If we take $p = 1$ then the condition (**S**$_b$**50**) is clearly satisfied.

Assume that T satisfies the condition (**S**$_b$**50**). Hence there exists a positive integer p such that

$$\mathscr{S}_b(T^px, T^px, T^py) < \max\{\mathscr{S}_b(x, x, y), \mathscr{S}_b(T^px, T^px, x), \mathscr{S}_b(T^py, T^py, y),$$
$$\mathscr{S}_b(T^py, T^py, x), \mathscr{S}_b(T^px, T^px, y)\},$$

for any $x, y \in X$ with $x \neq y$. If we take $p = q$ then the condition (**S**$_b$**75**) is satisfied. Also T satisfies the condition (**S**$_b$**100**) taking $p = p(x)$.

Suppose that T satisfies the condition (**S**$_b$**100**). Hence there exists a positive integer $p(x)$ for any given $x \in X$ such that

$$\mathscr{S}_b(T^{p(x)}x, T^{p(x)}x, T^{p(x)}y) < \max\{\mathscr{S}_b(x, x, y), \mathscr{S}_b(T^{p(x)}x, T^{p(x)}x, x),$$
$$\mathscr{S}_b(T^{p(x)}y, T^{p(x)}y, y), \mathscr{S}_b(T^{p(x)}y, T^{p(x)}y, x),$$
$$\mathscr{S}_b(T^{p(x)}x, T^{p(x)}x, y)\},$$

for any $y \in X$ with $x \neq y$. If we take $p = p(x) = p(x, y)$ then the condition (**S**$_b$**125**) is satisfied. The proof is completed.

The converses of the above implications in Proposition 5 are not always true as we have seen in the following examples.

Example 5 Let us consider the S_b-metric defined in Example 1 on $X = [0, 1]$. Let

$$Tx = \begin{cases} 0 \ ; \ x \in [0, 1]\,, x \neq \frac{1}{6} \\ 1 \ ; \qquad x = \frac{1}{6} \end{cases},$$

for all $x \in X$. Then T is a self-mapping on the S_b-metric space (X, S_b) and satisfies the condition (**S**$_b$**50**) for $p = 2$. But T does not satisfy the condition (**S**$_b$**25**). Indeed, for $x = \frac{1}{2}$, $y = \frac{1}{6} \in [0, 1]$ we obtain

$$\mathscr{S}_b(Tx, Tx, Ty) = \mathscr{S}_b\,(0, 0, 1) = \frac{1}{4},$$

$$\mathscr{S}_b(x, x, y) = \mathscr{S}_b\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{6}\right) = \frac{1}{36},$$

$$\mathscr{S}_b(Tx, Tx, x) = \mathscr{S}_b\left(0, 0, \frac{1}{2}\right) = \frac{1}{16},$$

$$\mathscr{S}_b(Ty, Ty, y) = \mathscr{S}_b\left(1, 1, \frac{1}{6}\right) = \frac{25}{144},$$

$$\mathscr{S}_b(Ty, Ty, x) = \mathscr{S}_b\left(1, 1, \frac{1}{2}\right) = \frac{1}{16},$$

$$\mathscr{S}_b(Tx, Tx, y) = \mathscr{S}_b\left(0, 0, \frac{1}{6}\right) = \frac{1}{144}$$

and so we get

$$\mathscr{S}_b(Tx, Tx, Ty) = \frac{1}{4} < \max\left\{\frac{1}{36}, \frac{1}{16}, \frac{25}{144}, \frac{1}{16}, \frac{1}{144}\right\} = \frac{25}{144}.$$

Example 6 Let us consider the S_b-metric defined in Example 1 on $X = \{1, 2, 3\}$. Let

$$Tx = \begin{cases} x + 1 \ ; \ x \in \{1, 2\} \\ \quad 2 \quad ; \quad x = 3 \end{cases},$$

for all $x \in X$. Then T is a self-mapping on the S_b-metric space $(X, \mathscr{S}_b)$ and satisfies the condition ($\mathbf{S}_b\mathbf{75}$) for $p = 1$ and $q = 2$. But T does not satisfy the condition ($\mathbf{S}_b\mathbf{50}$). Indeed, let us choose $x = 2$ and $y = 3$.

For $p = 1$ we have

$$\mathscr{S}_b(Tx, Tx, Ty) = \mathscr{S}_b\,(3, 3, 2) = \frac{1}{4},$$

$$\mathscr{S}_b(x, x, y) = \mathscr{S}_b\,(2, 2, 3) = \frac{1}{4},$$

$$\mathscr{S}_b(Tx, Tx, x) = \mathscr{S}_b\,(3, 3, 2) = \frac{1}{4},$$

$$\mathscr{S}_b(Ty, Ty, y) = \mathscr{S}_b\,(2, 2, 3) = \frac{1}{4},$$

$$\mathscr{S}_b(Ty, Ty, x) = \mathscr{S}_b\,(2, 2, 2) = 0,$$

$$\mathscr{S}_b(Tx, Tx, y) = \mathscr{S}_b\,(3, 3, 3) = 0$$

and so we get

$$\mathscr{S}_b(Tx, Tx, Ty) = \frac{1}{4} < \max\left\{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, 0, 0\right\} = \frac{1}{4}.$$

Hence the condition ($\mathbf{S}_b\mathbf{50}$) is not satisfied.

For $p = 2$ we have

$$\mathscr{S}_b(T^2x, T^2x, T^2y) = \mathscr{S}_b\,(2, 2, 3) = \frac{1}{4},$$

$$\mathscr{S}_b(x, x, y) = \mathscr{S}_b\,(2, 2, 3) = \frac{1}{4},$$

$$\mathscr{S}_b(T^2x, T^2x, x) = \mathscr{S}_b\,(2, 2, 2) = 0,$$

$$\mathscr{S}_b(T^2y, T^2y, y) = \mathscr{S}_b\,(3, 3, 3) = 0,$$

$$\mathscr{S}_b(T^2y, T^2y, x) = \mathscr{S}_b\,(3, 3, 2) = \frac{1}{4},$$

$$\mathscr{S}_b(T^2x, T^2x, y) = \mathscr{S}_b\,(2, 2, 3) = \frac{1}{4}$$

and so we get

$$\mathscr{S}_b(T^2x, T^2x, T^2y) = \frac{1}{4} < \max\left\{\frac{1}{4}, 0, 0, \frac{1}{4}, \frac{1}{4}\right\}.$$

Hence the condition ($\mathbf{S}_b\mathbf{50}$) is not satisfied. For $p \geq 3$ using similar arguments we can easily see that the condition ($\mathbf{S}_b\mathbf{50}$) is not satisfied.

Example 7 Let us consider the self-mapping T defined in Figure 4 on page 105 in [3] and the S_b-metric defined in Example 1. If we choose $x = \left(\frac{1}{n} + 1, 0\right)$, $y = \left(\frac{1}{n}, 0\right)$ for each n then the condition ($\mathbf{S}_b\mathbf{50}$) is not satisfied. A positive integer $p(x)$ can be chosen for any given $x \in X$ such that the condition ($\mathbf{S}_b\mathbf{100}$) is satisfied.

Example 8 Let $X = [0, 1] \cup \{5\}$ be the S_b-metric space with the S_b-metric defined in Example 1 and let

$$Tx = \begin{cases} \sqrt{x} \ ; & x \in [0, 1]\,, x \neq \frac{1}{4}, x \neq \frac{1}{5} \\ \frac{1}{5} \ ; & x = \frac{1}{4} \\ 5 \ ; & x = \frac{1}{5} \\ \frac{1}{4} \ ; & x = 5 \end{cases},$$

for all $x \in X$. Then T is a self-mapping on the S_b-metric space $(X, \mathscr{S}_b)$ and satisfies the condition ($\mathbf{S}_b\mathbf{125}$). But T does not satisfy the condition ($\mathbf{S}_b\mathbf{100}$).

Theorem 10 *Let $(X, \mathscr{S}_b)$ be an S_b-metric space, $\mathscr{S}_b$ be a symmetric S_b-metric with $b > 1$ and T be a self-mapping of X which satisfies the condition* $(\mathbf{S}_b\mathbf{125})$. *If T has a fixed point then it is unique.*

Proof Suppose that x and y are the fixed points of T such that $x, y \in X$ with $x \neq y$. Then there exists a positive integer $p = p(x, y)$ such that

$$\mathscr{S}_b(T^px, T^px, T^py) < \max\{\mathscr{S}_b(x, x, y), \mathscr{S}_b(T^px, T^px, x), \mathscr{S}_b(T^py, T^py, y), \mathscr{S}_b(T^py, T^py, x), \mathscr{S}_b(T^px, T^px, y)\},$$

by the condition $(\mathbf{S}_b\mathbf{125})$. Then by using the symmetry condition and the fact that $T^px = x$, $T^py = y$ we get

$$\mathscr{S}_b(T^px, T^px, T^py) = \mathscr{S}_b(x, x, y) < \mathscr{S}_b(x, x, y),$$

which is a contradiction. Consequently, the fixed point is unique.

For $b = 1$, Theorem 10 coincides with Theorem 4.

Following the ideas used in the proof of Theorem 10 we obtain the following corollary by Proposition 5.

Corollary 8 *Let $(X, \mathscr{S}_b)$ be an S_b-metric space, $\mathscr{S}_b$ be a symmetric S_b-metric with $b > 1$ and T be a self-mapping of X which satisfies the condition* $(\mathbf{S}_b\mathbf{25})$ *(resp.* $(\mathbf{S}_b\mathbf{50})$, $(\mathbf{S}_b\mathbf{75})$ *and* $(\mathbf{S}_b\mathbf{100})$*). If T has a fixed point then it is unique.*

For $b = 1$, Corollary 8 coincides with Corollaries 1 and 2.

If we consider the following example then it is clear that the symmetry condition have an important role for the self-mappings satisfying the condition $(\mathbf{S}_b\mathbf{125})$.

Example 9 Let $X = \mathbb{R}$ and the function $\mathscr{S}_b : X \times X \times X \to [0, \infty)$ be defined as

$$\begin{aligned} &\mathscr{S}_b(0, 0, 1) = 2, \\ &\mathscr{S}_b(1, 1, 0) = 4, \\ &\mathscr{S}_b(x, y, z) = 0 \text{ if } x = y = z, \\ &\mathscr{S}_b(x, y, z) = 1 \text{ otherwise,} \end{aligned}$$

for all $x, y, z \in \mathbb{R}$ [48]. Then the function $\mathscr{S}_b$ is an S_b-metric with $b \geq 2$ which is not symmetric and $(\mathbb{R}, \mathscr{S}_b)$ is an S_b-metric space. If we consider the self-mapping $T : \mathbb{R} \to \mathbb{R}$ defined by

$$Tx = \begin{cases} 1 \ ; \ x \in \{0, 1\} \\ 6 \ ; \ \text{otherwise} \end{cases},$$

for all $x \in \mathbb{R}$, then T has two fixed points $x_1 = 1$, $x_2 = 6$ and T does not satisfy the condition $(\mathbf{S}_b\mathbf{125})$ for the fixed points.

In order to obtain a fixed point theorem for a self-mapping of X satisfying the condition $(\mathbf{S}_b\mathbf{25})$, we give the following definition.

Definition 14 Let $(X, \mathscr{S}_b)$ be an S_b-metric space with $b \geq 1$ and T be a self-mapping of X.

1. For each $x \in X$ and each positive integer $n \geq 2$ satisfying
$$T^i x \neq T^j x \ (0 \leq i < j \leq n-1), \tag{8}$$
if we have
$$\mathscr{S}_b(T^n x, T^n x, T^i x) < \max_{1 \leq j \leq n} \left\{ \mathscr{S}_b(T^j x, T^j x, x) \right\} \ (i = 1, 2, \ldots, n-1), \tag{9}$$
then T is called a C_S^b-mapping on X.
2. For each $x \in X$ and each positive integer $n \geq 2$ with the condition (8) if we have
$$\mathscr{S}_b(T^n x, T^n x, T^i x) < \max_{1 \leq p < q \leq n} \left\{ \mathscr{S}_b(T^p x, T^p x, T^q x) \right\} \ (i = 1, 2, \ldots, n-1),$$
then T is called an L_S^b-mapping on X.

If we consider the case $b = 1$ then the notion of a C_S^b-mapping (resp. an L_S^b-mapping) coincides with the notion of a C_S-mapping (resp. an L_S-mapping).

We note that the symmetry condition (5) is not necessary in the following proposition and the case $b = 1$ was proved in Theorem 5.

Proposition 6 *Let $(X, \mathscr{S}_b)$ be an S_b-metric space with $b \geq 1$ and T be a self-mapping of X. If T satisfies the condition* $(\mathbf{S}_b\mathbf{25})$ *then T is a C_S^b-mapping.*

Proof Using mathematical induction and the condition $(\mathbf{S}_b\mathbf{25})$ the proof follows easily.

The converse of the above proposition is not always true. We give the following example.

Example 10 Let $X = [0, 1] \cup \{2, 5, 8\}$ be the S_b-metric space with the S_b-metric defined in Example 1 and let

$$Tx = \begin{cases} x & ; x \in [0, 1] \\ x - 3 & ; x \in \{5, 8\} \\ 1 & ; \quad x = 2 \end{cases},$$

for all $x \in X$. Then T is a C_S-mapping. Indeed, we have the following cases for $x \in \{2, 5, 8\}$.

Case 1 For $x = 2$ and $n = 2$ we get

$$\mathscr{S}_b(T^2 2, T^2 2, T2) = 0 < \max \left\{ \mathscr{S}_b(T^2 2, T^2 2, 2), \mathscr{S}_b(T2, T2, 2) \right\} = \frac{1}{4}.$$

For $n > 2$ using similar arguments it can be seen that (9) holds.

Case 2 For $x = 5$ and $n \in \{2, 3\}$ we get

$$\mathcal{S}_b(T^2 5, T^2 5, T5) = \frac{1}{4} < \max\left\{\mathcal{S}_b(T^2 5, T^2 5, 5), \mathcal{S}_b(T5, T5, 5)\right\} = 4$$

and

$$\max\left\{\mathcal{S}_b(T^3 5, T^3 5, T5), \mathcal{S}_b(T^3 5, T^3 5, T^2 5)\right\} = \frac{1}{4} <$$
$$\max\left\{\mathcal{S}_b(T^3 5, T^3 5, 5), \mathcal{S}_b(T^2 5, T^2 5, 5), \mathcal{S}_b(T5, T5, 5)\right\} = 4.$$

For $n > 3$ using similar arguments it can be seen that (9) holds.

Case 3 For $x = 8$ and $n \in \{2, 3, 4\}$ we get

$$\mathcal{S}_b(T^2 8, T^2 8, T8) = \frac{9}{4} < \max\left\{\mathcal{S}_b(T^2 8, T^2 8, 8), \mathcal{S}_b(T8, T8, 8)\right\} = 9,$$

$$\max\left\{\mathcal{S}_b(T^3 8, T^3 8, T8), \mathcal{S}_b(T^3 8, T^3 8, T^2 8)\right\} = 4 <$$
$$\max\left\{\mathcal{S}_b(T^3 8, T^3 8, 8), \mathcal{S}_b(T^2 8, T^2 8, 8), \mathcal{S}_b(T8, T8, 8)\right\} = \frac{49}{4}$$

and

$$\max\left\{\mathcal{S}_b(T^4 8, T^4 8, T8), \mathcal{S}_b(T^4 8, T^4 8, T^2 8), \mathcal{S}_b(T^4 8, T^4 8, T^3 8)\right\} = 4 <$$
$$\max\left\{\mathcal{S}_b(T^4 8, T^4 8, 8), \mathcal{S}_b(T^3 8, T^3 8, 8), \mathcal{S}_b(T^2 8, T^2 8, 8), \mathcal{S}_b(T8, T8, 8)\right\} = \frac{49}{4}.$$

For $n > 4$ using similar arguments it can be seen that (9) holds.

But T does not satisfy the condition (**S**$_b$**25**). Indeed, for $x = \frac{1}{4}, y = \frac{1}{5} \in [0, 1]$ we have

$$\mathcal{S}_b(Tx, Tx, Ty) = \mathcal{S}_b\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{5}\right) = \frac{1}{1600},$$

$$\mathcal{S}_b(x, x, y) = \mathcal{S}_b\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{5}\right) = \frac{1}{1600},$$

$$\mathcal{S}_b(Tx, Tx, x) = \mathcal{S}_b(x, x, x) = 0,$$

$$\mathcal{S}_b(Ty, Ty, y) = \mathcal{S}_b(y, y, y) = 0,$$

$$\mathscr{S}_b(Ty, Ty, x) = \mathscr{S}_b\left(\frac{1}{5}, \frac{1}{5}, \frac{1}{4}\right) = \frac{1}{1600},$$

$$\mathscr{S}_b(Tx, Tx, y) = \mathscr{S}_b\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{5}\right) = \frac{1}{1600}$$

and so

$$\mathscr{S}_b(Tx, Tx, Ty) = \frac{1}{1600} < \max\left\{\frac{1}{1600}, 0, 0, \frac{1}{1600}, \frac{1}{1600}\right\} = \frac{1}{1600}.$$

In the following proposition, the symmetry condition (5) is necessary.

Proposition 7 *Let $(X, \mathscr{S}_b)$ be an S_b-metric space and $\mathscr{S}_b$ be a symmetric S_b-metric with $b > 1$. Then the notions of a C_S^b-mapping and an L_S^b-mapping are equivalent.*

Proof Let T be an L_S^b-mapping and $x \in X$. Suppose that the condition (8) is satisfied for each positive integer $n \geq 2$. Hence we have

$$min\{\mathscr{S}_b(T^i x, T^i x, T^j x) : 0 \leq i < j \leq k-1\} > 0,$$

where $2 \leq k \leq n$. It is obvious that

$$\mathscr{S}_b(T^n x, T^n x, T^i x) < \underset{0 \leq p < q \leq n}{max}\{\mathscr{S}_b(T^p x, T^p x, T^q x)\},$$

where $i = 1, 2, \ldots, n-1$. Let

$$U_n = \underset{1 \leq i \leq n-1}{max}\{\mathscr{S}_b(T^n x, T^n x, T^i x)\}$$

and

$$V_n = \underset{1 \leq i \leq n}{max}\{\mathscr{S}_b(T^i x, T^i x, x)\}.$$

Then using the symmetry condition (5) we obtain

$$\begin{aligned}
U_n &= max\{\mathscr{S}_b(T^n x, T^n x, T^i x) : 1 \leq i \leq n-1\} \\
&< max\{V_n, max\{\mathscr{S}_b(T^p x, T^p x, T^q x) : 1 \leq p < q \leq n-1\}\} \\
&= max\{U_{n-1}, V_n, max\{\mathscr{S}_b(T^p x, T^p x, T^q x) : 1 \leq p < q \leq n-2\}\} \\
&\leq max\{V_n, V_{n-1}, max\{\mathscr{S}_b(T^p x, T^p x, T^q x) : 1 \leq p < q \leq n-2\}\} \\
&= max\{V_n, max\{\mathscr{S}_b(T^p x, T^p x, T^q x) : 1 \leq p < q \leq n-2\}\} \\
&\leq \cdots \\
&\leq max\{V_n, max\{\mathscr{S}_b(T^p x, T^p x, T^q x) : 1 \leq p < q \leq 2\}\} \\
&= max\{V_n, \mathscr{S}_b(Tx, Tx, T^2 x)\} = max\{V_n, \mathscr{S}_b(T^2 x, T^2 x, Tx)\} \\
&\leq max\{V_n, max\{\mathscr{S}_b(Tx, Tx, x), \mathscr{S}_b(T^2 x, T^2 x, x)\}\} = V_n.
\end{aligned}$$

Therefore T is a C_S^b-mapping.

Conversely, let T be a C_S^b-mapping and $x \in X$. Suppose that the condition (8) is satisfied for each $n \geq 2$. Using the definition of a C_S^b-mapping we get

$$\mathscr{S}_b(T^n x, T^n x, T^i x) < \max_{1 \leq j \leq n} \{\mathscr{S}_b(T^j x, T^j x, x)\},$$

where $i = 1, 2, \ldots, n-1$. If $1 \leq j \leq n$ then $0 \leq j-1 \leq n-1$. Let us choose q such that $0 \leq j-1 < q \leq n$. For $j-1 = 0$ we have $1 \leq q \leq n$ and

$$\mathscr{S}_b(T^n x, T^n x, T^i x) < \max_{1 \leq q \leq n} \{\mathscr{S}_b(T^q x, T^q x, x)\}.$$

If we take $j-1 = p$ then using the symmetry condition (5) we obtain

$$\mathscr{S}_b(T^n x, T^n x, T^i x) < \max_{0 \leq p < q \leq n} \{\mathscr{S}_b(T^q x, T^q x, T^p x)\} = \max_{0 \leq p < q \leq n} \{\mathscr{S}_b(T^p x, T^p x, T^q x)\}.$$

Consequently, T is an L_S^b-mapping.

Notice that if we take $b = 1$ in Proposition 7 then we get Proposition 3.

Theorem 11 *Let $(X, \mathscr{S}_b)$ be an S_b-metric space, $\mathscr{S}_b$ be a symmetric S_b-metric with $b > 1$ and T be a C_S^b-mapping. Then T has a fixed point in X if and only if there exist integers p and q, $p > q \geq 0$ and $x \in X$ satisfying*

$$T^p x = T^q x. \tag{10}$$

If the condition (10) is satisfied then $T^q x$ is a fixed point of T.

Proof Let $x_0 \in X$ be a fixed point of T, that is, $Tx_0 = x_0$. It is obvious that the condition (10) is satisfied for $p = 1, q = 0$.

Conversely, suppose that there exist the integers p and q such that $p > q \geq 0$ and $x \in X$ satisfying

$$T^p x = T^q x.$$

Let p be the smallest integer such that $T^k x = T^q x$ with $k > q$. If we put $T^q x = y$ and $n = p - q$ we have

$$T^n y = T^n T^q x = T^{p-q+q} x = T^p x = T^q x = y$$

and so n is the minimal integer such that $T^n y = y$ for $n \geq 1$. We prove that y is a fixed point of T. Assume that y is not a fixed point of T. Then $n \geq 2$ and $T^i y \neq T^j y$ for $0 \leq i < j \leq n-1$. By the definition of a C_S^b-mapping and the symmetry condition (5), we obtain

$$\mathscr{S}_b(T^i y, T^i y, y) = \mathscr{S}_b(T^i y, T^i y, T^n y)$$

$$= \mathscr{S}_b(T^n y, T^n y, T^i y) < \underset{1\leq j\leq n}{max}\{\mathscr{S}_b(T^j y, T^j y, y)\}$$

$$= \underset{1\leq j\leq n-1}{max}\{\mathscr{S}_b(T^j y, T^j y, y)\},$$

where $i = 1, 2, \ldots, n-1$. Then we get

$$\underset{1\leq i\leq n-1}{max}\{\mathscr{S}_b(T^i y, T^i y, y)\} < \underset{1\leq j\leq n-1}{max}\{\mathscr{S}_b(T^j y, T^j y, y)\}.$$

This is a contradiction. Therefore $T^q x = y$ is a fixed point of T.

Corollary 9 *Let $(X, \mathscr{S}_b)$ be an S_b-metric space, $\mathscr{S}_b$ be a symmetric S_b-metric with $b > 1$ and T be a self-mapping of X satisfying the condition* $(\mathbf{S}_b\mathbf{25})$ *(or T be an L_S^b-mapping). Then T has a fixed point in X if and only if there exist integers p and q, $p > q \geq 0$ and $x \in X$ satisfying the condition (10). Then $T^q x$ is a fixed point of T.*

We note that the case $b = 1$ was proved in [32]. Hence Theorem 11 (resp. Corollary 9) generalizes Theorem 6 (resp. Corollary 3 and Theorem 7) given in [32].

Theorem 12 *Let $(X, \mathscr{S}_b)$ be an S_b-metric space with $b > 1$, T be a self-mapping of X satisfying the condition* $(\mathbf{S}_b\mathbf{125})$ *and $x \in X$. Suppose that x is a periodic point of T with periodic index m. Then T has a fixed point x in $\{T^n x\}$ $(n \geq 0)$ if and only if for any $T^{n_1}x$, $T^{n_2}x \in \{T^n x\}$ $(n \geq 0)$ with $T^{n_1}x \neq T^{n_2}x$, there exist $T^{n_3}x$, $T^{n_4}x \in \{T^n x\}$ such that*

$$T^{p(T^{n_3}x, T^{n_4}x)}(T^{n_3}x) = T^{n_1}x \textit{ and } T^{p(T^{n_3}x, T^{n_4}x)}(T^{n_4}x) = T^{n_2}x.$$

Then the point x is the unique fixed point of T in X.

Proof The proof of the if part can be easily seen from the definition of periodic index taking $T^{n_3}x = T^{n_1}x$, $T^{n_4}x = T^{n_2}x$ and $p(T^{n_3}x, T^{n_4}x) = m$. Now we prove the only if part. If x is a periodic point of T with periodic index m, then we have

$$\{T^n x\} = \{x, Tx, \ldots, T^{m-1}x\}.$$

Assume that $x \neq Tx$. Then there exist $T^{n_1}x$, $T^{n_2}x \in \{T^n x\}$ with $T^{n_1}x \neq T^{n_2}x$ such that

$$\delta(\{T^n x\}) = \max_{0\leq k,l\leq m-1,k\neq l}\{\mathscr{S}_b(T^k x, T^k x, T^l x)\} = \mathscr{S}_b(T^{n_1}x, T^{n_1}x, T^{n_2}x).$$

By the hypothesis there exist $T^{n_3}x$, $T^{n_4}x \in \{T^n x\}$ such that

$$T^{p(T^{n_3}x, T^{n_4}x)}(T^{n_3}x) = T^{n_1}x \text{ and } T^{p(T^{n_3}x, T^{n_4}x)}(T^{n_4}x) = T^{n_2}x.$$

Since $T^{n_1}x \neq T^{n_2}x$ we obtain $T^{n_3}x \neq T^{n_4}x$. So we obtain

$$\begin{aligned}
&\delta(\{T^n x\}) = \mathscr{S}_b(T^{n_1}x, T^{n_1}x, T^{n_2}x)\\
&= \mathscr{S}_b(T^{p(T^{n_3}x,T^{n_4}x)}(T^{n_3}x), T^{p(T^{n_3}x,T^{n_4}x)}(T^{n_3}x), T^{p(T^{n_3}x,T^{n_4}x)}(T^{n_4}x))\\
&< \max\{\mathscr{S}_b(T^{n_3}x, T^{n_3}x, T^{n_4}x), \mathscr{S}_b(T^{n_1}x, T^{n_1}x, T^{n_3}x), \mathscr{S}_b(T^{n_2}x, T^{n_2}x, T^{n_4}x),\\
&\mathscr{S}_b(T^{n_2}x, T^{n_2}x, T^{n_3}x), \mathscr{S}_b(T^{n_1}x, T^{n_1}x, T^{n_4}x)\}\\
&\leq \delta(\{T^n x\}),
\end{aligned}$$

which is a contradiction. Therefore we get $x = Tx$. The uniqueness of the fixed point x can be seen from Theorem 10.

Notice that the symmetry condition is not necessary in Theorem 12 and the case $b = 1$ was proved in Theorem 8.

Now we give the following corollaries as a result of Theorem 12.

Corollary 10 *Let $(X, \mathscr{S}_b)$ be an S_b-metric space with $b > 1$, T be a self-mapping of X satisfying the condition* ($\mathbf{S}_b\mathbf{100}$) *and $x \in X$ be a periodic point of T. Then the following conditions are equivalent*:

1. *T has a unique fixed point in $\{T^n x\}$ $(n \geq 0)$,*
2. *There exists $T^{n_0}x \in \{T^n x\}$ $(n \geq 0)$ such that*

$$T^{p(T^{n_0}x)}(T^{n_0}x) = T^{n_1}x,$$

for any $T^{n_1}x \in \{T^n x\}$ $(n \geq 0)$, where $p(T^{n_0}x)$ is a positive integer.

Then the point x is the unique fixed point of T in X.

Notice that the case $b = 1$ was given in Corollary 4.

Corollary 11 *Let $(X, \mathscr{S}_b)$ be an S_b-metric space with $b > 1$, T be a self-mapping of X satisfying the condition* ($\mathbf{S}_b\mathbf{75}$) *and $x \in X$ be a periodic point of T. Then x is the unique fixed point of T if there exist $T^{n_3}x$, $T^{n_4}x \in \{T^n x\}$ $(n \geq 0)$ with $T^{n_3}x \neq T^{n_4}x$ such that*

$$T^p(T^{n_3}x) = T^{n_1}x \text{ and } T^q(T^{n_4}x) = T^{n_2}x,$$

for any $T^{n_1}x$, $T^{n_2}x \in \{T^n x\}$ $(n \geq 0)$ with $T^{n_1}x \neq T^{n_2}x$, where p and q are positive integers.

Notice that the case $b = 1$ was given in Corollary 5.

Corollary 12 *Let $(X, \mathscr{S}_b)$ be an S_b-metric space with $b > 1$ and T be a self-mapping of X satisfying the condition* ($\mathbf{S}_b\mathbf{50}$). *Then the following conditions are equivalent*:

1. *T has a fixed point in X,*
2. *There exists a periodic point $x \in X$ of T.*

Then the point x is the unique fixed point of T in X.

Notice that the case $b = 1$ was given in Corollary 6.

Now we give a new contractive condition as a generalization of the condition ($\mathbf{S}_b\mathbf{25}$) using the notion of diameter on S_b-metric spaces.

Definition 15 Let $(X, \mathscr{S}_b)$ be an S_b-metric space with $b \geq 1$, $x, y \in X$, T be a self-mapping of X, $U_x = \{T^n x : n \in \mathbb{N}\}$, $diam\{U_x\} < \infty$ and $diam\{U_y\} < \infty$. We define

$$(\mathbf{S}_b\mathbf{25a}) \quad \mathscr{S}_b(Tx, Tx, Ty) < diam\{U_x \cup U_y\},$$

for each $x, y \in X$ with $x \neq y$.

Proposition 8 *Let $(X, \mathscr{S}_b)$ be an S_b-metric space with $b \geq 1$ and T be a self-mapping of X. If T satisfies the condition* ($\mathbf{S}_b\mathbf{25}$), *then T satisfies the condition* ($\mathbf{S}_b\mathbf{25a}$).

Proof It can be easily seen from the definitions of the conditions ($\mathbf{S}_b\mathbf{25}$) and ($\mathbf{S}_b\mathbf{25a}$).

The converse of Proposition 8 is not always true as we have seen in the following example.

Example 11 Let us consider the S_b-metric defined in Example 1 on $X = (0, 1)$. Let

$$Tx = \begin{cases} x \ ; \ x \in (0,1), x \neq \frac{1}{2}, x \neq \frac{1}{3} \\ \frac{1}{3} \ ; \qquad x = \frac{1}{2} \\ \frac{1}{2} \ ; \qquad x = \frac{1}{3} \end{cases},$$

for all $x \in X$. Then T is a self-mapping on the S_b-metric space $(X, \mathscr{S}_b)$ and satisfies the condition ($\mathbf{S}_b\mathbf{25a}$) since $\sup\{(0, 1)\} = 1$. But T does not satisfy the condition ($\mathbf{S}_b\mathbf{25}$). Indeed, for $x = \frac{1}{2}$, $y = \frac{1}{3} \in (0, 1)$ we obtain

$$\mathscr{S}_b(Tx, Tx, Ty) = \mathscr{S}_b\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{2}\right) = \frac{1}{144},$$

$$\mathscr{S}_b(x, x, y) = \mathscr{S}_b\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{3}\right) = \frac{1}{144},$$

$$\mathscr{S}_b(Tx, Tx, x) = \mathscr{S}_b\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{2}\right) = \frac{1}{144},$$

$$\mathscr{S}_b(Ty, Ty, y) = \mathscr{S}_b\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{3}\right) = \frac{1}{144},$$

$$\mathscr{S}_b(Ty, Ty, x) = \mathscr{S}_b\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right) = 0,$$

$$\mathscr{S}_b(Tx, Tx, y) = \mathscr{S}_b\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) = 0$$

and so we get

$$\mathscr{S}_b(Tx, Tx, Ty) = \frac{1}{144} < \max\left\{\frac{1}{144}, \frac{1}{144}, \frac{1}{144}, 0, 0\right\} = \frac{1}{144}.$$

Let $T : X \to Y$ be a map from an S_b-metric space X to an S_b-metric space Y. Then T is continuous at $x \in X$ if and only if $Tx_n \to Tx$ whenever $x_n \to x$.

Let $(X, \mathscr{S}_b)$ be an S_b-metric space with $b \geq 1$. $(X, \mathscr{S}_b)$ is said to be compact if every sequence in X has a convergent subsequence.

Theorem 13 *Let $(X, \mathscr{S}_b)$ be a compact S_b-metric space with $b > 1$ and T be a continuous self-mapping of X satisfying the condition* ($\mathbf{S}_b$**25a**)*. Then T has a unique fixed point.*

Proof There exists a compact subset A of X containing TX since T is a continuous self-mapping of X and X is compact. Then $TA \subset A$ and so $B = \bigcap_{n=1}^{\infty} T^n A$ is a nonempty compact subset of X which is mapped by T onto itself.

We now show that B is a singleton consisting of the unique fixed point x of T. Suppose that B is not a singleton. Then we get $diam\{B\} > 0$. Hence there exist $x, y \in B$ with $\mathscr{S}_b(x, x, y) = diam\{B\}$ since B is a compact subset. Also there exist $x_0, y_0 \in B$ with $Tx_0 = x$, $Ty_0 = y$ since T maps B onto itself. Using the condition ($\mathbf{S}_b$**25a**) we obtain

$$diam\{B\} = \mathscr{S}_b(x, x, y) = \mathscr{S}_b(Tx_0, Tx_0, Ty_0) < diam\{B\},$$

which is a contradiction. Consequently, T has a unique fixed point.

The case $b = 1$ was proved in Theorem 9.

Corollary 13 *Let $(X, \mathscr{S}_b)$ be a compact S_b-metric space with $b > 1$ and T be a continuous self-mapping of X satisfying the condition* ($\mathbf{S_b}$**25**)*. Then T has a unique fixed point.*

The case $b = 1$ was proved in Corollary 7.

4 Some Generalizations of Nemytskii-Edelstein and Ćirić's Fixed-Point Theorems

In this section we investigate new generalizations of the Nemytskii-Edelstein fixed-point theorem and the Ćirić's fixed-point theorem.

4.1 Generalizations of Nemytskii-Edelstein Fixed-Point Theorem

In this subsection we obtain new generalizations of the classical Nemytskii-Edelstein fixed-point theorem for continuous self-mappings of a compact S_b-metric space.

At first, we note that a b-metric is not continuous in general [6]. If we consider Lemmas 5 and 6 then we deduce that an S_b-metric is not always continuous. We give the following theorem.

Theorem 14 *Let $(X, \mathscr{S}_b)$ be a compact S_b-metric space with the continuous S_b-metric function $(b > 1)$. If a self-mapping $T : X \to X$ satisfies*

$$\mathscr{S}_b(Tx, Tx, Ty) < \mathscr{S}_b(x, x, y), \tag{11}$$

for all $x, y \in X$ with $x \neq y$, then T has a unique fixed point.

Proof Let us define the function $\xi : X \to [0, 1)$ as

$$\xi(x) = \mathscr{S}_b(x, x, Tx).$$

Since X is compact, the function ξ takes on its minimum value. That is, there exists $x_0 \in X$ such that

$$\mathscr{S}_b(x_0, x_0, Tx_0) < \mathscr{S}_b(x, x, Tx),$$

for all $x \in X$. We now show that x_0 is a fixed point of T. Assume that $Tx_0 \neq x_0$. Using the condition (11), we obtain

$$\mathscr{S}_b(Tx_0, Tx_0, TTx_0) < \mathscr{S}_b(x_0, x_0, Tx_0),$$

which contradicts the minimality of $\mathscr{S}_b(x_0, x_0, Tx_0)$ among all numbers $\mathscr{S}_b(x, x, Tx)$. Hence x_0 is a fixed point of T, that is, $Tx_0 = x_0$.

Now we prove that the fixed point x_0 is unique. On the contrary, assume that y_0 is another fixed point of T. From the condition (11) we get

$$\mathscr{S}_b(x_0, x_0, y_0) = \mathscr{S}_b(Tx_0, Tx_0, Ty_0) < \mathscr{S}_b(x_0, x_0, y_0).$$

Therefore $x_0 = y_0$, that is, x_0 is the unique fixed point of T.

The case $b = 1$ was proved in [43] (see Theorem 3.3 on page 264). We note that Theorem 14 is a generalization of Theorem 3.3 which is called the Nemystkii-Edelstein fixed-point theorem on an S-metric space.

Now we give two new generalizations of the classical Nemytskii-Edelstein fixed-point theorem for continuous self-mappings of a compact S_b-metric space. Notice that if T satisfies the condition (11) then T satisfies the condition ($\mathbf{S}_b\mathbf{25}$). Indeed we have

$$\begin{aligned}\mathcal{S}_b(Tx,Tx,Ty) &< \mathcal{S}_b(x,x,y)\\ &\leq \max\left\{\begin{array}{c}\mathcal{S}_b(x,x,y),\mathcal{S}_b(Tx,Tx,x),\mathcal{S}_b(Ty,Ty,y),\\ \mathcal{S}_b(Ty,Ty,x),\mathcal{S}_b(Tx,Tx,y)\end{array}\right\},\end{aligned}$$

for all $x, y \in X$ with $x \neq y$. Therefore we can give the following generalizations:

1. Corollary 13 is a generalization of Theorem 14.
2. By Proposition 8, Theorem 13 is another generalization of Theorem 14.

Now we give an example of a continuous self-mapping which satisfies the conditions ($\mathbf{S}_b\mathbf{25}$) and ($\mathbf{S}_b\mathbf{25a}$) but does not satisfy the condition (11).

Example 12 Let $X = [0, 1]$ be the compact S_b-metric space with the S_b-metric given in Example 1. Let us define the function $T : X \to X$ as

$$Tx = \left\{\begin{array}{ll} x+\frac{3}{4} & ; x \in \left[0,\frac{1}{4}\right) \\ 1 & ; x \in \left[\frac{1}{4},1\right] \end{array}\right.,$$

for all $x \in X$. Then T is a continuous self-mapping on the compact S_b-metric space $X = [0, 1]$. It can be easily seen that T satisfies the conditions ($\mathbf{S}_b\mathbf{25}$) and ($\mathbf{S}_b\mathbf{25a}$). Hence T has a unique fixed point $x = 1$ in $[0, 1]$. But the condition (11) is not satisfied. Indeed we get

$$\mathcal{S}_b(Tx,Tx,Ty) = \frac{1}{4}|x-y|^2 < \mathcal{S}_b(x,x,y) = \frac{1}{4}|x-y|^2,$$

for $x, y \in \left[0,\frac{1}{4}\right)$.

4.2 Generalizations of Ćirić's Fixed-Point Theorem

In this subsection we give a new generalization of the Ćirić's fixed-point theorem.

Theorem 15 *Let $(X, \mathcal{S}_b)$ be a complete S_b-metric space with $b \geq 1$ and T be a self-mapping of X satisfying*

$$\mathcal{S}_b(Tx,Tx,Ty) \leq h\max\left\{\begin{array}{c}\mathcal{S}_b(x,x,y),\mathcal{S}_b(Tx,Tx,x),\mathcal{S}_b(Ty,Ty,y),\\ \mathcal{S}_b(Ty,Ty,x),\mathcal{S}_b(Tx,Tx,y)\end{array}\right\}, \tag{12}$$

for all $x, y \in X$ and some $0 \leq h < \frac{1}{2b^2+b}$. Then T has a unique fixed point.

Proof Let $x_0 \in X$ and the sequence $\{x_n\}$ be defined as follows:

$$Tx_0 = x_1,\ Tx_1 = x_2, \dots,\ Tx_n = x_{n+1}, \dots .$$

Suppose that $x_n \neq x_{n+1}$ for all n. From the condition (12) and Lemma 3, we obtain

$$\begin{aligned}\mathcal{S}_b(x_n, x_n, x_{n+1}) &= \mathcal{S}_b(Tx_{n-1}, Tx_{n-1}, Tx_n) \qquad (13)\\ &\leq h \max \left\{ \begin{array}{c} \mathcal{S}_b(x_{n-1}, x_{n-1}, x_n), \mathcal{S}_b(x_n, x_n, x_{n-1}), \mathcal{S}_b(x_{n+1}, x_{n+1}, x_n), \\ \mathcal{S}_b(x_{n+1}, x_{n+1}, x_{n-1}), \mathcal{S}_b(x_n, x_n, x_n) \end{array} \right\} \\ &= h \max \left\{ \begin{array}{c} \mathcal{S}_b(x_{n-1}, x_{n-1}, x_n), \mathcal{S}_b(x_n, x_n, x_{n-1}), \\ \mathcal{S}_b(x_{n+1}, x_{n+1}, x_n), \mathcal{S}_b(x_{n+1}, x_{n+1}, x_{n-1}) \end{array} \right\} \\ &\leq h \max \left\{ \begin{array}{c} \mathcal{S}_b(x_{n-1}, x_{n-1}, x_n), b\mathcal{S}_b(x_{n-1}, x_{n-1}, x_n), \\ \mathcal{S}_b(x_{n+1}, x_{n+1}, x_n), \mathcal{S}_b(x_{n+1}, x_{n+1}, x_{n-1}) \end{array} \right\}. \qquad (14)\end{aligned}$$

By the condition (S_b2), we get

$$\mathcal{S}_b(x_{n+1}, x_{n+1}, x_{n-1}) \leq 2b\mathcal{S}_b(x_{n+1}, x_{n+1}, x_n) + b\mathcal{S}_b(x_{n-1}, x_{n-1}, x_n). \qquad (15)$$

Using the inequalities (13), (15) and Lemma 3, we have

$$\begin{aligned}\mathcal{S}_b(x_n, x_n, x_{n+1}) &\leq 2bh\mathcal{S}_b(x_{n+1}, x_{n+1}, x_n) + bh\mathcal{S}_b(x_{n-1}, x_{n-1}, x_n)\\ &\leq 2b^2h\mathcal{S}_b(x_n, x_n, x_{n+1}) + bh\mathcal{S}_b(x_{n-1}, x_{n-1}, x_n)\end{aligned}$$

and so

$$(1 - 2b^2h)\mathcal{S}_b(x_n, x_n, x_{n+1}) \leq bh\mathcal{S}_b(x_{n-1}, x_{n-1}, x_n),$$

which implies

$$\mathcal{S}_b(x_n, x_n, x_{n+1}) \leq \frac{bh}{1 - 2b^2h}\mathcal{S}_b(x_{n-1}, x_{n-1}, x_n). \qquad (16)$$

Let $k = \frac{bh}{1-2b^2h}$. Then $k < 1$ since $2b^2h + bh < 1$. We note that $1 - 2b^2h \neq 0$ since $0 \leq h < \frac{1}{2b^2+b}$. Now repeating this process in the inequality (16), using the mathematical induction we get

$$\mathcal{S}_b(x_n, x_n, x_{n+1}) \leq k^n\mathcal{S}_b(x_0, x_0, x_1). \qquad (17)$$

We prove that the sequence $\{x_n\}$ is Cauchy. For all $n, m \in \mathbb{N}$ with $m > n$, using the inequality (17), the condition (S_b2) and Lemma 3, we obtain

$$\mathcal{S}_b(x_n, x_n, x_m) \leq \frac{2bk^n}{1 - b^2k}\mathcal{S}_b(x_0, x_0, x_1). \qquad (18)$$

Therefore $\lim\limits_{n,m\to\infty} S_b(x_n, x_n, x_m) = 0$ by the inequality (18) and so $\{x_n\}$ is a Cauchy sequence. From the completeness hypothesis, there exists $x \in X$ such that $\{x_n\} \to x$. Now we show that x is a fixed point of T. Assume that $Tx \neq x$. Then using the condition (12) and Lemma 3, we get

$$\begin{aligned}\mathscr{S}_b(x_n, x_n, Tx) &= \mathscr{S}_b(Tx_{n-1}, Tx_{n-1}, Tx) \\ &\leq h \max \left\{ \begin{array}{c} \mathscr{S}_b(x_{n-1}, x_{n-1}, x), \mathscr{S}_b(x_n, x_n, x_{n-1}), \mathscr{S}_b(Tx, Tx, x), \\ \mathscr{S}_b(Tx, Tx, x_{n-1}), \mathscr{S}_b(x_n, x_n, x) \end{array} \right\}\end{aligned}$$

and so taking limit for $n \to \infty$ we have

$$\mathscr{S}_b(x, x, Tx) \leq h\mathscr{S}_b(Tx, Tx, x) \leq hb\mathscr{S}_b(x, x, Tx),$$

which implies $\mathscr{S}_b(x, x, Tx) = 0$ and $Tx = x$ since $0 \leq h < \frac{1}{2b^2+b}$.

Finally we show that the point x is unique. On the contrary, assume that x and y be two fixed points of T. Using the condition (12) and Lemma 3, we obtain

$$\begin{aligned}\mathscr{S}_b(Tx, Tx, Ty) = \mathscr{S}_b(x, x, y) &\leq h \max \left\{ \begin{array}{c} \mathscr{S}_b(x, x, y), \mathscr{S}_b(x, x, x), \mathscr{S}_b(y, y, y), \\ \mathscr{S}_b(y, y, x), \mathscr{S}_b(x, x, y) \end{array} \right\} \\ &\leq hb\mathscr{S}_b(x, x, y),\end{aligned}$$

which implies $x = y$ since $0 \leq h < \frac{1}{2b^2+b}$. Consequently, x is a unique fixed point of T.

The case $b = 1$ was proved in [41] (see Corollary 2.21 on page 123). We note that Theorem 13 and Corollary 13 are the generalizations of Theorem 15 for continuous self-mappings of a compact S_b-metric space. If we consider the self-mapping defined in Example 12, then it can be easily checked that the inequality (12) is not satisfied for $x = 0$ and $y = \frac{1}{8}$.

5 Some Fixed-Circle Theorems

Recently, the notion of a fixed circle has been introduced on a metric and an S-metric space (see [33, 34, 38]). It is important to obtain new-fixed circle theorems on various metric spaces since there exist some applications of them to other disciplines. For example, some activation functions having a fixed circle was used complex valued neural networks (see [28] for more details). Therefore, our aim is to investigate some generalized existence and uniqueness conditions for fixed-circle theorems on S_b-metric spaces.

5.1 Fixed-Circle Theorems on Metric Spaces

In this section we give a brief survey about the fixed-circle problem on metric spaces. Additionally, we obtain a new fixed-circle result.

Definition 16 ([33]) Let (X, d) be a metric space and $C_{x_0,r} = \{x \in X : d(x_0, x) = r\}$ be a circle. For a self-mapping $T : X \to X$, if $Tx = x$ for every $x \in C_{x_0,r}$ then the circle $C_{x_0,r}$ is called a fixed circle of T.

Using the inequality (1), we give the following existence theorem for a self-mapping having a fixed circle.

Theorem 16 ([33]) *Let (X, d) be a metric space and $C_{x_0,r}$ be any circle on X. Let us define the mapping*

$$\varphi : X \to [0, \infty), \varphi(x) = d(x, x_0), \tag{19}$$

for all $x \in X$. If there exists a self-mapping $T : X \to X$ satisfying

$(C1)$ $d(x, Tx) \leq \varphi(x) - \varphi(Tx)$
and
$(C2)$ $d(Tx, x_0) \geq r$,
for each $x \in C_{x_0,r}$, then the circle $C_{x_0,r}$ is a fixed circle of T.

Remark 2 ([33]) We note that Theorem 2 guarantees the existence of a fixed point while Theorem 16 guarantees the existence of a fixed circle. In the case where the circle $C_{x_0,r}$ has only one element Theorem 16 is a special case of Theorem 2.

Now we recall another known existence theorems.

Theorem 17 ([33]) *Let (X, d) be a metric space and $C_{x_0,r}$ be any circle on X. Let the mapping φ be defined as in (19). If there exists a self-mapping $T : X \to X$ satisfying*

$(C1)^*$ $d(x, Tx) \leq \varphi(x) + \varphi(Tx) - 2r$
and
$(C2)^*$ $d(Tx, x_0) \leq r$,
for each $x \in C_{x_0,r}$, then $C_{x_0,r}$ is a fixed circle of T.

Theorem 18 ([33]) *Let (X, d) be a metric space and $C_{x_0,r}$ be any circle on X. Let the mapping φ be defined as in (19). If there exists a self-mapping $T : X \to X$ satisfying*

$(C1)^{**}$ $d(x, Tx) \leq \varphi(x) - \varphi(Tx)$
and
$(C2)^{**}$ $hd(x, Tx) + d(Tx, x_0) \geq r$,
for each $x \in C_{x_0,r}$ and some $h \in [0, 1)$, then $C_{x_0,r}$ is a fixed circle of T.

Now we prove a new fixed-circle theorem.

Theorem 19 *Let (X, d) be a metric space and $C_{x_0,r}$ be any circle on X. Let the mapping φ be defined as in (19). If there exists a self-mapping $T : X \to X$ satisfying*

$(C1)^{***}$ $d(x, Tx) \leq \varphi(x) + \varphi(Tx) - 2r$
and
$(C2)^{***}$ $d(x, Tx) + d(Tx, x_0) \leq r$,
for each $x \in C_{x_0,r}$, then $C_{x_0,r}$ is a fixed circle of T.

Proof Let $x \in C_{x_0,r}$. Then using the conditions $(C1)^{***}$, $(C2)^{***}$ and the triangle inequality, we obtain

$$\begin{aligned} d(x, Tx) &\leq \varphi(x) + \varphi(Tx) - 2r \\ &= d(x, x_0) + d(Tx, x_0) - 2r \\ &\leq d(x, Tx) + d(Tx, x_0) + d(Tx, x_0) - 2r \\ &\leq 2d(x, Tx) + 2d(Tx, x_0) - 2r \\ &\leq 2r - 2r = 0 \end{aligned}$$

and so

$$d(x, Tx) = 0,$$

which implies $Tx = x$. Consequently, $C_{x_0,r}$ is a fixed circle of T.

Example 13 ([33]) Let (X, d) be a metric space, $C_{x_0,r}$ be any circle on X and α be a constant such that

$$d(\alpha, x_0) \neq r.$$

If we define the self-mapping $T : X \to X$ as

$$Tx = \begin{cases} x \ ; x \in C_{x_0,r} \\ \alpha \ ; \text{otherwise} \end{cases},$$

for all $x \in X$, then it can be easily seen that the conditions $(C1)$ and $(C2)$ (resp. the conditions $(C1)^*$, $(C2)^*$, the conditions $(C1)^{**}$, $(C2)^{**}$ and the conditions $(C1)^{***}$, $(C2)^{***}$) are satisfied. Clearly $C_{x_0,r}$ is a fixed circle of T.

Let $I_X : X \to X$ be the identity map defined as $I_X(x) = x$ for all $x \in X$. Notice that the identity map satisfies the conditions $(C1)$ and $(C2)$ (resp. $(C1)^*$ and $(C2)^*$, $(C1)^{**}$ and $(C2)^{**}$, $(C1)^{***}$ and $(C2)^{***}$) in Theorem 16 (resp. Theorem 17, Theorem 18 and Theorem 19). Now we investigate a condition which excludes the I_X in Theorems 16–19. We give the following theorem.

Theorem 20 ([33]) *Let (X, d) be a metric space and $C_{x_0,r}$ be any circle on X. Let the mapping φ be defined as in (19). If a self-mapping $T : X \to X$ satisfies the* *condition*

$$(I_d) \qquad d(x, Tx) \leq \frac{\varphi(x) - \varphi(Tx)}{h},$$

for all $x \in X$ and some $h > 1$, then $T = I_X$ and $C_{x_0,r}$ is a fixed circle of T.

Notice that the converse statement of this theorem is also true. Hence if a self-mapping T in Theorem 16 (resp. Theorem 17, Theorem 18 and Theorem 19) does not satisfy the condition (I_d) given in Theorem 20 then T can not be the identity map.

Notice that the fixed circle $C_{x_0,r}$ is not necessarily unique in Theorem 16 (resp. Theorem 17, Theorem 18 and Theorem 19).

Proposition 9 ([33]) *Let (X, d) be a metric space. For any given circles $C_{x_0,r}$ and $C_{x_1,\rho}$, there exists at least one self-mapping T of X such that T fixes the circles $C_{x_0,r}$ and $C_{x_1,\rho}$.*

Corollary 14 ([33]) *Let (X, d) be a metric space. For any given circles $C_{x_1,r_1}, \cdots, C_{x_n,r_n}$, there exists at least one self-mapping T of X such that T fixes the circles $C_{x_1,r_1}, \cdots, C_{x_n,r_n}$.*

Then it is a natural problem to investigate the uniqueness of the fixed circles obtained in the above fixed-circle theorems. Now we investigate the uniqueness conditions for the fixed circles in Theorem 16.

Theorem 21 ([33]) *Let (X, d) be a metric space and $C_{x_0,r}$ be any circle on X. Let $T : X \to X$ be a self-mapping satisfying the conditions $(C1)$ and $(C2)$ given in Theorem 16. If the contractive condition*

$$(C3) \qquad d(Tx, Ty) \leq hd(x, y), \tag{20}$$

is satisfied for all $x \in C_{x_0,r}$, $y \in X \setminus C_{x_0,r}$ and some $h \in [0, 1)$ by T, then $C_{x_0,r}$ is the unique fixed circle of T.

Notice that the uniqueness of the fixed circle in Theorems 17–19 can be also obtained using the contractive condition $(C3)$. More generally it is possible to use an appropriate contractive condition for the uniqueness of the obtained fixed circle theorems (see [33] for more details).

5.2 Fixed-Circle Theorems on S-Metric Spaces

In this section we recall the following fixed-circle definition on an S-metric space. Some comparisons of circles were given on metric and S-metric spaces. Then we give a survey of the known results about the fixed-circle problem on S-metric spaces. Also we give a new example of a self-mapping having a fixed circle.

Definition 17 ([34]) Let $(X, \mathscr{S})$ be an S-metric space and $x_0 \in X$, $r \in (0, \infty)$. The circle centered at x_0 with radius r is defined by

$$C^S_{x_0,r} = \{x \in X : \mathscr{S}(x, x, x_0) = r\}.$$

Definition 18 ([34]) Let $(X, \mathscr{S})$ be an S-metric space, $C^S_{x_0,r} = \{x \in X : \mathscr{S}(x, x, x_0) = r\}$ be a circle on X and $T : X \to X$ be a self-mapping. If $Tx = x$ for all $x \in C^S_{x_0,r}$ then the circle $C^S_{x_0,r}$ is called a fixed circle of T.

Now we recall the following propositions and corollaries.

Proposition 10 ([38]) *Let $(X, \mathscr{S})$ be an S-metric space such that $\mathscr{S}$ is generated by a metric d. Then any circle $C^S_{x_0,r}$ on the S-metric space is the circle $C_{x_0,\frac{r}{2}}$ on the metric space (X, d).*

Corollary 15 ([38]) *The circle $C_{x_0,r}$ on a metric space (X, d) is the circle $C^S_{x_0,2r}$ on the S-metric space generated by d.*

Proposition 11 ([38]) *Let (X, d_S) be a metric space such that d_S is generated by an S-metric $\mathscr{S}$. Then any circle $C_{x_0,r}$ on the metric space (X, d_S) is the circle $C^S_{x_0,\frac{r}{2}}$ on the S-metric space $(X, \mathscr{S})$.*

Corollary 16 ([38]) *The circle $C^S_{x_0,r}$ on an S-metric space $(X, \mathscr{S})$ is the circle $C_{x_0,2r}$ on the metric space (X, d_S) where d_S is generated by $\mathscr{S}$.*

In the following theorems we see some fixed-circle results on S-metric spaces.

Theorem 22 ([38]) *Let $(X, \mathscr{S})$ be an S-metric space and $C^S_{x_0,r}$ be any circle on X. Let us define the mapping*

$$\varphi : X \to [0, \infty),\ \varphi(x) = \mathscr{S}(x, x, x_0), \tag{21}$$

for all $x \in X$. If there exists a self-mapping $T : X \to X$ satisfying

$(SC1)$ $\mathscr{S}(x, x, Tx) \leq \varphi(x) - \varphi(Tx)$
and

$(SC2)$ $\mathscr{S}(Tx, Tx, x_0) \geq r$,
for each $x \in C^S_{x_0,r}$, then $C^S_{x_0,r}$ is a fixed circle of T.

Theorem 23 ([38]) *Let $(X, \mathscr{S})$ be an S-metric space and $C^S_{x_0,r}$ be any circle on X. Let the mapping φ be defined as in (21). If there exists a self-mapping $T : X \to X$ satisfying*

$(SC1)^*$ $\mathscr{S}(x, x, Tx) \leq \varphi(x) + \varphi(Tx) - 2r$
and

$(SC2)^*$ $\mathscr{S}(Tx, Tx, x_0) \leq r$,
for each $x \in C^S_{x_0,r}$, then $C^S_{x_0,r}$ is a fixed circle of T.

Theorem 24 ([34]) *Let $(X, \mathscr{S})$ be an S-metric space and $C^S_{x_0,r}$ be any circle on X. Let the mapping φ be defined as in (21). If there exists a self-mapping $T : X \to X$ satisfying*

$(SC1)^{**}$ $\mathscr{S}(x, x, Tx) \leq \varphi(x) - \varphi(Tx)$
and
$(SC2)^{**}$ $h\mathscr{S}(x, x, Tx) + \mathscr{S}(Tx, Tx, x_0) \geq r,$
for each $x \in C^S_{x_0,r}$ and some $h \in [0, 1)$, then $C^S_{x_0,r}$ is a fixed circle of T.

Theorem 25 ([34]) *Let $(X, \mathscr{S})$ be an S-metric space and $C^S_{x_0,r}$ be any circle on X. Let the mapping φ be defined as in (21). If there exists a self-mapping $T : X \to X$ satisfying*

$(SC1)^{***}$ $\mathscr{S}(x, x, Tx) \leq \varphi(x) + \varphi(Tx) - 2r$
and
$(SC2)^{***}$ $\mathscr{S}(x, x, Tx) + \mathscr{S}(Tx, Tx, x_0) \leq r,$
for each $x \in C^S_{x_0,r}$, then $C^S_{x_0,r}$ is a fixed circle of T.

Example 14 Let $X = \mathbb{C}$ and $(\mathbb{C}, \mathscr{S})$ be the S-metric space with the S-metric defined in (3). Let us consider the circle $C^S_{0,3}$ and define the self-mapping $T : \mathbb{C} \to \mathbb{C}$ by

$$Tz = \begin{cases} \frac{9}{4\overline{z}} & ; z \neq 0 \\ 0 & ; z = 0 \end{cases},$$

for all $z \in \mathbb{C}$. Then it can be easily seen that the conditions $(SC1)$ and $(SC2)$ (resp. the conditions $(SC1)^*$, $(SC2)^*$, the conditions $(SC1)^{**}$, $(SC2)^{**}$ and the conditions $(SC1)^{***}$, $(SC2)^{***}$) are satisfied. Clearly $C^S_{0,3}$ is the fixed circle of T. But, if we define the self-mapping $T : \mathbb{C} \to \mathbb{C}$ by

$$Tz = \begin{cases} \frac{9}{4z} & ; z \neq 0 \\ 0 & ; z = 0 \end{cases},$$

for all $z \in \mathbb{C}$, then the self-mapping T satisfies the condition $(SC2)^*$ but does not satisfy the condition $(SC1)^*$. Clearly T does not fix the circle $C^S_{0,3}$. Especially, T maps the circle $C^S_{0,3}$ onto itself while fixes the points $z_1 = \frac{3}{2}$ and $z_2 = -\frac{3}{2}$ only.

Notice that the identity map satisfies the conditions $(SC1)$ and $(SC2)$ (resp. $(SC1)^*$ and $(SC2)^*$, $(SC1)^{**}$ and $(SC2)^{**}$, $(SC1)^{***}$ and $(SC2)^{***}$) in Theorem 22 (resp. Theorem 23, Theorem 24 and Theorem 25). Now we determine a condition which excludes the I_X in Theorem 22, Theorem 23, Theorem 24 and Theorem 25. We give the following theorem.

Theorem 26 ([34]) *Let $(X, \mathscr{S})$ be an S-metric space and $C^S_{x_0,r}$ be any circle on X. Let the mapping φ be defined as in (21). If there exists a self-mapping $T : X \to X$ satisfying the condition*

$$(I_S) \qquad \mathscr{S}(x,x,Tx) \le \frac{\varphi(x)-\varphi(Tx)}{h},$$

for all $x \in X$ and some $h > 2$, then $T = I_X$ and $C^S_{x_0,r}$ is a fixed circle of T.

Let $(X, \mathscr{S})$ be an S-metric space. For any given circles $C^S_{x_0,r}$ and $C^S_{x_1,\rho}$ on X, we notice that there exists at least one self-mapping T of X such that T fixes both of the circles $C^S_{x_0,r}$ and $C^S_{x_1,\rho}$. Indeed, let us define the mappings $\varphi_1, \varphi_2 : X \to [0,\infty)$ as

$$\varphi_1(x) = \mathscr{S}(x,x,x_0)$$

and

$$\varphi_2(x) = \mathscr{S}(x,x,x_1),$$

for all $x \in X$. If we define the self-mapping $T : X \to X$ as

$$Tx = \begin{cases} x \ ; \ x \in C^S_{x_0,r} \cup C^S_{x_1,\rho} \\ \alpha \ ; \qquad \text{otherwise} \end{cases},$$

for all $x \in X$, where α is a constant satisfying $\mathscr{S}(\alpha,\alpha,x_0) \neq r$ and $\mathscr{S}(\alpha,\alpha,x_1) \neq \rho$, it can be easily seen that the self-mapping $T : X \to X$ satisfies the conditions $(SC1)$ and $(SC2)$ (resp. $(SC1)^*$ and $(SC2)^*$, $(SC1)^{**}$ and $(SC2)^{**}$, $(SC1)^{***}$ and $(SC2)^{***}$) in Theorem 22 (resp. Theorem 23, Theorem 24 and Theorem 25) for the circles $C^S_{x_0,r}$ and $C^S_{x_1,\rho}$ using the mappings φ_1 and φ_2, respectively. Hence T fixes both of the circles $C^S_{x_0,r}$ and $C^S_{x_1,\rho}$. The number of the fixed circles can be extended to any positive integer n using the same arguments.

In the following theorem, a uniqueness condition of a fixed circle was given.

Theorem 27 ([38]) *Let $(X, \mathscr{S})$ be an S-metric space and $C^S_{x_0,r}$ be any circle on X. Let $T : X \to X$ be a self-mapping satisfying the conditions $(SC1)$ and $(SC2)$ given in Theorem 22. If the contractive condition $(S25)$ is satisfied for all $x \in C^S_{x_0,r}$, $y \in X \backslash C^S_{x_0,r}$ by T then $C^S_{x_0,r}$ is the unique fixed circle of T.*

Notice that the uniqueness of the fixed circle in Theorem 23, Theorem 24 and Theorem 25 can be also obtained using the contractive condition $(S25)$. Furthermore it is possible to use appropriate contractive conditions for the uniqueness of the obtained fixed circle theorems (see [34, 38] for more details).

5.3 Fixed-Circle Theorems on S_b-Metric Spaces

In this section we give new fixed-circle theorems on an S_b-metric space with the geometric interpretation. Also we obtain some illustrative examples for our results.

Definition 19 Let $(X, \mathscr{S}_b)$ be an S_b-metric space with $b \geq 1$ and $x_0 \in X$, $r \in (0, \infty)$. The circle centered at x_0 with radius r is defined by

$$C^{S_b}_{x_0,r} = \{x \in X : \mathscr{S}_b(x, x, x_0) = r\} .$$

Example 15 Let $X = \mathbb{R}$ and the function $\mathscr{S} : X \times X \times X \to [0, \infty)$ be defined as

$$\mathscr{S}(x, y, z) = |arccotx - arccoty| + |arccotx + arccoty - 2arccotz| ,$$

for all $x, y, z \in \mathbb{R}$. Then $\mathscr{S}$ is an S-metric which is not generated by any metric and the pair $(\mathbb{R}, \mathscr{S})$ is an S-metric space. If we consider the function $\mathscr{S}_b : X \times X \times X \to [0, \infty)$ defined as

$$\mathscr{S}_b(x, y, z) = \mathscr{S}(x, y, z)^3,$$

then the function $\mathscr{S}_b$ is an S_b-metric with $b = 16$.

In the following example we extend the S_b-metric defined in the previous example to the three dimensional case and give an example of a circle on this S_b-metric space using mathematica [50].

Example 16 Let us consider the set $X = \mathbb{R}^3$ and the function $\mathscr{S}_b : X \times X \times X \to [0, \infty)$ be defined as

$$\mathscr{S}_b(x, y, z) = \sum_{i=1}^{3} \left(|arccotx_i - arccoty_i| + |arccotx_i + arccoty_i - 2arccotz_i|\right)^3 ,$$

for all $x = (x_1, x_2, x_3)$, $y = (y_1, y_2, y_3)$, $z = (z_1, z_2, z_3) \in X$. Then $\mathscr{S}_b$ is an S_b-metric with $b = 16$ and $(\mathbb{R}^3, \mathscr{S}_b)$ is an S_b-metric space. If we choose $x_0 = (0, 0, 0) = 0$ and $r = 8\pi$, then we get the circle

$$\begin{aligned} C^{S_b}_{0,8\pi} &= \{x \in X : \mathscr{S}_b(x, x, x_0) = 8\pi\} \\ &= \left\{ x \in X : \sum_{i=1}^{3} \left| arccotx_i - \frac{\pi}{2} \right|^3 = \pi \right\} , \end{aligned}$$

as shown in Fig. 1.

Using the *arctan* function, we obtain the following example of an S_b-metric space.

Example 17 Let us consider the set $X = \mathbb{R}^3$ and the function $\mathscr{S}_b : X \times X \times X \to [0, \infty)$ be defined as

$$\mathscr{S}_b(x, y, z) = \sum_{i=1}^{3} \left(|arctanx_i - arctany_i| + |arctanx_i + arctany_i - 2arctanz_i|\right)^3 ,$$

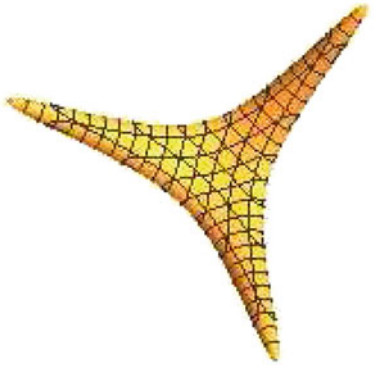


Fig. 1 The circle $C^{S_b}_{0,8\pi}$ in Example 16

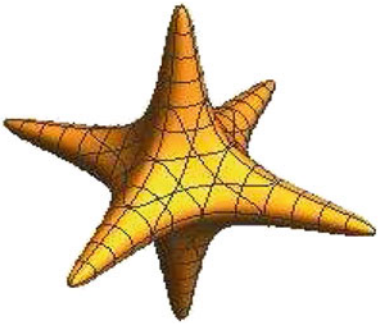


Fig. 2 The circle $C^{S_b}_{0,8\pi}$ in Example 17

for all $x = (x_1, x_2, x_3)$, $y = (y_1, y_2, y_3)$, $z = (z_1, z_2, z_3) \in X$. Then $\mathscr{S}_b$ is an S_b-metric with $b = 16$ and $(\mathbb{R}^3, \mathscr{S}_b)$ is an S_b-metric space. If we choose $x_0 = (0, 0, 0) = 0$ and $r = 8\pi$, then we get the circle

$$C^{S_b}_{0,8\pi} = \{x \in X : \mathscr{S}_b(x, x, x_0) = 8\pi\}$$
$$= \left\{x \in X : \sum_{i=1}^{3} |arctanx_i|^3 = \pi\right\},$$

as shown in Fig. 2.

Definition 20 Let $(X, \mathscr{S}_b)$ be an S_b-metric space with $b \geq 1$, $C^{S_b}_{x_0,r}$ be a circle on X and $T : X \to X$ be a self-mapping. If $Tx = x$ for all $x \in C^{S_b}_{x_0,r}$ then the circle $C^{S_b}_{x_0,r}$ is called as a fixed circle of T.

Now we give the following existence and uniqueness theorems for the fixed circles of self-mappings on an S_b-metric space.

Theorem 28 *Let $(X, \mathscr{S}_b)$ be an S_b-metric space with $b \geq 1$ and $C^{S_b}_{x_0,r}$ be a circle on X. Let us define the mapping*

$$\varphi : X \to [0, \infty), \varphi(x) = \mathscr{S}_b(x, x, x_0), \tag{22}$$

for all $x \in X$. If there exists a self-mapping $T : X \to X$ satisfying

(CS_b1) $\mathscr{S}_b(x, x, Tx) \leq \varphi(x) - \varphi(Tx)$
and

(CS_b2) $\mathscr{S}_b(Tx, Tx, x_0) \geq r$,
for each $x \in C^{S_b}_{x_0,r}$, then the circle $C^{S_b}_{x_0,r}$ is a fixed circle of T.

Proof Let us consider the mapping φ defined in (22) for a given circle $C^{S_b}_{x_0,r}$ and let $x \in C^{S_b}_{x_0,r}$ be any point. Now we show that $Tx = x$ whenever $x \in C^{S_b}_{x_0,r}$. Using the condition (CS_b1) we get

$$\begin{aligned} \mathscr{S}_b(x, x, Tx) &\leq \varphi(x) - \varphi(Tx) \\ &= \mathscr{S}_b(x, x, x_0) - \mathscr{S}_b(Tx, Tx, x_0) \\ &= r - \mathscr{S}_b(Tx, Tx, x_0). \end{aligned} \tag{23}$$

Because of the condition (CS_b2), the point Tx should be lie on or exterior of the circle $C^{S_b}_{x_0,r}$. Then we have two cases.

If $\mathscr{S}_b(Tx, Tx, x_0) > r$ then using the inequality (23) we have a contradiction. Therefore it should be $\mathscr{S}_b(Tx, Tx, x_0) = r$. In this case, using the inequality (20) we obtain

$$\mathscr{S}_b(x, x, Tx) \leq r - \mathscr{S}_b(Tx, Tx, x_0) = r - r = 0$$

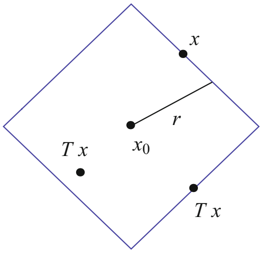
and so $Tx = x$. Therefore we have $Tx = x$ for all $x \in C^{S_b}_{x_0,r}$. Consequently, the self-mapping T fixes the circle $C^{S_b}_{x_0,r}$.

Remark 3

1) Notice that the condition (CS_b1) guarantees that Tx is not in the exterior of the circle $C^{S_b}_{x_0,r}$ for each $x \in C^{S_b}_{x_0,r}$. Similarly, the condition (CS_b2) guarantees that Tx is not in the interior of the circle $C^{S_b}_{x_0,r}$ for each $x \in C^{S_b}_{x_0,r}$. Consequently, $Tx \in C^{S_b}_{x_0,r}$ for each $x \in C^{S_b}_{x_0,r}$ and so we have $T(C^{S_b}_{x_0,r}) \subset C^{S_b}_{x_0,r}$ (see Figs. 3, 4, and 5).
2) If we take $b = 1$ in Theorem 28 then we get Theorem 22.

Now we give an example of a self-mapping which has a fixed circle.

Fig. 3 The geometric description of the condition (CS_b1)



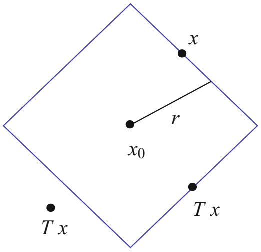


Fig. 4 The geometric description of the condition (CS_b2)

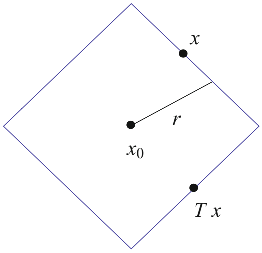


Fig. 5 The geometric description of the condition $(CS_b1) \cap (CS_b2)$

Example 18 Let $(X, \mathscr{S}_b)$ be an S_b-metric space with $b \geq 1$ and $C_{x_0,r}^{S_b}$ be any circle on X. Let us define the self-mapping $T : X \to X$ as

$$Tx = \begin{cases} x & ; \ x \in C_{x_0,r}^{S_b} \\ x_0 & ; \ \text{otherwise} \end{cases},$$

for all $x \in X$. It can be easily checked that the conditions (CS_b1) and (CS_b2) are satisfied. Then $C_{x_0,r}^{S_b}$ is a fixed circle of T.

In the following example, we give an example of a self-mapping which satisfies the condition (CS_b1) and does not satisfy the condition (CS_b2).

Example 19 Let $X = \mathbb{R}$ and the function $\mathscr{S}_b$ be the S_b-metric with $b = 4$ defined by

$$\mathscr{S}_b(x, y, z) = (|x - z| + |y - z|)^2,$$

for all $x, y, z \in \mathbb{R}$. Let $(\mathbb{R}, \mathscr{S}_b)$ be the corresponding S_b-metric space. Let us consider the circle $C_{0,4}^{S_b} = \{-1, 1\}$ and define the self-mapping $T : \mathbb{R} \to \mathbb{R}$ as

$$Tx = \begin{cases} 0 \ ; & x \in C_{0,4}^{S_b} \\ 4 \ ; & \text{otherwise} \end{cases},$$

for all $x \in \mathbb{R}$. Then the self-mapping T satisfies the condition (CS_b1) but does not satisfy the condition (CS_b2). Clearly, T does not fix the circle $C_{0,4}^{S_b}$.

Now we give an example of a self-mapping which satisfies the condition (CS_b2) and does not satisfy the condition (CS_b1).

Example 20 Let $(X, \mathscr{S}_b)$ be an S_b-metric space with $b \geq 1$ and $C^{S_b}_{x_0,r}$ be any circle on X. Let β be chosen such that $\mathscr{S}_b(\beta, \beta, x_0) = \rho > r$ and consider the self-mapping $T : X \to X$ defined by $Tx = \beta$ for all $x \in X$. Then the self-mapping T satisfies the condition (CS_b2) but does not satisfy the condition (CS_b1). Clearly, T does not fix the circle $C^{S_b}_{x_0,r}$.

Theorem 29 *Let $(X, \mathscr{S}_b)$ be an S_b-metric space with $b \geq 1$ and $C^{S_b}_{x_0,r}$ be a circle on X. Let the self-mapping φ be defined as in (22). If there exists a self-mapping $T : X \to X$ satisfying*

$(CS_b1)^*$ $\mathscr{S}_b(x, x, Tx) \leq \varphi(x) + \varphi(Tx) - 2r$
and
$(CS_b2)^*$ $\mathscr{S}_b(Tx, Tx, x_0) \leq r$,
for each $x \in C^{S_b}_{x_0,r}$, then the circle $C^{S_b}_{x_0,r}$ is a fixed circle of T.

Proof Let us consider the mapping φ defined in (22) for a given circle $C^{S_b}_{x_0,r}$ and let $x \in C^{S_b}_{x_0,r}$ be any point. Using the condition $(CS_b1)^*$ we obtain

$$\begin{aligned} \mathscr{S}_b(x, x, Tx) &\leq \varphi(x) + \varphi(Tx) - 2r \\ &= \mathscr{S}_b(x, x, x_0) + \mathscr{S}_b(Tx, Tx, x_0) - 2r \\ &= \mathscr{S}_b(Tx, Tx, x_0) - r. \end{aligned} \tag{24}$$

Because of the condition $(CS_b2)^*$, the point Tx should be lie on or interior of the circle $C^{S_b}_{x_0,r}$.

If $\mathscr{S}_b(Tx, Tx, x_0) < r$ then using the inequality (24) we have a contradiction. Therefore it should be $\mathscr{S}_b(Tx, Tx, x_0) = r$. If $\mathscr{S}_b(Tx, Tx, x_0) = r$ then using the inequality (24) we get

$$\mathscr{S}_b(x, x, Tx) \leq \mathscr{S}_b(Tx, Tx, x_0) - r = r - r = 0$$

and so we find $Tx = x$. Consequently, $C^{S_b}_{x_0,r}$ is a fixed circle of T.

Remark 4

1) Notice that the condition $(CS_b1)^*$ guarantees that Tx is not in the interior of the circle $C^{S_b}_{x_0,r}$ for each $x \in C^{S_b}_{x_0,r}$. Similarly, the condition $(CS_b2)^*$ guarantees that Tx is not in the exterior of the circle $C^{S_b}_{x_0,r}$ for each $x \in C^{S_b}_{x_0,r}$. Consequently, $Tx \in C^{S_b}_{x_0,r}$ for each $x \in C^{S_b}_{x_0,r}$ and so we have $T(C^{S_b}_{x_0,r}) \subset C^{S_b}_{x_0,r}$ (see Figs. 6, 7, and 8).
2) If we take $b = 1$ in Theorem 29 then we get Theorem 23.

Now we give an example of a self-mapping having a fixed-circle.

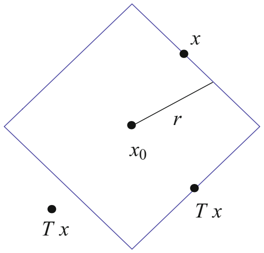


Fig. 6 The geometric description of the condition $(CS_b1)^*$

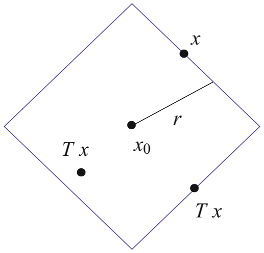


Fig. 7 The geometric description of the condition $(CS_b2)^*$

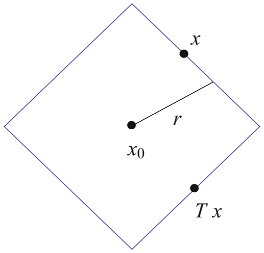


Fig. 8 The geometric description of the condition $(CS_b1)^* \cap (CS_b2)^*$

Example 21 Let $(\mathbb{R}, \mathscr{S}_b)$ be the S_b-metric space with $b = 4$ defined in Example 19. Let us consider the circle $C_{0,64}^{S_b} = \{-4, 4\}$ on $\mathbb{R}$ and define the self-mapping $T : \mathbb{R} \to \mathbb{R}$ as

$$Tx = \begin{cases} \frac{16}{x} \ ; \ x \in C_{0,64}^{S_b} \\ 0 \ ; \ \text{otherwise} \end{cases},$$

for all $x \in \mathbb{R}$. Then the self-mapping T satisfies the conditions $(CS_b1)^*$ and $(CS_b2)^*$. It can be easily checked that $C_{0,64}^{S_b}$ is a fixed circle of T.

We give an example of a self-mapping which satisfies the condition $(CS_b1)^*$ and does not satisfy the condition $(CS_b2)^*$.

Example 22 Let $(\mathbb{R}, \mathscr{S}_b)$ be the S_b-metric space with $b = 4$ defined in Example 19. Let us consider the circle $C_{0,1}^{S_b} = \left\{-\frac{1}{2}, \frac{1}{2}\right\}$ on $\mathbb{R}$ and define the self-mapping $T : \mathbb{R} \to \mathbb{R}$ as

$$Tx = \begin{cases} -\frac{3}{2} \ ; \ x = -\frac{1}{2} \\ \frac{3}{2} \ ; \ x = \frac{1}{2} \\ 3 \ ; \ \text{otherwise} \end{cases},$$

for all $x \in \mathbb{R}$. Then the self-mapping T satisfies the condition $(CS_b1)^*$ but does not satisfy the condition $(CS_b2)^*$. Clearly, T does not fix the circle $C_{0,1}^{S_b}$.

We give an example of a self-mapping which satisfies the condition $(CS_b2)^*$ and does not satisfy the condition $(CS_b1)^*$.

Example 23 Let $(X, \mathscr{S}_b)$ be any S_b-metric space with $b \geq 1$ and $C_{x_0,r}^{S_b}$ be any circle on X. Let β be chosen such that $\mathscr{S}_b(\beta, \beta, x_0) = \rho < r$ and consider the self-mapping $T : X \to X$ defined by $Tx = \beta$ for all $x \in X$. Then the self-mapping T satisfies the condition $(CS_b2)^*$ but does not satisfy the condition $(CS_b1)^*$. Clearly, T does not fix the circle $C_{x_0,r}^{S_b}$.

Theorem 30 *Let $(X, \mathscr{S}_b)$ be an S_b-metric space with $b \geq 1$ and $C_{x_0,r}^{S_b}$ be a circle on X. Let the self-mapping φ be defined as in (22). If there exists a self-mapping $T : X \to X$ satisfying*

$(CS_b1)^{**}$ $\mathscr{S}_b(x, x, Tx) \leq \varphi(x) - \varphi(Tx)$
and
$(CS_b2)^{**}$ $h\mathscr{S}_b(x, x, Tx) + \mathscr{S}_b(Tx, Tx, x_0) \geq r$,
for each $x \in C_{x_0,r}^{S_b}$ and some $h \in [0, 1)$, then the circle $C_{x_0,r}^{S_b}$ is a fixed circle of T.

Proof Let us consider the mapping φ defined in (22) for a given circle $C_{x_0,r}^{S_b}$ and let $x \in C_{x_0,r}^{S_b}$ be any point. Using the conditions $(CS_b1)^{**}$ and $(CS_b2)^{**}$ we obtain

$$\begin{aligned}\mathscr{S}_b(x, x, Tx) &\leq \varphi(x) - \varphi(Tx) = \mathscr{S}_b(x, x, x_0) - \mathscr{S}_b(Tx, Tx, x_0)\\ &= r - \mathscr{S}_b(Tx, Tx, x_0)\\ &\leq h\mathscr{S}_b(x, x, Tx) + \mathscr{S}_b(Tx, Tx, x_0) - \mathscr{S}_b(Tx, Tx, x_0)\\ &= h\mathscr{S}_b(x, x, Tx),\end{aligned}$$

which is a contradiction since $h \in [0, 1)$. Therefore we get $Tx = x$ and $C_{x_0,r}^{S_b}$ is a fixed circle of T.

Remark 5

1) Notice that the condition $(CS_b1)^{**}$ guarantees that Tx is not in the exterior of the circle $C_{x_0,r}^{S_b}$ for each $x \in C_{x_0,r}^{S_b}$. Similarly, the condition $(CS_b2)^{**}$ guarantees that Tx should be lie on or exterior or interior of the circle $C_{x_0,r}^{S_b}$. Consequently, Tx should be lie on or interior of the circle $C_{x_0,r}^{S_b}$ (see Fig. 9).
2) If we take $b = 1$ in Theorem 30 then we get Theorem 24.

Example 24 Let $(\mathbb{R}, \mathscr{S}_b)$ be the S_b-metric space with $b = 4$ defined in Example 19. Let us consider the circle $C_{0,36}^{S_b} = \{-3, 3\}$ on $\mathbb{R}$ and define the self-mapping $T : \mathbb{R} \to \mathbb{R}$ as

$$Tx = \begin{cases} e^{x+3} - 4 \ ; & x = -3 \\ 6 - x \ ; & x = 3 \\ 0 \ ; & \text{otherwise} \end{cases},$$

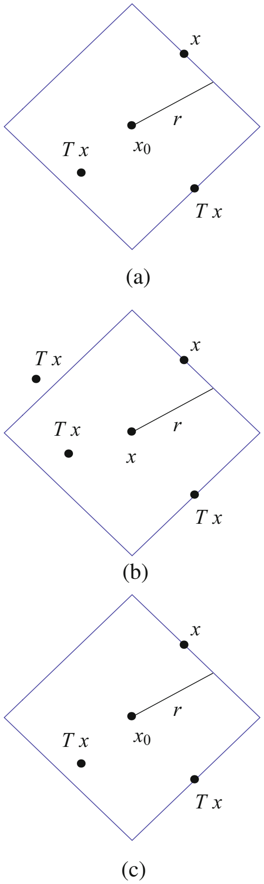


Fig. 9 The geometric interpretation of the conditions $(CS_b1)^{**}$ and $(CS_b2)^{**}$. (**a**) The condition $(CS_b1)^{**}$. (**b**) The condition $(CS_b2)^{**}$. (**c**) The condition $(CS_b1)^{**} \cap (CS_b2)^{**}$

for all $x \in \mathbb{R}$. Then the self-mapping T satisfies the conditions $(CS_b1)^{**}$ and $(CS_b2)^{**}$. Hence $C_{0,36}^{S_b}$ is a fixed circle of T.

Now we give an example of a self-mapping which satisfies the condition $(CS_b1)^{**}$ but does not satisfy the condition $(CS_b2)^{**}$.

Example 25 Let $(\mathbb{R}, \mathscr{S}_b)$ be the S_b-metric space with $b = 4$ defined in Example 19. Let us consider the circle $C_{2,16}^{S_b} = \{0, 4\}$ on $\mathbb{R}$ and define the self-mapping $T : \mathbb{R} \to \mathbb{R}$ as

$$Tx = \begin{cases} 2 \ ; \ x \in C_{2,16}^{S_b} \\ 18 \ ; \ \text{otherwise} \end{cases},$$

for all $x \in \mathbb{R}$. Then the self-mapping T satisfies the condition $(CS_b1)^{**}$ but does not satisfy the condition $(CS_b2)^{**}$. Clearly, T does not fix the circle $C_{2,16}^{S_b}$.

We give an example of a self-mapping which satisfies the condition $(CS_b2)^{**}$ and does not satisfy the condition $(CS_b1)^{**}$.

Example 26 Let $(\mathbb{R}, \mathcal{S}_b)$ be the S_b-metric space with $b = 4$ defined in Example 19. Let us consider the circle $C_{0,9}^{S_b} = \left\{-\frac{3}{2}, \frac{3}{2}\right\}$ on X and define the self-mapping $T : \mathbb{R} \to \mathbb{R}$ as $Tx = 9$ for all $x \in \mathbb{R}$. Then the self-mapping T satisfies the condition $(CS_b2)^{**}$ but does not satisfy the condition $(CS_b1)^{**}$. Clearly, T does not fix the circle $C_{0,9}^{S_b}$.

Theorem 31 *Let $(X, \mathcal{S}_b)$ be an S_b-metric space with $b \geq 1$ and $C_{x_0,r}^{S_b}$ be a circle on X. Let the self-mapping φ be defined as in (22). If there exists a self-mapping $T : X \to X$ satisfying*

$(CS_b1)^{***}$ $\quad \mathcal{S}_b(x, x, Tx) \leq \varphi(x) + \varphi(Tx) - 2r$

and

$(CS_b2)^{***}$ $\quad b\mathcal{S}_b(x, x, Tx) + \left(\frac{1+b^2}{2}\right)\mathcal{S}_b(Tx, Tx, x_0) \leq r,$

for each $x \in C_{x_0,r}^{S_b}$, then the circle $C_{x_0,r}^{S_b}$ is a fixed circle of T.

Proof Let us consider the mapping φ defined in (22) for a given circle $C_{x_0,r}^{S_b}$ and let $x \in C_{x_0,r}^{S_b}$ be any point. Using the conditions $(CS_b1)^{***}$, $(CS_b2)^{***}$ and (S_b2) we get

$$\begin{aligned}
\mathcal{S}_b(x, x, Tx) &\leq \varphi(x) + \varphi(Tx) - 2r = \mathcal{S}_b(x, x, x_0) + \mathcal{S}_b(Tx, Tx, x_0) - 2r \\
&\leq b\left[2\mathcal{S}_b(x, x, Tx) + \mathcal{S}_b(x_0, x_0, Tx)\right] + \mathcal{S}_b(Tx, Tx, x_0) - 2r \\
&= 2b\mathcal{S}_b(x, x, Tx) + b\mathcal{S}_b(x_0, x_0, Tx) + \mathcal{S}_b(Tx, Tx, x_0) - 2r \\
&\leq 2b\mathcal{S}_b(x, x, Tx) + b^2\mathcal{S}_b(Tx, Tx, x_0) + \mathcal{S}_b(Tx, Tx, x_0) - 2r \\
&= 2b\mathcal{S}_b(x, x, Tx) + (1 + b^2)\mathcal{S}_b(Tx, Tx, x_0) - 2r \\
&\leq 2r - 2r = 0
\end{aligned}$$

and so $\mathcal{S}_b(x, x, Tx) = 0$ which implies $Tx = x$. Consequently, $C_{x_0,r}^{S_b}$ is a fixed circle of T.

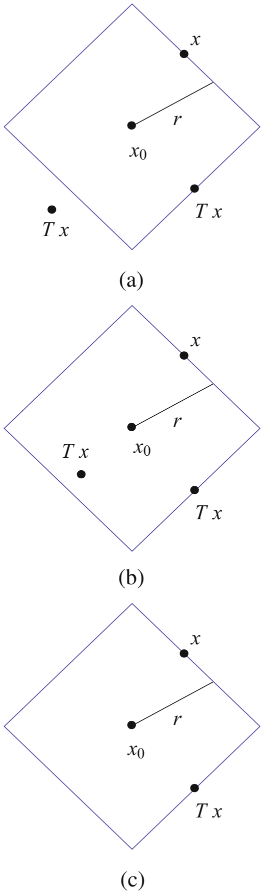
Remark 6

1) Notice that the condition $(CS_b1)^{***}$ guarantees that Tx is not in the interior of the circle $C_{x_0,r}^{S_b}$ for each $x \in C_{x_0,r}^{S_b}$. Similarly, the condition $(CS_b2)^{***}$ guarantees that Tx is not in the exterior of the circle $C_{x_0,r}^{S_b}$ for each $x \in C_{x_0,r}^{S_b}$. Consequently, $Tx \in C_{x_0,r}^{S_b}$ for each $x \in C_{x_0,r}^{S_b}$ and so we get $T(C_{x_0,r}^{S_b}) \subset C_{x_0,r}^{S_b}$ (see Fig. 10).
2) If we take $b = 1$ in Theorem 31 then we get Theorem 25.

Example 27 Let $X = \mathbb{R}$ and the function $\mathcal{S}_b : X \times X \times X \to [0, \infty)$ be defined by

$$\mathcal{S}_b(x, y, z) = \frac{1}{3}\left(|x - z| + |x + z - 2y|\right),$$

Fig. 10 The geometric interpretation of the conditions $(CS_b1)^{**}$ and $(CS_b2)^{**}$. (**a**) The condition $(CS_b1)^{***}$. (**b**) The condition $(CS_b2)^{***}$. (**c**) The condition $(CS_b1)^{***} \cap (CS_b2)^{***}$


(a)

(b)

(c)

for all $x, y, z \in \mathbb{R}$. Then $(\mathbb{R}, \mathscr{S}_b)$ is an S_b-metric space with $b = 1$. Let us consider the circle $C_{3,3}^{S_b} = \left\{-\frac{3}{2}, \frac{15}{2}\right\}$ and define the self-mapping $T : \mathbb{R} \to \mathbb{R}$ as

$$Tx = \begin{cases} x \ ; \ x \in \left\{-\frac{3}{2}, \frac{15}{2}\right\} \\ 3 \ ; \quad \text{otherwise} \end{cases},$$

for all $x \in \mathbb{R}$. Then the self-mapping T satisfies the conditions $(CS_b1)^{***}$ and $(CS_b2)^{***}$. Clearly, $C_{3,3}^{S_b}$ is a fixed circle of T.

If we consider the self-mapping T defined in Example 22, then it can be easily checked that the self-mapping T satisfies the condition $(CS_b1)^{***}$ but does not satisfy the condition $(CS_b2)^{***}$ for the unit circle $C_{0,1}^{S_b}$.

Now we give an example of a self-mapping which satisfies the condition $(CS_b2)^{***}$ and does not satisfy the condition $(CS_b1)^{***}$.

Example 28 Let $(X, \mathscr{S}_b)$ be any S_b-metric space with $b \geq 1$ and $C^{S_b}_{x_0,r}$ be any circle on X. If we consider the self-mapping $T : X \to X$ defined by $Tx = x_0$ for all $x \in X$. Then the self-mapping T satisfies the condition $(CS_b2)^{***}$ but does not satisfy the condition $(CS_b1)^{***}$. Clearly, T does not fix the circle $C^{S_b}_{x_0,r}$.

Notice that the identity map I_X satisfies the conditions (CS_b1) and (CS_b2) (resp. $(CS_b1)^*$ and $(CS_b2)^*$, $(CS_b1)^{**}$ and $(CS_b2)^{**}$, $(CS_b1)^{***}$ and $(CS_b2)^{***}$) in Theorem 28 (resp. Theorem 29, Theorem 30 and Theorem 31). In the following theorem we present a condition which excludes I_X in Theorems 28–31.

Theorem 32 *Let $(X, \mathscr{S}_b)$ be an S_b-metric space with $b \geq 1$ and $C^{S_b}_{x_0,r}$ be a circle on X. Let the self-mapping φ be defined as in (22). A self-mapping $T : X \to X$ satisfies the condition*

$$(I_{S_b}) \qquad \mathscr{S}_b(x, x, Tx) \leq \frac{\varphi(x) - \varphi(Tx)}{h},$$

for all $x \in X$ and some $h > 2b$ if and only if $T = I_X$.

Proof Let $x \in X$ and $Tx \neq x$. Then using the inequality (I_{S_b}) and the condition (S_b2) we get

$$\begin{aligned} h\mathscr{S}_b(x, x, Tx) &\leq \varphi(x) - \varphi(Tx) \\ &= \mathscr{S}_b(x, x, x_0) - \mathscr{S}_b(Tx, Tx, x_0) \\ &\leq b\,[2\mathscr{S}_b(x, x, Tx) + \mathscr{S}_b(x_0, x_0, Tx)] - \mathscr{S}_b(Tx, Tx, x_0) \\ &\leq 2b\mathscr{S}_b(x, x, Tx) + b\mathscr{S}_b(x_0, x_0, Tx) - b\mathscr{S}_b(x_0, x_0, Tx) \\ &= 2b\mathscr{S}_b(x, x, Tx) \end{aligned}$$

and so

$$(h - 2b)\,\mathscr{S}_b(x, x, Tx) \leq 0,$$

which is a contradiction since $h > 2b$. Hence we obtain $Tx = x$ and $T = I_X$. Conversely, it is clear that the identity map I_X satisfies the condition (I_{S_b}).

Hence if a self-mapping T in Theorem 28 (resp. Theorem 29, Theorem 30 and Theorem 31) does not satisfy the condition (I_{S_b}) given in Theorem 32 then T can not be the identity map.

Finally we investigate the uniqueness of the fixed circles in theorems obtained above on an S_b-metric space. We note that the fixed circle $C^{S_b}_{x_0,r}$ is not necessarily unique in Theorem 28 (resp. Theorem 29, Theorem 30 and Theorem 31). We can give the following proposition.

Proposition 12 *Let $(X,\mathscr{S}_b)$ be an S_b-metric space with $b \geq 1$. For any given circles $C^{S_b}_{x_0,r}$ and $C^{S_b}_{x_1,\rho}$, there exists at least one self-mapping T of X such that T fixes the circles $C^{S_b}_{x_0,r}$ and $C^{S_b}_{x_1,\rho}$.*

Proof Let $C^{S_b}_{x_0,r}$ and $C^{S_b}_{x_1,\rho}$ be any circles on X. Let us define the self-mapping $T : X \to X$ as

$$Tx = \begin{cases} x \ ; \ x \in C^{S_b}_{x_0,r} \cup C^{S_b}_{x_1,\rho} \\ \beta \ ; \qquad \text{otherwise} \end{cases}, \tag{25}$$

for all $x \in X$, where β is a constant satisfying $\mathscr{S}_b(\beta,\beta,x_0) \neq r$ and $\mathscr{S}_b(\beta,\beta,x_1) \neq \rho$. Let us define the mappings $\varphi_1, \varphi_2 : X \to [0,\infty)$ as

$$\varphi_1(x) = \mathscr{S}_b(x,x,x_0)$$

and

$$\varphi_2(x) = \mathscr{S}_b(x,x,x_1),$$

for all $x \in X$. Then it can be easily checked that the conditions (CS_b1) and (CS_b2) are satisfied by T for the circles $C^{S_b}_{x_0,r}$ and $C^{S_b}_{x_1,\rho}$ with the mappings $\varphi_1(x)$ and $\varphi_2(x)$, respectively. Clearly, $C^{S_b}_{x_0,r}$ and $C^{S_b}_{x_1,\rho}$ are the fixed circles of T by Theorem 28.

Finally we note that the self-mapping T defined in (25) satisfies the conditions $(CS_b1)^*$ and $(CS_b2)^*$ (resp. $(CS_b1)^{**}$ and $(CS_b2)^{**}$, $(CS_b1)^{***}$ and $(CS_b2)^{***}$) for the circles $C^{S_b}_{x_0,r}$ and $C^{S_b}_{x_1,\rho}$ with the mappings $\varphi_1(x)$ and $\varphi_2(x)$, respectively.

Corollary 17 *Let $(X,\mathscr{S}_b)$ be an S_b-metric space with $b \geq 1$. For any given circles $C^{S_b}_{x_1,r_1}, \ldots, C^{S_b}_{x_n,r_n}$, there exists at least one self-mapping T of X such that T fixes the circles $C^{S_b}_{x_1,r_1}, \ldots, C^{S_b}_{x_n,r_n}$.*

Hence it is important to investigate the uniqueness of the fixed circles. Now we determine a uniqueness condition for the circles in Theorem 28 (resp. Theorem 29, Theorem 30 and Theorem 31).

Theorem 33 *Let $(X,\mathscr{S}_b)$ be an S_b-metric space with $b \geq 1$, $C^{S_b}_{x_0,r}$ be a circle on X and $T : X \to X$ be a self-mapping which fixes the circle $C^{S_b}_{x_0,r}$. If the contractive condition* ($\mathbf{S}_b\mathbf{25}$) *is satisfied for all $x \in C^{S_b}_{x_0,r}$, $y \in X \backslash C^{S_b}_{x_0,r}$ by T, then $C^{S_b}_{x_0,r}$ is the unique fixed circle of T.*

Proof Suppose that there exist two fixed circles $C^{S_b}_{x_0,r}$ and $C^{S_b}_{x_1,\rho}$ of the self-mapping T. Let $x \in C^{S_b}_{x_0,r}$, $y \in C^{S_b}_{x_1,\rho}$ and $x \neq y$ be any arbitrary points. We show that $\mathscr{S}_b(x,x,y) = 0$ and so $x = y$. Using the condition ($\mathbf{S}_b\mathbf{25}$) we obtain

$$\mathscr{S}_b(x,x,y) = \mathscr{S}_b(Tx,Tx,Ty) < max\{\mathscr{S}_b(x,x,y), \mathscr{S}_b(Tx,Tx,x),$$

$$\mathscr{S}_b(Ty, Ty, y), \mathscr{S}_b(Ty, Ty, x), \mathscr{S}_b(Tx, Tx, y)\}$$
$$= \mathscr{S}_b(x, x, y),$$

which is a contradiction. Consequently, it should be $x = y$ for all $x \in C_{x_0,r}^{S_b}$, $y \in C_{x_1,\rho}^{S_b}$ and so $C_{x_0,r}^{S_b}$ is the unique fixed circle of T.

More generally it is possible to use appropriate contractive conditions for the uniqueness of the obtained fixed-circle theorems.

References

1. A. Aghajani, M. Abbas, J.R. Roshan, Common fixed point of generalized weak contractive mappings in partially ordered G_b-metric spaces. Filomat **28**(6), 1087–1101 (2014)
2. A. Azam, B. Fisher, M. Khan, Common fixed point theorems in complex valued metric spaces. Numer. Funct. Anal. Optim. **32**(3), 243–253 (2011)
3. D.F. Bailey, Some theorems on contractive mappings. J. Lond. Math. Soc. **41**, 101–106 (1996)
4. I.A. Bakhtin, The contraction mapping principle in almost metric space. Funct. Anal., Ulianowsk. Gos. Ped. Ins. **30**, 26–37 (1989)
5. S. Banach, Sur les operations dans les ensembles abstraits et leur application aux equations integrals. Fundam. Math. **2**, 133–181 (1922)
6. M. Boriceanu, M. Bota, A. Petrusel, Multivalued fractals in *b*-metric spaces. Cent. Eur. J. Math **8**(2), 367–377 (2010)
7. J. Caristi, Fixed point theorems for mappings satisfying inwardness conditions. Trans. Am. Math. Soc. **215**, 241–251 (1976)
8. B. Chaudhary, S. Nanda, *Functional Analysis with Applications* (Wiley Eastern Limited, New Delhi, 1989)
9. K. Ciesielski, On Stefan Banach and some of his results. Banach J. Math. Anal. **1**(1), 1–10 (2007)
10. L.B. Ćirić, A generalization of Banach's contraction principle. Proc. Am. Math. Soc. **45**, 267–27 (1974)
11. E.T. Copson, *Metric Space* (Universal Bookstall, New Delhi, 1996)
12. B.C. Dhage, Generalized metric spaces mappings with fixed point. Bull. Calcutta Math. Soc. **84**, 329–336 (1992)
13. M. Edelstein, On fixed and periodic points under contractive mappings. J. Lond. Math. Soc. **37**, 74–79 (1962)
14. Ö. Ege, Complex valued G_b-metric spaces. J. Comput. Anal. Appl. **21**(2), 363–368 (2016)
15. S. G*ä*hler, 2-metrische R*ä*ume und iher topoloische Struktur. Math. Nachr. **26**, 115–148 (1963)
16. A. Gupta, Cyclic contraction on *S*-metric space. Int. J. Anal. Appl. **3**(2), 119–130 (2013)
17. N.T. Hieu, N.T. Ly, N.V. Dung, A generalization of ciric quasi-contractions for maps on *S*-metric spaces. Thai J. Math. **13**(2), 369–380 (2015)
18. P.K. Jain, A. Khalil, *Metric Space* (Narosa Publishing House, New Delhi, 1996)
19. G.A. Jones, D. Singerman, *Complex Functions an Algebraic and Geometric Viewpoint* (Cambridge University Press, New York, 1987)
20. A.N. Kolmogorov, S.V. Fomin, *Elements of the Theory of Functions and Functional Analysis* (Dover Publication, New York, 1957)
21. A.N. Kolmogorov, S.V. Fomin, *Introductory Real Analysis* (Dover Publication, New York, 1970)
22. E. Kreyszig, *Introductory Functional Analysis with Applications* (Wiley, New York, 1978)

23. D.P. Mandic, The use of Möbius transformations in neural networks and signal processing, in *Proceedings of Neural Networks for Signal Processing X 1 and 2*, pp. 185–194 (2000)
24. J.E. Marsden, *Elementary Classica1 Analysis* (W. H. Freeman and Company, SanFrancisco, 1974)
25. N.M. Mlaiki, Common fixed points in complex S-metric space. Adv. Fixed Point Theory **4**(4), 509–524 (2014)
26. Z. Mustafa, B. Sims, A new approach to generalized metric spaces. J. Nonlinear Convex Anal. **7**, 289–297 (2006)
27. V.V. Nemytskii, The fixed point method in analysis. Usp. Mat. Nauk **1**, 141–174 (1936) [in Russian]
28. N. Özdemir, B.B. İskender, N.Y. Özgür, Complex valued neural network with Möbius activation function. Commun. Nonlinear Sci. Numer. Simul. **16**(12), 4698–4703 (2011)
29. N.Y. Özgür, N. Taş, *Some Generalizations of Fixed Point Theorems on S-Metric Spaces*. Essays in Mathematics and Its Applications in Honor of Vladimir Arnold (Springer, New York, 2016)
30. N.Y. Özgür, N. Taş, Some generalizations of the Banach's contraction principle on a complex valued S-metric space. J. New Theory **2**(14), 26–36 (2016)
31. N.Y. Özgür, N. Taş, Some new contractive mappings on S-metric spaces and their relationships with the mapping ($S25$). Math. Sci. **11**, 7 (2017). https://doi.org/10.1007/s40096-016-0199-4
32. N.Y. Özgür, N. Taş, Some fixed point theorems on S-metric spaces. Mat. Vesnik **69**(1), 39–52 (2017)
33. N.Y. Özgür, N. Taş, Some fixed-circle theorems on metric spaces. Bull. Malays. Math. Sci. Soc. (2017). https://doi.org/10.1007/s40840-017-0555-z
34. N.Y. Özgür, N. Taş, Fixed-circle problem on S-metric spaces with a geometric viewpoint. arXiv:1704.08838 [math.MG]
35. N.Y. Özgür, N. Taş, Common fixed point results on complex valued S-metric spaces (submitted for publication)
36. N.Y. Özgür, N. Taş, Some generalized fixed-point theorems on complex valued S-metric spaces (submitted for publication)
37. N.Y. Özgür, N. Taş, The Picard theorem on S-metric spaces, Acta Math. Sci. (in press)
38. N.Y. Özgür, N. Taş, Çelik, U.: Some fixed-circle results on S-metric spaces. Bull. Math. Anal. Appl. **9**(2), 10–23 (2017)
39. N.Y. Özgür, N. Taş, A note on "Best proximity point results in S-metric spaces" with some topological aspects (submitted for publication)
40. B.E. Rhoades, A comparison of various definitions of contractive mappings. Trans. Am. Math. Soc. **226**, 257–290 (1977)
41. S. Sedghi, N.V. Dung, Fixed point theorems on S-metric spaces. Mat. Vesnik **66**(1), 113–124 (2014)
42. S. Sedghi, N. Shobe, H. Zhou, A common fixed point theorem in D*-metric spaces. Fixed Point Theory Appl. **2007**, Article ID 27906, 13 pp. (2007)
43. S. Sedghi, N. Shobe, A. Aliouche, A generalization of fixed point theorems in S-metric spaces. Mat. Vesnik **64**(3), 258–266 (2012)
44. S. Sedghi, A. Gholidahneh, T. Došenović, J. Esfahani, S. Radenović, Common fixed point of four maps in S_b-metric spaces. J. Linear Topol. Algebra **5**(2), 93–104 (2016)
45. H. Siddique, *Functional Analysis with Applications* (Tata McGraw-Hill Publishing Company, New Delhi, 1986)
46. N. Souayah, N. Mlaiki, A fixed point theorem in S_b-metric spaces. J. Math. Comput. Sci. **16**, 131–139 (2016)
47. N. Taş, N.Y. Özgür, On parametric S-metric spaces and fixed-point type theorems for expansive mappings. J. Math. **2016**, Article ID 4746732, 6 pp. (2016) https://doi.org/10.1155/2016/4746732
48. N. Taş, N.Y. Özgür, New generalized fixed point results on S_b-metric spaces. arXiv:1703.01868 [math.GN]
49. M. Ughade, D. Turkoglu, S.R. Singh, R.D. Daheriya, Some fixed point theorems in A_b-metric space. Br. J. Math. Comput. Sci. **19**(6), 1–24 (2016)
50. Wolfram Research, Inc., Mathematica, Trial Version, Champaign, IL (2016)

Finite-Difference Modeling of Nonlinear Phenomena in Time-Domain Electromagnetics: A Review

Theodoros T. Zygiridis and Nikolaos V. Kantartzis

1 Introduction

The investigation of electromagnetic (EM) problems is directly connected to the solution of Maxwell's equations, which is, in most cases, obtained using computational methods, as analytical techniques can be practically applied only when specific or ideal conditions are satisfied. Especially when dynamic phenomena need to be considered, the corresponding studies are commonly performed in the time domain, and the finite-difference time-domain (FDTD) method constitutes one of the most popular choices [1, 2]. Of course, other alternatives also exist, such as the finite-element time-domain method [3], the discontinuous Galerkin time-domain approach [4], or techniques based on integral equation methods [5]. The FDTD scheme is well-known for its attractive features, which include the simplicity of implementation, the explicit character (i.e. no matrix inversions are required), the ability to model a wide range of material properties, its parallelization potential on multi-core systems, etc. Furthermore, a number of extensions and improvements of the standard formulation have been proposed through the years, including high-order formulations [6], subgridding techniques [7], perfectly-matched layers for absorbing boundary conditions [8], hybrid schemes with other discretization methods [9], thin-wire formulations [10], surface-impedance boundary conditions [11], etc.

T. T. Zygiridis (✉)
Department of Informatics and Telecommunications Engineering, University of Western Macedonia, Kozani, Greece
e-mail: tzygiridis@uowm.gr

N. V. Kantartzis
Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece
e-mail: kant@ece.auth.gr

T. M. Rassias (ed.), *Applications of Nonlinear Analysis*, Springer Optimization and Its Applications 134, https://doi.org/10.1007/978-3-319-89815-5_29

In its original formulation, the standard FDTD method is well-suited for linear EM problems. In fact, the vast majority of pertinent FDTD applications and the corresponding extensions have been mainly targeted towards cases without nonlinearities. On the other hand, several nonlinear EM problems are of significant engineering interest, including (but not restricted to) configurations operating at optical frequencies, such as optical fibers, switches, resonators, filters, couplers, multiplexers, and splitters, which constitute fundamental parts of modern and future communication, signal-processing, and transmission systems. The constantly increasing research regarding the aforementioned applications has triggered the development of FDTD approaches for the study of the corresponding nonlinear EM phenomena. Apart from the above-mentioned advantages, FDTD formulations do not require any special conditions for their application, in order to provide consistent results. For this reason, the full-wave FDTD solutions are commonly more preferable than other approximate (less generic) models, such as the nonlinear Schrödinger equation (an asymptotic envelope equation) [12] or the beam-propagation technique (based on the paraxial wave equation) [13].

More specifically, the most common case of EM nonlinearities pertains to the response of materials. In such cases, the material constitutive parameters exhibit a dependence on the intensity of the electric or the magnetic field. This behavior gives rise to a complicated form for the polarization **P**, which is related to the electric-field intensity **E** and dielectric displacement **D** via

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P} \tag{1}$$

where ϵ_0 is the electric permittivity of vacuum. The behavior of nonlinear media can be described via the following general formula for the polarization's components [14]:

$$P_i = \sum_j \chi_{ij}^{(1)} E_j + 2 \sum_{j,k} \chi_{ijk}^{(2)} E_j E_k + 4 \sum_{j,k,l} \chi_{ijkl}^{(3)} E_j E_k E_l + \ldots \tag{2}$$

where $\chi^{(1)}$ stands for the linear susceptibility and $\chi^{(2)}, \chi^{(3)}, \ldots$ denote nonlinear susceptibilities with increasing orders.[1] Two of the most characteristic pertinent phenomena are the Kerr effect and Raman scattering, which are related to the third-order nonlinear susceptibility models. Regarding the latter, the corresponding relation between the polarization and the electric-field intensity is described in the time domain by

$$\begin{aligned} \mathbf{P}_{NL}(\mathbf{r}, t) = \epsilon_0 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \chi^{(3)}(t - \tau_1, t - \tau_2, t - \tau_3) \mathbf{E}(\mathbf{r}, \tau_1) \\ \times \mathbf{E}(\mathbf{r}, \tau_2) \mathbf{E}(\mathbf{r}, \tau_3) d\tau_1 d\tau_2 d\tau_3 \end{aligned} \tag{3}$$

[1] In the simple—quite common—case of linear materials, it is $\mathbf{P} = \epsilon_0 \chi^{(1)} \mathbf{E}$.

which clearly poses significant challenges in terms of implementation in the context of an efficient time-domain computational scheme. Apart from that, other forms of nonlinearities will be also considered in the present analysis.

In this chapter, we provide a review of selected FDTD-related works that present formulations suitable for EM problems exhibiting nonlinear behavior in at least one of their aspects. Although the pertinent literature has become quite vast over the years, it is our opinion that the current collection covers a significant number of important contributions for nonlinear problems, and extends previous related reviews such as those found in [14–16]. Specifically, [15] summarizes the main extensions of the FDTD method to nonlinear optics up to 1997, while [14] presents a more general review of complex material models, where a small part is devoted to the works pertinent to nonlinear media. Furthermore, [16] compares the performance of two existing FDTD approaches that are suitable for modeling the instantaneous Kerr effect. Without proceeding to a high level of detail, we provide all the necessary information that one needs to have at hand, in order to understand the main gist of each methodology and recognize the applications that the examined approach is suitable for. In addition, a quick reference to the test problems that the examined methods were implemented to is given. With the current study, a sufficient description of the specific research area is provided, which can be useful for those not previously familiar with nonlinear EM problems. It can also serve as a starting point for researchers that wish to engage into the computational study of nonlinear EM phenomena using FDTD techniques. Before proceeding to the main part of this work, a short description of the original FDTD formulation that is suitable for linear problems is provided.

2 FDTD Discretization of Linear Electromagnetic Problems

We begin by briefly introducing the standard FDTD methodology for linear EM phenomena, which is currently considered a widely accepted numerical tool for performing reliable simulation studies. In the case of linear, isotropic, and non-dispersive materials, Maxwell's equations take the form

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} - \mathbf{M}_c - \mathbf{M}_s \tag{4}$$

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} + \mathbf{J}_c + \mathbf{J}_s \tag{5}$$

where $\mathbf{E}$ is the electric-field intensity, $\mathbf{H}$ is the magnetic-field intensity, $\mathbf{D} = \epsilon_r \epsilon_0 \mathbf{E}$ is the electric-flux density, $\mathbf{B} = \mu_r \mu_0 \mathbf{H}$ is the magnetic-flux density, $\mathbf{J}_c = \sigma \mathbf{E}$ and $\mathbf{M}_c = \tilde{\sigma} \mathbf{H}$ denote the electric and magnetic conductivity current densities, respectively, $\mathbf{J}_s$ and $\mathbf{M}_s$ denote the electric- and magnetic-current source terms, respectively, σ is the electric conductivity, $\tilde{\sigma}$ stands for the magnetic conductivity, ϵ_0 is the electric permittivity in free space, μ is the magnetic permeability in free space,

ϵ_r is the relative electric permittivity, and μ_r is the relative magnetic permeability. The standard FDTD formulation considers the field values located at nodes that are organized according to a dual staggered spatial grid, at distances of Δx, Δy, Δz along x, y, z axes, respectively. Specifically, the primary grid is used for the electric field components, with the following arrangement: E_x is located at $\left(i+\frac{1}{2}, j, k\right)$ nodes, E_y at $\left(i, j+\frac{1}{2}, k\right)$, and E_z at $\left(i, j, k+\frac{1}{2}\right)$ nodes.[2] In a similar fashion, the magnetic-field components are located at nodes of the secondary grid, and are organized according to: H_x on $\left(i, j+\frac{1}{2}, k+\frac{1}{2}\right)$, H_y on $\left(i+\frac{1}{2}, j, k+\frac{1}{2}\right)$, and H_z on $\left(i+\frac{1}{2}, j+\frac{1}{2}, k\right)$. The time axis is also discretized with steps equal to Δt, with the electric components computed at time instants described by $(n\Delta t)$, while the magnetic components are computed at $\left(n+\frac{1}{2}\right)$ time instants.

In order to discretize Maxwell's equations, second-order finite-difference approximations are implemented for both space and time derivatives. Consider, for example, the equation

$$\epsilon \frac{\partial E_x}{\partial t} = \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} - \sigma E_x - J_{sx} \tag{6}$$

which is one of the six scalar equations that the two vector equations (4), (5) can be decomposed. The approximating formulae for the three appearing derivatives at node $\left(i+\frac{1}{2}, j, k\right)$ and time instant $\left(n+\frac{1}{2}\right)$ are:

$$\frac{\partial E_x}{\partial t} \simeq \frac{E_x|^{n+1}_{i+\frac{1}{2},j,k} - E_x|^{n}_{i+\frac{1}{2},j,k}}{\Delta t} \tag{7}$$

$$\frac{\partial H_z}{\partial y} \simeq \frac{H_z|^{n+\frac{1}{2}}_{i+\frac{1}{2},j+\frac{1}{2},k} - H_z|^{n+\frac{1}{2}}_{i+\frac{1}{2},j-\frac{1}{2},k}}{\Delta y} \tag{8}$$

$$\frac{\partial H_y}{\partial z} \simeq \frac{H_y\Big|^{n+\frac{1}{2}}_{i+\frac{1}{2},j,k+\frac{1}{2}} - H_y\Big|^{n+\frac{1}{2}}_{i+\frac{1}{2},j,k-\frac{1}{2}}}{\Delta z} \tag{9}$$

In addition, the following averaging formula for the conductivity current is introduced:

$$\sigma E_x \simeq \frac{1}{2} \sigma|_{i+\frac{1}{2},j,k} \left(E_x|^{n+1}_{i+\frac{1}{2},j,k} + E_x|^{n}_{i+\frac{1}{2},j,k} \right) \tag{10}$$

[2] The convention $f\,(i\Delta x, j\Delta y, k\Delta z, n\Delta t) = f|^n_{i,j,k}$ is used in this work.

Once the aforementioned expressions are substituted in (6), the update equation for E_x at each time instant is obtained:

$$\begin{aligned} E_x\big|^{n+1}_{i+\frac{1}{2},j,k} &= \frac{2\epsilon|_{i+\frac{1}{2},j,k} - \sigma|_{i+\frac{1}{2},j,k}\Delta t}{2\epsilon|_{i+\frac{1}{2},j,k} + \sigma|_{i+\frac{1}{2},j,k}\Delta t}\, E_x\big|^{n}_{i+\frac{1}{2},j,k} \\ &+ \frac{2\Delta t}{\Delta y\left(2\epsilon|_{i+\frac{1}{2},j,k} + \sigma|_{i+\frac{1}{2},j,k}\Delta t\right)}\left(H_z\big|^{n+\frac{1}{2}}_{i+\frac{1}{2},j+\frac{1}{2},k} - H_z\big|^{n+\frac{1}{2}}_{i+\frac{1}{2},j-\frac{1}{2},k}\right) \\ &- \frac{2\Delta t}{\Delta z\left(2\epsilon|_{i+\frac{1}{2},j,k} + \sigma|_{i+\frac{1}{2},j,k}\Delta t\right)}\left(H_y\big|^{n+\frac{1}{2}}_{i+\frac{1}{2},j,k+\frac{1}{2}} - H_y\big|^{n+\frac{1}{2}}_{i+\frac{1}{2},j,k-\frac{1}{2}}\right) \\ &- \frac{2\Delta t}{2\epsilon|_{i+\frac{1}{2},j,k} + \sigma|_{i+\frac{1}{2},j,k}\Delta t}\, J_{s,x}\big|^{n+\frac{1}{2}}_{i+\frac{1}{2},j,k} \end{aligned} \tag{11}$$

The same procedure is applied to the remaining five equations, so that an equal number of discrete update formulae is derived, which determine the problem's interior discretization scheme. In this way, the required components are calculated in an explicit fashion, without necessitating—computationally expensive—system solutions in every iteration. Furthermore, the update procedure is conditionally stable, with the time-step size bounded by the well-known stability limit

$$\Delta t \leq \frac{1}{c_0\sqrt{\frac{1}{\Delta x^2} + \frac{1}{\Delta y^2} + \frac{1}{\Delta z^2}}} \tag{12}$$

where $c_0 = 1/\sqrt{\mu\epsilon}$ is the free-space speed of light.

As far as the boundary conditions are concerned, these are implemented according to the physics of the problem under consideration. For instance, absorbing boundary conditions are applied in the case of open (radiation) problems, homogeneous Dirichlet conditions in case of waves guided by metallic structures, periodic conditions for infinite configurations, symmetry conditions for size reduction of the computational domain, etc.

3 FDTD Methodologies for Nonlinear Problems

3.1 Integration of Nonlinear Maxwell's Equations in 1D Setups

The first attempts to exploit the FDTD algorithm for the solution of the nonlinear Maxwell's equations can be traced back to 1992 and the works reported in [17, 18]. Specifically, the methodology presented therein investigates the simple case of

1D wave propagation along the x-axis considering nonlinear instantaneous effects, where E_z and H_y are the components of the electric-and magnetic-field intensities, respectively. In this manner, the study of optical solitons with extended bandwidths becomes possible. The equations that describe the problem under investigation are:

$$\mu_0 \frac{\partial H_y}{\partial t} = \frac{\partial E_z}{\partial x} \tag{13}$$

$$\frac{\partial D_z}{\partial t} = \frac{\partial H_y}{\partial x} \tag{14}$$

$$D_z = \epsilon_0 \epsilon_\infty E_z + P_z \tag{15}$$

where ϵ_∞ is the relative permittivity at infinite frequency (the rest of the terms have been explained previously). The polarization comprises a linear and a nonlinear part:

$$P_z = P_z^L + P_z^{NL} \tag{16}$$

The (first-order) linear term is described by

$$P_z^L = \int_{-\infty}^{+\infty} \chi^{(1)}(t-\tau) E_z(x,\tau)\, d\tau \tag{17}$$

where $\chi^{(1)}$ is the first-order susceptibility function. For the latter, Lorentz linear dispersion is considered in [18], according to

$$\chi^{(1)}(t) = \frac{\omega_p^2}{v_0} e^{-\delta t/2} \sin(v_0 t) \tag{18}$$

with the corresponding permittivity described by

$$\epsilon(\omega) = \epsilon_\infty + \chi^{(1)}(\omega) = \epsilon_\infty + \frac{\omega_0^2 (\epsilon_s - \epsilon_\infty)}{\omega_0^2 - j\delta\omega - \omega^2} \tag{19}$$

where $\omega_p^2 = \omega_0^2 (\epsilon_s - \epsilon_\infty)$ and $\nu_0^2 = \omega_0^2 - \delta^2/4$. The non-linear term P_z^{NL} depends on the third-order susceptibility function according to [19]

$$P_z^{NL} = \epsilon_0 \chi^{(3)} E_z(x,t) \int_{-\infty}^{+\infty} g(t-\tau) E_z^2(x,\tau)\, d\tau \tag{20}$$

where

$$g(t) = \alpha\delta(t) + (1-\alpha) g_R(t) \tag{21}$$

$$g_R(t) = \frac{\tau_1^2 + \tau_2^2}{\tau_1 \tau_2^2} e^{-\frac{t}{\tau_2}} \sin\left(\frac{t}{\tau_1}\right) \tag{22}$$

In (21), the delta function represents Kerr non-resonant virtual electronic transitions up to 1 fs, while the remaining part models transient Raman scattering. Parameter α acts as a relative weight for Kerr and Raman interactions. After defining the functions

$$F(t) = \epsilon_0 \int_0^t \chi^{(1)}(t-\tau) E_z(x,\tau)\, d\tau \tag{23}$$

$$G(t) = \epsilon_0 \int_0^t g_R(t-\tau) E_z^2(x,\tau) d\tau \tag{24}$$

and differentiating the above two formulae, a couple of second-order nonlinear equations is obtained, which are discretized with conventional finite differences and solved simultaneously (Eqs. (13) and (14) are dealt with in a standard manner). For instance, the equation pertinent to F is

$$\frac{1}{\omega_0^2}\frac{d^2F}{dt^2} + \frac{\delta}{\omega_0^2}\frac{dF}{dt} + \left(1 + \frac{\epsilon_s - \epsilon_\infty}{\epsilon_\infty + \alpha\chi^{(3)}E_z^2}\right)F + \frac{(\epsilon_s - \epsilon_\infty)(1-\alpha)\chi^{(3)}E_z}{\epsilon_\infty + \alpha\chi^{(3)}E_z^2}G = \frac{\epsilon_s - \epsilon_\infty}{\epsilon_\infty + \alpha\chi^{(3)}E_z^2}D_z \tag{25}$$

whose finite-difference analogue is

$$\begin{aligned}
&\frac{1}{\omega_0^2}\frac{F|_i^{n+1} - 2\,F|_i^n + F|_i^{n-1}}{\Delta t^2} + \frac{\delta}{\omega_0^2}\frac{F|_i^{n+1} - F|_i^{n-1}}{2\Delta t} \\
&+ \left(1 + \frac{\epsilon_s - \epsilon_\infty}{\epsilon_\infty + \alpha\chi^{(3)}\ E_z^2|_i^n}\right)\frac{F|_i^{n+1} + F|_i^{n-1}}{2} \\
&+ \frac{(\epsilon_s - \epsilon_\infty)(1-\alpha)\chi^{(3)}\ E_z|_i^n}{\epsilon_\infty + \alpha\chi^{(3)}\ E_z^2|_i^n}\frac{G|_i^{n+1} + G|_i^{n-1}}{2} \\
&= \frac{\epsilon_s - \epsilon_\infty}{\epsilon_\infty + \alpha\chi^{(3)}\ E_z^2|_i^n}\frac{D_z|_i^{n+1} + D_z|_i^{n-1}}{2}
\end{aligned} \tag{26}$$

A similar result is obtained for the update of G. Evidently, values at two previous time-steps need to be stored, and the latest value of E_z is obtained from

$$E_z = \frac{D_z - F - (1-\alpha)\,\chi^{(3)}E_zG}{\epsilon_0\left(\epsilon_\infty + \alpha\chi^{(3)}E_z^2\right)} \tag{27}$$

The aforementioned nonlinear equation can be handled via an iterative Newton procedure. The propagation of a single soliton, as well as the collision of two solitons moving in different directions were simulated for the first time with this methodology in [18].

3.2 Multi-Dimensional Formulation with Applications

A FDTD methodology that solves Maxwell's equations in a multi-dimensional (2D) framework is presented in [20], where Lorentz linear dispersion as well as Raman nonlinearity models are incorporated. Specifically, the former is described by

$$\frac{\partial^2 \mathbf{P}^L}{\partial t^2} + \Gamma_L \frac{\partial \mathbf{P}^L}{\partial t} + \omega_L^2 \mathbf{P}^L = \epsilon_0 \chi_0 \omega_L^2 \mathbf{E} \tag{28}$$

while the latter is described by

$$\frac{\partial^2 \chi^{NL}}{\partial t^2} + \Gamma_R \frac{\partial \chi^{NL}}{\partial t} + \omega_R^2 \chi^{NL} = \epsilon_R \omega_R^2 |\mathbf{E}|^2 \tag{29}$$

considering that $\mathbf{P} = \mathbf{P}^L + \mathbf{P}^{NL} = \mathbf{P}^L + \epsilon_0 \chi^{NL}\mathbf{E}$. The aforementioned equations are combined with

$$\frac{\partial}{\partial t}(\mu_0 \mathbf{H}) = -\nabla \times \mathbf{E} \tag{30}$$

$$\frac{\partial}{\partial t}(\epsilon_L \mathbf{E}) = \nabla \times \mathbf{H} - \frac{\partial \mathbf{P}}{\partial t} \tag{31}$$

where ϵ_L is the linear permittivity. The proposed formulation differentiates from earlier works that rely on the introduction of the effective quantities

$$\epsilon_{eff} = \epsilon_L + \epsilon_0 \chi^{NL}, \quad \sigma_{eff} = \epsilon_0 \frac{\partial}{\partial t} \chi^{NL} \tag{32}$$

The proposed methodology was applied in a problem of scattering of a pulsed Gaussian beam, which is normally incident on a linear-nonlinear interface. It was noted that propagation within the nonlinear medium resulted in self-focusing of the beams, and the linear diffraction region and nonlinear effects were identified via the intensity patterns in the focus region.

3.3 Transient Analysis Within a Nonlinear Magnetic Sheet

The FDTD solution of problems involving wave propagation within a significantly conducting nonlinear magnetic medium is discussed in [21]. In essence, the case of a saturable ferromagnetic material is considered, whose differential permeability that describes the $B - H$ characteristic curve is modeled by

$$d\mu(H) = \frac{\partial B}{\partial H} = \mu_m + \frac{B_s}{H_c} e^{-|H|/H_c} \tag{33}$$

where μ_m, B_s, H_c are known constants. The presence of the magnetic nonlinearity affects only the magnetic-field update equation,

$$\frac{\partial E_z}{\partial y} = -\frac{\partial B_x}{\partial t} = -d\mu(H_x)\frac{\partial H_x}{\partial t} \tag{34}$$

which is easily transformed into a FDTD update equation using standard approximations:

$$H_x|_j^{n+1} = H_x|_j^n - \frac{\Delta t}{d\mu\Big(H_x|_j^n\Big)\,\Delta y}\left(E_z|_{j+\frac{1}{2}}^{n+\frac{1}{2}} - E_z|_{j-\frac{1}{2}}^{n+\frac{1}{2}}\right) \tag{35}$$

Owing to the high conductivity of the considered medium, the current density term in the corresponding discrete equation is computed from the most recent value of E, in order to ensure stable simulations. Furthermore, to avoid computations in free space that would weaken the stability, the fields are calculated only inside the magnetic material. Such an approach necessitates the implementation of proper boundary conditions, which are translated into second-order, one-sided, finite-difference formulae, e.g.

$$H_x|_0^n = \frac{1}{F}\left[2\,E_{inc}|^n + \frac{1}{2\sigma\Delta y}\left(4\,H_x|_1^n - H_x|_2^n\right)\right] \tag{36}$$

where $F = \eta_0 + 3/(2\sigma\Delta y)$ ($\eta_0 = \sqrt{\mu_0/\epsilon_0}$). It was also shown that the nonlinearity plays the most significant role in the reduction of the maximum allowable time-step size. In fact, due to the combination of the material's nonlinearity and high conductivity, the time-step size had to be reduced by 10–30 times with respect to the Courant stability criterion, with smaller time-steps needed when the material is governed by more intense nonlinearity.

3.4 Incorporating Active Device Models

The work of Toland et al. [22] presents a methodology that enables simulating realistic devices incorporating active and nonlinear regions. Let us consider an active device, like a resonant tunneling diode, with a junction capacitance C, a series resistance R and a nonlinear current source F. The latter are translated into the distributed parameters E and $F(Vs)$, considered for each grid cell. This actually implies the cells within the active region can be treated as a voltage source, and the total effect from all nodes in the active region is characteristic of the physical device. It is well-known that voltage/current sources can be introduced into an FDTD algorithm without difficulty, yet with a special attention to the overall stability of the resulting scheme. Hence, we start from the fact that the voltage source is a solution of

$$\frac{dV_s}{dt} + \frac{V_s}{RC} + \frac{F(V_s)}{C} = -\frac{V_{in}}{RC} \tag{37}$$

For this equation to be solved, forward time average differencing can be utilized. To this aim, the nonlinear current source is expanded in a Taylor series, as

$$F(V_s|^{n+1}) \approx F(V_s|^n) + \frac{dF(V_s|^n)}{dV}\left[V_s|^{n+1} - V_s|^n\right] \tag{38}$$

leading to

$$V_s|^{n+1} = A_1\, V_s|^n - A_2 F(V_s|^n) - A_3 \Delta y \left[E_y\big|^{n+1} + E_y\big|^n\right] \tag{39}$$

where

$$A_1 = \frac{2RC - \Delta t\,[1 - RdF(V_s)/dV]}{\beta}, \quad A_2 = \frac{2R\Delta t}{\beta}, \quad A_3 = \frac{\Delta t}{\beta} \tag{40}$$

with $\beta = 2RC + \Delta t\,[1 + RdF(V_s)/dV]$. This outcome can then be plugged into the FDTD algorithm. For example, bearing in mind a y-directed source, E_y is updated according to

$$\begin{aligned} E_y\big|^{n+1} &= \frac{\epsilon/\Delta t - 0.5(1 - A_3)/R}{\epsilon/\Delta t + 0.5(1 - A_3)/R}\, E_y\big|^n \\ &+ \frac{1}{\Delta z\,[\epsilon/\Delta t + 0.5(1 - A_3)/R]}\left(H_x|^{n+1/2}_{i,j+\frac{1}{2},k+\frac{1}{2}} - H_x|^{n+1/2}_{i,j+\frac{1}{2},k-\frac{1}{2}}\right) \\ &+ \frac{1}{\Delta x\,[\epsilon/\Delta t + 0.5(1 - A_3)/R]}\left(H_z|^{n+1/2}_{i+\frac{1}{2},j+\frac{1}{2},k} - H_z|^{n+1/2}_{i-\frac{1}{2},j+\frac{1}{2},k}\right) \\ &- \frac{1}{2R\Delta y\,[\epsilon/\Delta t + 0.5(1 - A_3)/R]}\left[(1 + A_1)\, V_s|^n - A_2 F(V_s|^n)\right] \end{aligned} \tag{41}$$

The use of (39) in conjunction with (41), guarantees that no instabilities are to be generated due to the nonlinear nature of the involved elements.

3.5 A Discrete Model for Magnetic Diffusion Problems

A possible alternative for problems involving slowly changing waveforms or extended diffusion times is developed in [23]. Such problems pose certain computational difficulties, due to the limiting Courant stability condition. As a study example, we may consider an aluminum enclosure. At first, a transient field will cause the appearance of eddy currents on the enclosure, which should lead to the cancellation of the external excitation. If the aluminum were considered to be a perfect conductor, the eddy currents would be characterized by long duration, similar to that of the transient field. Of course, these eddy currents are practically characterized by a specific time constant, which depends on a number of factors, such as the enclosure volume, the thickness and the conductivity of the wall. The decay time can be of the order of 0.1 s, in case of an aluminum enclosure the size of a large truck. If the duration of a magnetic pulse is much longer than this value, the enclosure will behave as being magnetically transparent. However, before this transparency takes place, any magnetic memory inside the enclosure will be scrambled. In essence, the magnetic field penetrating the enclosure cannot be treated solely in the context of a simple diffusion problem, at least in three dimensions. In fact, when the magnetic field develops a component that is normal to the surface of the enclosure, the corresponding problem should not be characterized as diffusive. Furthermore, the fields in the enclosure satisfy Maxwell's system, and the appearance of eddy currents are a direct result of this property.

Problems involving steel enclosures are not much different than those with aluminum enclosures. The main difference is that the steel's permeability is likely to prevent it from establishing a full magnetic transparency. Furthermore, nonlinear effects will take place, considering the nonlinear relationship between B, H. On the other hand, both enclosures will initially exclude interior magnetic fields, and will allow magnetic fields to slowly appear into the interior. For studying these phenomena, an implicit FDTD approach has been suggested.

Let us now consider a wave traveling along $+x$, with only H_y and E_z components. As usual, air is represented by (ϵ_0, μ_0) and walls by (ϵ, μ, σ). Moreover, E will be computed at the air-wall interface. Regarding the grid, the x-axis is divided into nodal points separated in air by $\Delta x/2$, while the spatial resolution in the walls becomes $\delta x/2$. In air, the equation

$$\frac{\partial H_y}{\partial x} = -\epsilon_0 \frac{\partial E_z}{\partial t} \tag{42}$$

can take the Crank-Nicolson FDTD form

$$E|_i^{n+1} = E|_i^n - \frac{\Delta t}{\epsilon_0 \Delta x}\left[\lambda\left(H|_{i+1}^{n+1} - H|_{i-1}^{n+1}\right) + (1-\lambda)\left(H|_{i+1}^{n} - H|_{i-1}^{n}\right)\right] \tag{43}$$

Here, λ is a parameter that is set between $1/2$ (center differences) and 1 (forward differences) for implicit schemes. Equation (43) can be written in tridiagonal form, as

$$-a_i\ H|_{i+1}^{n+1} + b_i\ E|_i^{n+1} - c_i\ H|_{i-1}^{n+1} = d_i \tag{44}$$

where

$$a_i = -\frac{\lambda \Delta t}{\epsilon_0 \Delta x}, \quad b_i = 1, \quad c_i = \frac{\lambda \Delta x}{\epsilon_0 \Delta t}, \quad d_i = E|_i^n - (1-\lambda)\frac{\lambda \Delta t}{\mu_0 \Delta x}(H|_{i+1}^n - H|_{i-1}^n) \tag{45}$$

A corresponding set of equations can be deduced at even i with E, H, and μ_0, ϵ_0 interchanged. For $\lambda \neq 0$, the two equation families allow all field samples to be calculated simultaneously. If $\lambda = 0$ is selected, (43) becomes an explicit

FDTD expression. In this case, $E|_i$ and $H|_i$ cannot be all advanced at the same time. Moreover, at positions within the wall, (42) is modified by adding the term $-\sigma E$ on the right side, and (44) needs to be altered by replacing (a, b, c, d) with (A, B, C, D), where

$$A_i = -\left(\frac{\epsilon}{\Delta t} - \Lambda\sigma\right)^{-1} \frac{\lambda}{\delta x} = -C_i, \qquad B_i = 1,$$

$$D_i = \frac{\frac{\epsilon}{\Delta t} - (1 - \Lambda)\sigma}{\frac{\epsilon}{\Delta t} + \Lambda\sigma} E|_i^n - \frac{1}{\left(\frac{\epsilon}{\Delta t} + \Lambda\sigma\right)\delta x}\left[(1-\lambda)\left(H|_{i+1}^n - H|_{i-1}^n\right)\right] \tag{46}$$

Here, Λ is another parameter that can be adjusted. It is noted that the coefficients in (46) will produce numerical instabilities for "explicit" Λ values, if Δt is selected higher than the Courant limit. Therefore, a consistent choice is to select Λ between 1/2 (center-differenced loss) and 1 (forward-differenced loss).

3.6 Calculation of Photonic Band Structures

The FDTD methodology of [24] is suitable for studying the photonic band structure of a dielectric material, when Kerr-type nonlinearities are present. The algorithm assumes field distributions (**B** and **E**) that are characterized by the wave vectors **k**. The time-dependent Maxwell's system is integrated to provide $B(\mathbf{k}, t)$. For a specific **k**, Maxwell's equations will also determine the appropriate frequency value. In the case of nonlinear Kerr media, it is reminded that $\mathbf{D} = \left(\epsilon + \chi|\mathbf{E}|^2\right)\mathbf{E}$. An analytical solution for **E** is available for such a system. First, we take the square of the magnitude the aforementioned formula, to find an equation for χ, i.e.

$$|\chi|^2 x^3 + 2\,\mathrm{Re}\{\epsilon * \chi\} x^2 + |\epsilon|^2 x - |\mathbf{D}|^2 = 0 \tag{47}$$

The solution to this cubic equation is known and its knowledge extremely useful, as it can save computational time. The main question pertains to which root to consider, among the available ones. The case we are studying (ϵ, χ have the same sign) is quite simple, as there exists one positive real root, and x must be a positive real. After x has been determined, **E** is given by

$$\mathbf{E} = \frac{\mathbf{D}}{\epsilon + \chi x} \tag{48}$$

Computations are carried out by solving the system on a 3D lattice, with the curl computed in **k** space. Such an approach can be more reliable than the standard finite-difference approximation of the curl normally used in the FDTD method. Furthermore, implementation of a staggered spatial grid (which adds programming complexity) is avoided, and the periodic boundary condition is automatically fulfilled, without requiring further modifications. The calculations evolve as follows (linear case):

$$\mathbf{B}(\mathbf{k}, t + \Delta t) = \mathbf{B}(\mathbf{k}, t) - jc\Delta t\mathbf{k} \times \mathbf{E}(\mathbf{k}, t + \Delta t/2) \tag{49}$$

$$\mathbf{D}(\mathbf{k}, t + \Delta t/2) = \mathbf{D}(\mathbf{k}, t - \Delta t/2) + jc\Delta t\mathbf{k} \times \mathbf{B}(\mathbf{k}, t + \Delta t) \tag{50}$$

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{D}(\mathbf{r}, t)/\epsilon(\mathbf{r}) \tag{51}$$

while (51) is replaced by the appropriate equation, for the nonlinear case. The split-time scheme not only assures higher accuracy, but also is more economical in terms of memory, as it is not necessary to preserve the field values at two subsequent steps. Finally, $\mathbf{B}(\mathbf{k}, t)$ is Fourier transformed to provide $\mathbf{B}(\mathbf{k}, \omega)$. To ensure that all modes have been considered, $\mathbf{B}(\mathbf{k}, t)$ is multiplied with a Gaussian function, with a width half of that of $1/\Delta\omega$ ($\Delta\omega$ is the frequency resolution). This action broaden the peaks. As a source-free solution is required, the two divergence conditions need to be satisfied. However, we still may select an arbitrary initial condition, as the corresponding fields at later times will automatically satisfy the divergence conditions (this can be verified by taking the divergence of Maxwell's equations). If the initial fields display nonzero divergence, they can be considered as fields originating from static charges. Hence, they will be identified only at $\omega = 0$. Note that the resolution ($\Delta\mathbf{k}$ and $\Delta\omega$) of the calculation is mainly affected by two factors: the grid size and the duration of integration.

3.7 Z-Transform-Based FDTD Formulation

The study of nonlinear phenomena using the FDTD method is also the subject of [25], where techniques borrowed from digital filtering theory are implemented. Similar to other works, the Kerr effect is described by the polarization term

$$P_K(t) = \epsilon_0 \chi_0^{(3)} \alpha E^3(t) \tag{52}$$

and the term due to Raman scattering is given by

$$P_R(t) = \epsilon_0 \chi_0^{(3)} (1 - \alpha) E(t) \int_0^t g_R(t - \tau) E^2(\tau)\, d\tau \tag{53}$$

where

$$g_R(\omega) = \frac{1}{1 + j2\delta_{NL}\left(\frac{\omega}{\omega_{NL}}\right) - \left(\frac{\omega}{\omega_{NL}}\right)^2} \tag{54}$$

By defining the integral

$$I_R(t) = \epsilon_0 \chi_0^{(3)} (1 - \alpha) \int_0^t g_R(t - \tau) E^2(\tau)\, d\tau \tag{55}$$

the implementation of the Z-transform yields

$$I_R(z) = \epsilon_0 \frac{\gamma_R \Delta t e^{-\alpha_R \Delta t} \sin(\beta_R \Delta t) z^{-1}}{1 - 2e^{-\alpha_R \Delta t} \cos(\beta_R \Delta t) z^{-1} + e^{-2\alpha_R \Delta t} z^{-2}} E^2(z) \tag{56}$$

where $\alpha_R, \beta_R, \gamma_R$ are known constant coefficients. If the compact notation $I_R(z) = \epsilon_0 z^{-1} S_R(z)$ is introduced (which is equivalent to $I_R|^n = \epsilon_0 S_R|^{n-1}$ in the time domain), then the update of S_R is performed via

$$S_R|^n = cnl_1\, S_R|^{n-1} - cnl_2\, S_R|^{n-2} + cnl_3 \left(E^2\right)\Big|^n \tag{57}$$

where cnl_1, cnl_2, cnl_3 are known coefficients. As the implementation of time discretization dictates

$$P_R(t) = E(t) I_R(t) \leftrightarrow P_R|^n = E|^n\, I_R|^n$$

then $P_R|^n$ is computed from the preceding value of S_R:

$$P_R|^n = \epsilon_0\, E|^n\, S_R|^{n-1} \tag{58}$$

and then the current value of the latter is obtained from E^2, according to (57).

Regarding the Kerr effect, after considering a proper Taylor expansion of E^3, we are led to an update equation of the form

$$\left(E|^n\right)^3 = 3\left(E|^{n-1}\right)^2 \left(E|^n\right) - 2\left(E|^{n-1}\right)^3 \tag{59}$$

which enables the computation of the corresponding polarization term $P_K|^n$ from (52). Evidently, the resulting equation dictates that P_K depends on the electric-field value at two different time-steps.

Finally, to calculate the E field from the individual polarization terms, we start from

$$\epsilon_0 \epsilon_\infty\, E|^n = D|^n - P_L|^n - P_R|^n - P_K|^n \tag{60}$$

where P_L denotes the linear polarization, for which a similar, Z-transform based approach, is applied. Once all substitutions have been made, and all $E|^n$ terms have been collected, we end up with

$$E|^n = \frac{\frac{1}{\epsilon_0} D|^n - S_L|^{n-1} + 2\chi_0^{(3)} \alpha \left(E|^{n-1}\right)^3}{\epsilon_0 + \chi_0^{(3)} (1-\alpha)\, S_R|^{n-1} + 3\chi_0^{(3)} \alpha \left(E|^{n-1}\right)^2} \tag{61}$$

The methodology of [25] was implemented in the 1D calculation of the reflection coefficient from a nonlinear material with reasonably good accuracy. Furthermore, the potential of this approach to simulate soliton propagation was exemplified, in case of pulses with sufficiently large amplitude.

3.8 Auxiliary-Differential-Equation Approach for Absorbing and Gain Media

A fully explicit FDTD methodology capable of modeling wave propagation in certain types of nonlinear media is presented in [26]. The proposed approach is similar to the auxiliary-differential-equation (ADE) method used for dispersive media, and incorporates the atomic rate equations, which correspond to the time evolution of the atomic energy level populations, when the effect of applied signals is taken into account. Consequently, nonlinear gain and absorption effects can be included, and the approach is reliable over a non-trivial range of different signal strengths.

First, it is shown in the considered work that the electric polarization in real atomic transitions satisfies

$$\frac{d^2\mathbf{P}}{dt^2} + \Delta\omega_a \frac{d\mathbf{P}}{dt} + \omega_\alpha^2\mathbf{P} = \kappa\Delta N\mathbf{E} \tag{62}$$

where $\Delta\omega_a$ is the total energy decay rate that corresponds to the actual linewidth of the transition, ω_a is the resonance frequency of the material that is related to the atomic energy levels, κ is a constant related to, among others, the mass and the charge of an electron, and ΔN represents the instantaneous population difference. The time variation of the latter becomes significant in case of high signal intensities and signals displaying rapid variations. Furthermore, for an ideal two-level system, the population difference ΔN satisfies

$$\frac{d\Delta N}{dt} = -\frac{2}{\hbar\omega_a}\mathbf{E}\cdot\frac{d\mathbf{P}}{dt} - \frac{\Delta N - \Delta N_0}{\tau_{21}} \tag{63}$$

where ΔN_0 is the population difference at thermal equilibrium, $\hbar$ denotes the reduced Planck's constant, and τ_{21} stands for the atoms' lifetime in the upper energy level.

Considering the simple case of 1D propagation, the standard finite-difference update equation of the E_x component is

$$E_x|_k^{n+1} = E_x|_k^{n+1} - \frac{\Delta t}{\epsilon_0\Delta z}\left(H_y|_{k+\frac{1}{2}}^{n+\frac{1}{2}} - H_y|_{k-\frac{1}{2}}^{n+\frac{1}{2}}\right) - \frac{1}{\epsilon_0}\left(P_x|_k^{n+1} - P_x|_k^n\right) \tag{64}$$

and the macroscopic polarization is obtained via

$$P_x|_k^{n+1} = \frac{2\Delta t^2}{2\Delta\omega_a\Delta t}\left[\kappa\ \Delta N|_k^n\ E_x|_k^n + \left(\frac{2}{\Delta t^2} - \omega_a^2\right) P_x|_k^n + \left(\frac{\Delta\omega}{2\Delta t} - \frac{1}{\Delta t^2}\ P_x|_k^{n-1}\right)\right] \tag{65}$$

In a similar fashion, the equation regarding $\varDelta N$ is discretized as follows:

$$\varDelta N|_k^{n+1} = \frac{2\tau_{12}\varDelta t}{2\tau_{21} + \varDelta t}\left[\varDelta N|_k^n \left(\frac{1}{\varDelta t} - \frac{1}{2\tau_{21}}\right)\right.$$
$$\left. + \frac{\varDelta N_0}{\tau_{21}} - \frac{\left(E_x|_k^{n+1} + E_x|_k^n\right)\left(P_x|_k^{n+1} - P_x|_k^n\right)}{\varDelta t \hbar \omega_a}\right] \tag{66}$$

To ensure numerical stability as well as satisfactory accuracy, the size of the time-step is determined by $\varDelta t \leq \frac{T_a}{100}$, where T_a is the time period related to the material resonance.

This approach was validated by considering a problem involving wave propagation in a two-level system of atoms, considering Gaussian-pulse excitation. Good agreement with existing theoretical models for the case of small-signal frequency response was observed. Moreover, population dynamics in the presence of considerable fields were also simulated satisfactorily using the aforementioned technique.

3.9 Hybrid Implicit-Explicit Modeling

A hybridization of two computational schemes is proposed in [27] for modeling 2D waveguiding structures, which is suitable for problems with small, nonlinear inclusions within larger, linear areas. The main feature of this algorithm is the partial elimination of the restrictive time-step stability limit, by implementing a partially implicit discretization approach in the nonlinear parts of the configuration under study.

If a 2D computational domain is described by a set of xz-axes, the E_y component satisfies the wave equation

$$\nabla_{xz}^2 E_y - \mu_0\epsilon_0 \frac{\partial^2}{\partial t^2}\left(\epsilon_r E_y\right) - \mu_0 \frac{\partial}{\partial t}\left(\sigma E_y\right) = 0 \tag{67}$$

where $\nabla_{xz}^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial z^2}$. In nonlinear regions, the Kerr-type nonlinearity is expressed via $\epsilon_r = \epsilon_{r,L} + \alpha\left|E_y\right|^2$. The discretization of (67) in linear areas is performed using a forward-difference formula for the first temporal derivative, and central difference approximations for the second temporal and spatial derivatives, resulting in the following expression:

$$\frac{\mu_0 \sigma|_{i,k}}{\varDelta t}\left(E|_{i,k}^{n+1} - E|_{i,k}^n\right) + \frac{\mu_0\epsilon_0 \epsilon_r|_{i,k}}{\varDelta t^2}\left(E|_{i,k}^{n+1} - 2\,E|_{i,k}^n + E|_{i,k}^{n-1}\right) =$$
$$= \frac{1}{\varDelta x^2}\left(E|_{i+1,k}^n - 2\,E|_{i,k}^n + E|_{i-1,k}^n\right) + \frac{1}{\varDelta z^2}\left(E|_{i,k+1}^n - 2\,E|_{i,k}^n + E|_{i,k-1}^n\right) \tag{68}$$

Hence, the linear problem is treated with an explicit scheme, whose stability is dictated by the standard limit. On the other hand, a weighted-averaging time-stepping process is applied in nonlinear regions, according to:

$$\begin{aligned}\frac{\mu_0}{\Delta t}\Big[\left(\sigma E_y\right)\Big|_{i,k}^{n+1}-\left(\sigma E_y\right)\Big|_{i,k}^{n}\Big]+\frac{\mu_0\epsilon_0}{\Delta t^2}\Big[\left(\epsilon_r E_y\right)\Big|_{i,k}^{n+1}-2\left(\epsilon_r E_y\right)\Big|_{i,k}^{n}+\left(\epsilon_r E_y\right)\Big|_{i,k}^{n-1}\Big]=\\ =\sum_{\ell,m_\ell} c_\ell\Bigg[\frac{1}{\Delta x^2}\left(E_y\Big|_{i+1,k}^{m_\ell}-2\,E_y\Big|_{i,k}^{m_\ell}+E_y\Big|_{i-1,k}^{m_\ell}\right)\\ +\frac{1}{\Delta z^2}\left(E_y\Big|_{i,k+1}^{m_\ell}-2\,E_y\Big|_{i,k}^{m_\ell}+E_y\Big|_{i,k-1}^{m_\ell}\right)\Bigg]\end{aligned} \tag{69}$$

where $\ell = 1, 2, 3$ and $m_1 = n-1, m_2 = n, m_3 = n+1$. As seen, the spatial derivatives can be averaged over three successive time-steps. Note that the explicit scheme is obtained simply by setting $c_2 = 1, c_1 = c_3 = 0$ in the aforementioned formula. In case of highly conducting materials, the partially implicit scheme with $c_1 = c_2/2 = c_3 = 1/4$ is selected instead, which exhibits better stability properties than the fully explicit approach and, at the same time, does not suffer from artificial amplitude attenuation, unlike the fully implicit method. Consequently, the overall stability is not affected and is still directly related to the discretization approach applied in the linear regions. To deal with the implicit updates, the authors in [27] implement the Newton-Raphson's iterative technique, using the field values at the current time instant as the initial guess for the subsequent time-step. The developed method was applied in problems involving either slab waveguides with weak or moderate nonlinearities, or nonlinear distributed Bragg resonators with 40 grating periods.

3.10 3D Optical Pulse Simulation Using a Moving Reference Frame

A finite-difference methodology that is particularly suited for the efficient 3D modeling of single-mode propagation in optical fibers over large distances is presented in [28]. Using normalized units, the dielectric displacement can be expressed in the following form:

$$\mathbf{D} = \epsilon_r \mathbf{E} + \mathbf{P}_L + \mathbf{P}_{NL} \tag{70}$$

where for the nonlinear polarization, the Kerr effect is modeled via $P_{NL} = \chi^{(3)} E^3$. As already mentioned, the proper manipulation of Taylor expansion of E^3 produces $(E|^n)^3 \simeq 3(E|^{n-1})^2 E|^n - 2(E|^{n-1})^3$, a formula which was also introduced in [25]. We eventually end up with the expression

$$E|^n = \frac{D|^n + 2\chi_0^{(3)}\left(E|^{n-1}\right)^3}{\epsilon_r + 3\chi_0^{(3)}\left(E|^{n-1}\right)^2} \tag{71}$$

Note that the linear polarization term is computed with a formula similar to (61).

The implementation of the proposed methodology exploits several strategies, in an attempt to improve the algorithm's overall efficiency. First, the spatial mesh density is selected to be higher along the direction of propagation, and reduced in the transverse direction, where the pulses are expected to change much slower. Second, the symmetry of the transverse field is taken into consideration, and only one quadrant of the computational space actually needs to be simulated, after applying the proper boundary conditions. Finally, the pulse is always held in the middle of the computational domain, by computing the average position of the pulse on the axis located at the center of the fiber's core and properly displacing the field values in the mesh, with respect to a pre-selected spatial buffer. Due to the moving reference frame, a wavelet transform is applied, in order to track the changes of the pulse's shape, while a Fourier transform at the central pulse frequency is implemented, for the assessment of the pulse's speed and attenuation.

3.11 Decoupled FDTD Algorithms for 2D Photonic Crystals

The methodology presented in [29] is suitable for the analysis of arbitrary 2D structured material configurations, which lead to second-harmonic (SH) generation.[3] In essence, an artificial separation of the fundamental field (FF) and the SH is accomplished, which partially sacrifices generality, but enables less time-consuming simulations.

Considering an incident H-polarized FF, standard (linear) FDTD updating is implemented for the calculation for the FF. The nonlinearity is not considered for the FF, but only for the SH, which is not associated to the FF. The SH field is updated according to

$$\begin{aligned} E_z^{\mathrm{SH}}\Big|_{i,j}^{n+1} &= E_z^{\mathrm{SH}}\Big|_{i,j}^{n} \\ &+ \frac{\Delta t}{\epsilon^{\mathrm{SH}}\big|_{i,j}\Delta}\left(H_y^{\mathrm{SH}}\Big|_{i+\frac{1}{2},j}^{n+\frac{1}{2}} - H_y^{\mathrm{SH}}\Big|_{i-\frac{1}{2},j}^{n+\frac{1}{2}} + H_x^{\mathrm{SH}}\Big|_{i,j-\frac{1}{2}}^{n+\frac{1}{2}} - H_x^{\mathrm{SH}}\Big|_{i,j+\frac{1}{2}}^{n+\frac{1}{2}}\right) \\ &- \frac{1}{\epsilon^{\mathrm{SH}}\big|_{i,j}}\left(P_z^{(2)}\Big|_{i,j}^{n+\frac{1}{2}} - P_z^{(2)}\Big|_{i,j}^{n-\frac{1}{2}}\right) \end{aligned} \tag{72}$$

where the last term, which is proportional to the time-derivative of the second-order polarization, represents the nonlinearity source. The polarization is obtained from the FF values, in a manner that depends on the properties of the considered nonlinear material. For instance, the case of a defective nonlinear photonic crystal is examined in [29], where the polarization update has the form

[3] Second-harmonic generation is a phenomenon, according to which a wave within a nonlinear medium can produce another wave with twice the frequency of the former.

$$\begin{aligned} P_z^{(2)}\Big|_{i,j}^{n+\frac{1}{2}} = P_z^{(2)}\Big|_{i,j}^{n-\frac{1}{2}} + \frac{\epsilon_0\, d|_{i,j}}{2} \Bigg[& \left(E_x^{\mathrm{FF}}\Big|_{i+\frac{1}{2},j+1}^{n+\frac{1}{2}} \right)^2 + \left(E_x^{\mathrm{FF}}\Big|_{i+\frac{1}{2},j}^{n+\frac{1}{2}} \right)^2 \\ & - \left(E_x^{\mathrm{FF}}\Big|_{i+\frac{1}{2},j+1}^{n-\frac{1}{2}} \right)^2 - \left(E_x^{\mathrm{FF}}\Big|_{i+\frac{1}{2},j}^{n-\frac{1}{2}} \right)^2 \\ & + \left(E_y^{\mathrm{FF}}\Big|_{i,j+\frac{1}{2}}^{n+\frac{1}{2}} \right)^2 + \left(E_y^{\mathrm{FF}}\Big|_{i+1,j+\frac{1}{2}}^{n+\frac{1}{2}} \right)^2 - \left(E_y^{\mathrm{FF}}\Big|_{i,j+\frac{1}{2}}^{n-\frac{1}{2}} \right)^2 - \left(E_y^{\mathrm{FF}}\Big|_{i+1,j+\frac{1}{2}}^{n-\frac{1}{2}} \right)^2 \Bigg] \end{aligned} \tag{73}$$

where d stands for the nonlinear susceptibility. The examined structure supports waveguide modes at both FF and SH frequencies, as it is designed to exhibit photonic bandgaps in the proximity of the FF frequency in the case of H-polarization, and near the SH frequency for E-polarization. The configuration's transmission diagrams was computed, which displayed maxima at the expected wavelengths.

3.12 A High-Order Extension of the Nonlinear ADE-FDTD Technique

The work of [30] presents a reformulation of the ADE technique, for problems concerning optical pulse propagating within linear Lorentz and nonlinear Kerr and Raman media. The main difference compared to other approaches is that the ADE approach is applied to the polarization, rather than to the polarization currents. The technique was originally developed for linear cases only [7], and this work extended it to nonlinear cases as well. Furthermore, unlike the conventional formulation of the FDTD method, the authors of [30] introduced spatial approximations of the form

$$\frac{\partial f}{\partial u}\bigg|_i \simeq \frac{1}{\Delta u} \sum_{\ell} c_\ell \left(f|_{i+\ell+\frac{1}{2}} - f|_{i-\ell-\frac{1}{2}} \right) \tag{74}$$

where c_ℓ, $\ell = 0, 1, \ldots$ are the stencil coefficients of wavelet-based schemes that correspond to Deslauriers–Dubuc interpolating bases. The above-mentioned approximations are capable of ensuring higher-order accuracy for spatial derivatives. The paper considered third-order nonlinear polarization, consisting of both the Kerr (P_K) and the Raman (P_R) nonlinearities. The latter is treated with a simple ADE technique, which is also consistent with the implementation of the anisotropic perfectly matched layer (PML) absorbing boundary condition of [31]. As the electric-flux density is related to the electric-field intensity and the polarization terms via

$$D_y\Big|^{n+1} = \epsilon_0\epsilon_\infty E_y\Big|^{n+1} + P_D|^{n+1} + P_L|^{n+1} + P_K|^{n+1} + P_R|^{n+1} \tag{75}$$

(P_D, P_L correspond to Debye and Lorentz dispersion, respectively) the final form of the update equation, similar to [25], becomes

$$E_y\big|^{n+1} = \frac{D_y\big|^{n+1} - a_D\, P_D|^n - b_D\, E_y\big|^n - \, P_L|^{n+1}}{\epsilon_0 \left[\epsilon_\infty + b_D + \alpha \chi_0^{(3)} \left(E_y\big|^{n+1} \right)^2 + \, S|^{n+1} \right]} \tag{76}$$

where a_D, b_D are known coefficients, and S denotes an auxiliary variable related to Raman nonlinearity. Evidently, the aforementioned formula is nonlinear and needs to be solved via an iterative scheme

The typical problems that were studied for numerical verification considered mainly 2D geometries and demonstrated spatio-temporal soliton propagation in optical media. In addition, the computational savings due to the high-order approximations in terms of memory requirements and computing times were clearly demonstrated.

3.13 A Vector ADE-FDTD Method for Nonlinear Problems

A FDTD method that features a general vector ADE approach is developed in [32] for 2D setups, where the electric field does not feature just a single vector component, and is suitable for propagation problems in dispersive nonlinear materials. The polarization current is considered to comprise three terms, $\mathbf{J} = \mathbf{J}_{\text{Lorentz}} + \mathbf{J}_{\text{Kerr}} + \mathbf{J}_{\text{Raman}}$. The linear Lorentz polarization model is the sum of contributions from different resonances,

$$\mathbf{J}_{\text{Lorentz}} = \sum_{\ell=1}^{3} \mathbf{J}_{\text{Lorentz},\ell} \tag{77}$$

In phasor representation, each term can be written as

$$\dot{\mathbf{J}}_{\text{Lorentz},\ell} = \epsilon_0 \beta \ell \omega_\ell^2 \frac{j\omega}{\omega_\ell^2 - \omega^2} \dot{\mathbf{E}} \tag{78}$$

After multiplying the above equation with $\left(\omega_\ell^2 - \omega^2\right)$, applying the inverse transform, and discretizing the resulting expression, we end up with

$$\mathbf{J}_{\text{Lorentz},\ell}\big|^{n+1} = \alpha_\ell\, \mathbf{J}_{\text{Lorentz},\ell}\big|^{n} - \mathbf{J}_{\text{Lorentz},\ell}\big|^{n-1} + \frac{\gamma_\ell}{2\Delta t}\left(\mathbf{E}|^{n+1} - \mathbf{E}|^{n-1}\right) \tag{79}$$

where α_ℓ, γ_ℓ are known coefficients. However, the current density is required at $\left(n + \frac{1}{2}\right)$ time instants, hence a simple averaging process is applied that yields

$$\mathbf{J}_{\text{Lorentz},\ell}\big|^{n+\frac{1}{2}} = \frac{1}{2}\left[(1+\alpha_\ell)\, \mathbf{J}_{\text{Lorentz},\ell}\big|^{n} - \mathbf{J}_{\text{Lorentz},\ell}\big|^{n-1} + \frac{\gamma_\ell}{2\Delta t}\left(\mathbf{E}|^{n+1} - \mathbf{E}|^{n-1}\right)\right] \tag{80}$$

Regarding the nonlinear Kerr polarization, it satisfies

$$\mathbf{J}_{\mathrm{Kerr}} = \frac{\partial \mathbf{P}_{\mathrm{Kerr}}}{\partial t} = \frac{\partial}{\partial t}\left(\alpha \epsilon_0 \chi_0^{(3)} |\mathbf{E}|^2 \mathbf{E}\right) \tag{81}$$

and it is discretized according to

$$\mathbf{J}_{\mathrm{Kerr}}|^{n+\frac{1}{2}} = \frac{\alpha \epsilon_0 \chi_0^{(3)}}{\Delta t}\left\{\left(\left|\mathbf{E}|^{n+1}\right|\right)^2 \mathbf{E}|^{n+1} - \left(\left|\mathbf{E}|^{n}\right|\right)^2 \mathbf{E}|^{n}\right\} \tag{82}$$

As far as the nonlinear Raman polarization is concerned, an auxiliary variable is introduced for the convolution

$$S(t) = \chi_{\mathrm{Raman}}^{(3)}(t) * |\mathbf{E}(t)|^2 \overset{FT}{\leftrightarrow} S(\omega) = \chi_{\mathrm{Raman}}^{(3)}(\omega) \mathscr{F}\left\{|\mathbf{E}(t)|^2\right\} \tag{83}$$

where

$$\chi_{\mathrm{Raman}}^{(3)}(\omega) = \frac{(1-\alpha)\chi_0^{(3)} \omega_{\mathrm{Raman}}^2}{\omega_{\mathrm{Raman}}^2 + 2j\omega\delta_{\mathrm{Raman}} - \omega^2} \tag{84}$$

and $\mathscr{F}$ denotes Fourier transform. Transforming back to the time domain produces

$$\frac{\partial^2 S}{\partial t^2} + 2\delta_{\mathrm{Raman}} \frac{\partial S}{\partial t} + \omega_{\mathrm{Raman}}^2 = (1-\alpha)\, \chi_0^{(3)} \omega_{\mathrm{Raman}}^2 |\mathbf{E}|^2 \tag{85}$$

which, when discretized, leads to the update equation

$$S|^{n+1} = \frac{2 - \omega_{\mathrm{Raman}}^2 \Delta t^2}{\delta_{\mathrm{Raman}} \Delta t + 1} S|^{n} + \frac{\delta_{\mathrm{Raman}} \Delta t - 1}{\delta_{\mathrm{Raman}} \Delta t + 1} S|^{n-1} + \frac{(1-\alpha)\chi_0^{(3)} \omega_{\mathrm{Raman}}^2 \Delta t^2}{\delta_{\mathrm{Raman}} \Delta t + 1}\left(\left|\mathbf{E}|^{n}\right|\right)^2 \tag{86}$$

Finally, the Raman polarization term at the time instant $\left(n + \frac{1}{2}\right)$ is updated according to

$$\mathbf{J}_{\mathrm{Raman}}|^{n+\frac{1}{2}} = \frac{\epsilon_0}{\Delta t}\left(\mathbf{E}|^{n+1} S|^{n+1} - \mathbf{E}|^{n} S|^{n}\right) \tag{87}$$

Taking into account all the aforementioned quantities, the update of the electric-field intensity at $(n+1)$ must be performed via

$$\nabla \times \mathbf{H}|^{n+\frac{1}{2}} - \frac{\epsilon_0}{\Delta t}\left(\mathbf{E}|^{n+1} - \mathbf{E}|^{n}\right) - \mathbf{J}_{\mathrm{Lorentz}}|^{n+\frac{1}{2}} - \mathbf{J}_{\mathrm{Kerr}}|^{n+\frac{1}{2}} - \mathbf{J}_{\mathrm{Raman}}|^{n+\frac{1}{2}} = 0 \tag{88}$$

Evidently, the aforementioned formula describes a nonlinear system of coupled equations. First, $E_x|^{n+1}$, $E_y\big|^{n+1}$ are updated from (80), (82), (87), and (88). Then, the authors of [32] suggest the implementation of a multi-dimensional Newton's method, and define an objective function vector, according to

$$\begin{aligned}\begin{bmatrix} x \\ y \end{bmatrix} &= \nabla \times \mathbf{H}|^{n+\frac{1}{2}} - \frac{\epsilon_0}{\Delta t}\left(\mathbf{E}|^{n+1} - \mathbf{E}|^{n}\right) \\ &\quad - \frac{1}{2}\sum_{\ell=1}^{3}\left[(1+\alpha_\ell)\,\mathbf{J}_{\text{Lorentz},\ell}\big|^{n} - \mathbf{J}_{\text{Lorentz},\ell}\big|^{n-1} + \frac{\gamma_\ell}{2\Delta t}\left(\mathbf{E}|^{n+1}-\mathbf{E}|^{n-1}\right)\right] \\ &\quad - \frac{\alpha\epsilon_0\chi_0^{(3)}}{\Delta t}\left\{\left(\left|\mathbf{E}|^{n+1}\right|\right)^2 \mathbf{E}|^{n+1} - \left(\left|\mathbf{E}|^{n}\right|\right)^2\mathbf{E}|^{n}\right\} + \frac{\epsilon_0}{\Delta t}\left(\mathbf{E}|^{n+1}S^{n+1} - \mathbf{E}|^{n}S^{n}\right)\end{aligned} \tag{89}$$

Next, if the m-th guesses for $E_x|^{n+1}$, $E_y\big|^{n+1}$ are represented by $e_x^{(m+1)}, e_y^{(m+1)}$, then Newton's approach updates the guesses according to

$$\begin{bmatrix} e_x^{(m+1)} \\ e_y^{(m+1)} \end{bmatrix} = \begin{bmatrix} e_x^{(m)} \\ e_y^{(m)} \end{bmatrix} - \mathbf{J}^{-1}\left(\begin{bmatrix} x \\ y \end{bmatrix}\right)\Bigg|^{(m)} \tag{90}$$

until both objective functions attain values that are sufficiently close to zero ($\mathbf{J}$ stands for the Jacobian $\partial(x, y)/\partial(e_x, e_y)$). Temporal and spatial solitons in dispersive nonlinear material were modeled, in the context of numerically demonstrating the potential of the suggested approach.

3.14 Nonlinear FDTD Approach with Exponential Integrators

Electromagnetic problems with general nonlinear polarizations are studied with a Krylov-subspace-based operator-exponential method in [33], by following different strategies for the different (linear, nonlinear) parts of the involved differential equations. Specifically, the linear part is treated with a high-accuracy approach, while the nonlinear part is evaluated by means of standard high-order techniques.

In order to apply such an strategy, the governing equations need to be expressed as

$$\frac{\partial}{\partial t}\boldsymbol{\Psi} = \mathscr{H}\boldsymbol{\Psi} + \mathscr{N}(\boldsymbol{\Psi}, t) \tag{91}$$

where $\boldsymbol{\Psi} = [\mathbf{E}\ \mathbf{H}]^{\mathrm{T}}$. The $\mathscr{H}$ part is treated via an accurate exponential integrator, while $\mathscr{N}$ that denotes the nonlinear behavior, is updated with a sufficiently accurate standard approach. Specifically, the two scaled ($c_0 = 1$) splitting terms are

$$\mathscr{H} = \begin{bmatrix} -\sigma_e & \frac{1}{\epsilon_r}\nabla\times \\ -\frac{1}{\mu_r}\nabla\times & -\sigma_m \end{bmatrix}, \qquad \mathscr{N}(\boldsymbol{\Psi}) = \begin{bmatrix} \left(C(\mathbf{E}) - \frac{1}{\epsilon}\right)\nabla\times\mathbf{H} \\ 0 \end{bmatrix}$$

where $C(\mathbf{E})$ is defined as follows: starting from the constitutive relation

$$\mathbf{D} = \epsilon_0 \left(\epsilon_r + \chi^{(3)}|\mathbf{E}|^2\right) \mathbf{E} \tag{92}$$

the corresponding temporal derivatives are related via

$$\frac{\partial \mathbf{E}}{\partial t} = C\left(\mathbf{E}\right) \frac{\partial \mathbf{D}}{\partial t}$$

where it is found that

$$C\left(\mathbf{E}\right) = \frac{\left[\epsilon_r^2 + 4\chi^{(3)}\epsilon_r|\mathbf{E}|^2 + 3\left(\chi^{(3)}\right)^2|\mathbf{E}|^4\right]\mathbf{I} - 2\chi^{(3)}\epsilon_{nl}\mathbf{A}}{\left(\epsilon_{nl}\right)^2\left(\epsilon_r + 3\chi^{(3)}|\mathbf{E}|^2\right)} \tag{93}$$

In the above equation, $\epsilon_{nl} = \epsilon_r + \chi^{(3)}|\mathbf{E}|^2$ and

$$\mathbf{A} = \begin{bmatrix} E_x^2 & E_xE_y & E_xE_z \\ E_xE_y & E_y^2 & E_yE_z \\ E_xE_z & E_yE_z & E_z^2 \end{bmatrix} \tag{94}$$

Note that the operator $\mathscr{H}$ does not depend on the electric or the magnetic field. The discretization of the linear and nonlinear parts can be performed via standard methodologies, such as finite differences, finite elements, etc.

The most crucial part of this methodology is the time integration of the involved equation. Based on a classical fourth-order Runge-Kutta scheme, the authors of [33] first propose the implementation of a Lawson exponential integrator, according to the following scheme:

$$\mathbf{Y}_1 = \boldsymbol{\Psi}\left(t|^{n-1}\right) \tag{95}$$

$$\mathbf{Y}_2 = \frac{\Delta t}{2} e^{\Delta t \mathscr{H}} \mathscr{N}\left(\mathbf{Y}_1,\, t|^{n-1}\right) + e^{\frac{\Delta t \mathscr{H}}{2}} \mathbf{Y}_1 \tag{96}$$

$$\mathbf{Y}_3 = \frac{\Delta t}{2} \mathscr{N}\left(\mathbf{Y}_2,\, t|^{n-2}\right) + e^{\frac{\Delta t \mathscr{H}}{2}} \mathbf{Y}_1 \tag{97}$$

$$\mathbf{Y}_4 = \Delta t e^{\frac{\Delta t \mathscr{H}}{2}} \mathscr{N}\left(\mathbf{Y}_3,\, t|^{n-\frac{1}{2}}\right) e^{\Delta t \mathscr{H}} \mathbf{Y}_1 \tag{98}$$

$$\begin{aligned} \boldsymbol{\Psi}\left(t|^{n}\right) = \frac{\Delta t}{6} \Big[& e^{\Delta t \mathscr{H}} \mathscr{N}\left(\mathbf{Y}_1,\, t|^{n-1}\right) + 2e^{\frac{\Delta t \mathscr{H}}{2}} \mathscr{N}\left(\mathbf{Y}_2,\, t|^{n-\frac{1}{2}}\right) \\ & + 2e^{\frac{\Delta t \mathscr{H}}{2}} \mathscr{N}\left(\mathbf{Y}_3,\, t|^{n-\frac{1}{2}}\right) + \mathscr{N}\left(\mathbf{Y}_4,\, t|^{n}\right) \Big] + e^{\Delta t \mathscr{H}} \mathbf{Y}_1 \end{aligned} \tag{99}$$

To ensure that the Lawson integrator is fourth-order accurate, the matrix exponential is computed through Krylov-subspace techniques, with a Krylov-subspace dimension equal to 6. The authors mention that higher-order integrators can be realized, provided that the corresponding increase in the Krylov-subspace dimension is ensured, but at the same time significantly increasing the memory requirements compared to the standard FDTD approach.

Another fourth-order approach proposed in [33] is based on the Rosenbrock-Wanner exponential integrators. Considering the general first-order initial-value problem

$$\frac{\partial y}{\partial t} = f(y), \quad y(0) = 0 \tag{100}$$

Rosenbrock-Wanner methods emerge from the linearization of the result of applying the implicit Euler discretization scheme to the aforementioned equation. Specifically, the authors in [33] apply the following fourth-order multi-step approach:

$$k_1 = \phi\left(\frac{1}{2}\Delta t A\right) f(y_n) \tag{101}$$

$$k_2 = \phi\left(\Delta t A\right) f(y_n) \tag{102}$$

$$w_3 = \frac{3}{8}\left(k_1 + k_2\right) \tag{103}$$

$$u_3 = y_n + \Delta t w_3 \tag{104}$$

$$d_3 = f(u_3) - f(y_n) - \Delta t A w_3 \tag{105}$$

$$k_3 = \phi\left(\frac{1}{2}\Delta t A\right) d_3 \tag{106}$$

$$y_{n+1} = y_n + \Delta t\left(k_2 + \frac{16}{27}k_3\right) \tag{107}$$

where

$$d_i = f(u_i) - f(y_n) - \Delta t A \sum_{j=1}^{s} \alpha_{ij} k_j \tag{108}$$

s is the number of steps, $A = f'(y_n)$ represents the Jacobian that results from the linearization, $\phi(A) = I/(I - A)$ is the characteristic of the implicit Rosenbrock-Wanner methods, and α_{ij} are free parameters. The performance of this approach relies on the exact computation of the Jacobian at each time-step, otherwise only first-order accuracy is ensured (however, only two Krylov subspaces are required for a single time-step). Despite its computational overhead, the authors mention that the suggested approach offers specific advantages that render it a quite competitive solver for nonlinear Maxwell's equations.

3.15 A Unified Nonlinear FDTD Formulation

A FDTD formulation that is capable of including various linear and nonlinear kinds of dispersion is presented in [34], which also facilitates the implementation of unidirectional sources, such as Gaussian beams. The nonlinear polarization term satisfies $\mathbf{P}(t) = \epsilon_0 S(t)\mathbf{E}(t)$, where $S(t)$ is computed as the convolution of $\chi^{(3)}(t)$ and $E^2(t)$. The quantity S can be obtained from the recursive relation

$$S|^{n+1} = A\left(E^2\right)\Big|^n + B\,S|^n + C\,S|^{n-1} \tag{109}$$

By properly changing the values of the parameters $\chi^{(3)}$, A, B, and C, Kerr as well as Raman nonlinearities can be modeled. For a general dispersive Kerr nonlinear medium, we have

$$\mathbf{D} = \epsilon_0\,(\epsilon_\infty + S)\,\mathbf{E} + \mathbf{P}_L + \epsilon_0\chi_0^{(3)}E^2\mathbf{E} \tag{110}$$

When no material dispersion is taken into account, we have

$$\mathbf{D} = \epsilon_0\left(\epsilon_r + \chi_{0,K}^{(3)}E^2\right)\mathbf{E} \tag{111}$$

which can be used for the numerical update of the electric-field intensity. For the more general dispersive nonlinear Kerr medium, it is

$$\mathbf{D} = \epsilon_0\,(\epsilon_\infty + S)\,\mathbf{E} + \mathbf{P}_L + \epsilon_0\chi_{0,K}^{(3)}E^2\mathbf{E} \tag{112}$$

On the other hand, in case of absent instantaneous Kerr nonlinearity and linear dispersion, we end up with the simple update

$$\mathbf{E}|^{n+1} = \frac{1}{\epsilon_0\left(\epsilon_\infty + S|^{n+1}\right)}\mathbf{D}|^{n+1} \tag{113}$$

Finally, the authors described a general approach for the implementation of unidirectional sources, featuring arbitrary profiles, shapes, and radiation towards different angles.

3.16 Modeling Cold-Plasma Maxwell's Equations with a Time-Split Technique

The authors of [35] discuss a methodology for cold-plasma equations, when an external EM excitation is present. Specifically, the nonlinear Drude model for modeling nonlinear dispersive media is extracted from the cold-plasma equations, and associated with Maxwell's system. The cold-plasma equations for the electron density n_e and velocity $\mathbf{u}_e$, and Maxwell's equations for fields $\mathbf{E}$, $\mathbf{B}$ are given by

$$\frac{\partial n_e}{\partial t} + \nabla \cdot (n_e \mathbf{u}_e) = 0 \tag{114}$$

$$\frac{\partial \mathbf{u}_e}{\partial t} + (\mathbf{u}_e \cdot \nabla)\mathbf{u}_e = \frac{q_e}{m_e}(\mathbf{E} + \mathbf{u}_e \times \mathbf{B}) \tag{115}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{116}$$

$$\epsilon_0 \nabla \cdot \mathbf{D} = \rho \frac{\partial \mathbf{B}}{\partial t} = -\nabla \times \mathbf{E} \tag{117}$$

$$\epsilon_0 \frac{\partial \mathbf{E}}{\partial t} = \frac{1}{\mu_0}\nabla \times \mathbf{B} - \mathbf{J} \tag{118}$$

where m_e, q_e are the electron mass and charge, respectively. The electron number density and velocity field are represented by $n_e(\mathbf{r})$ and $\mathbf{u}_e(\mathbf{r})$. The first equation is the continuity equation, and the second one is the generalized Newton's second law. The charge density ρ and current density $\mathbf{J}$ are defined as $\rho = q_e(n_e - n_0)$ and $\mathbf{J} = q_e n_e \mathbf{u}_e$, respectively, where n_0 is the (assumed constant) positive ion density . To secure charge neutrality, the electron density equals n_0 before the exciting field appears. After rewriting the initial equations in terms of ρ and $\mathbf{J}$, we obtain

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot \mathbf{J} \tag{119}$$

$$\frac{\partial \mathbf{J}}{\partial t} + \sum_k \frac{\partial}{\partial x_k} \frac{\mathbf{J} J_k}{q_e n_e} = \frac{q_e}{m_e}\left(q_e n_e \mathbf{E} + \mathbf{J} \times \mathbf{B}\right) = \frac{1}{\tau}\mathbf{J} \tag{120}$$

where τ is the phenomenological damping time constant. The aforementioned equations can be reduced to

$$\frac{\partial \mathbf{J}}{\partial t} = -\frac{1}{\tau}\mathbf{J} + \epsilon_0 \omega_p^2 \mathbf{E} + \frac{q_e}{m_e}\left(\rho \mathbf{E} + \mathbf{J} \times \mathbf{B}\right) - \sum_k \frac{\partial}{\partial x_k}\left(\frac{\mathbf{J} J_k}{\rho + \epsilon_0 m_e \omega_p^2 / q_e}\right) \tag{121}$$

where $\omega(\mathbf{r}) = \sqrt{q_e^2 n_0(\mathbf{r})/(\epsilon_0 m_e)}$ is the plasma frequency.

In order to develop a discrete model for the nonlinear Drude equation (121), a time-split semi-implicit finite-difference approach is proposed. Specifically, the initial problem is divided into three subproblems:

$$\frac{\partial \mathbf{J}}{\partial t} = -\frac{1}{\tau}\mathbf{J} + \epsilon_0 \omega_p^2 \mathbf{E} + \frac{q_e}{m_e}\epsilon_0 (\nabla \cdot \mathbf{E})\mathbf{E} \tag{122}$$

$$\frac{\partial \mathbf{J}}{\partial t} = \frac{q_e}{m_e}\mathbf{J} \times \mathbf{B} \tag{123}$$

$$\frac{\partial \mathbf{J}}{\partial t} = -\sum_k \frac{\partial}{\partial x_k}\left(\frac{\mathbf{J} J_k}{\rho + \epsilon_0 m_e \omega_p^2 / q_e}\right) \tag{124}$$

First, using (122), $\mathbf{J}$ is updated from time-step $n-1/2$ to $n+1/2$ utilizing an implicit approach with respect to $\mathbf{J}$, and explicit differencing for the remaining terms. Hence,

$$\frac{\mathbf{J}^{(1)} - \mathbf{J}|^{n-1/2}}{\Delta t} = -\frac{\mathbf{J}^{(1)} + \mathbf{J}|^{n-1/2}}{2\tau} + \epsilon_0 \omega_p^2 \mathbf{E}|^n + \frac{q_e}{m_e}\epsilon_0 (\nabla \cdot \mathbf{E}|^n)\,\mathbf{E}|^n \tag{125}$$

where $\mathbf{J}^{(1)}$ is the intermediate updated value of $\mathbf{J}$ at $n + 1/2$. The $\nabla \cdot \mathbf{E}^n$ term is evaluated in a standard fashion. As it is required at the same mesh points as the electric field, the divergence must be interpolated. Then, the finite-difference equation can be solved explicitly:

$$\mathbf{J}^{(1)} = \frac{\tau - \Delta t/2}{\tau + \Delta t/2}\mathbf{J}^{n-1/2} + \frac{\tau \Delta t}{\tau + \Delta t/2}\epsilon_0 \left[\omega_p^2 \mathbf{E}^n + \frac{q_e}{m_e}(\nabla \cdot \mathbf{E}^n)\mathbf{E}^n\right] \tag{126}$$

Second, an implicit scheme is applied to (122). Before updating this equation, the components of $\mathbf{J}$ are computed at the cell centers via interpolation. Similarly, $\mathbf{H}$ is interpolated at the cell center, and also at time instant n. The advantage of this collocation is that it eliminates the need to solve large systems. The resulting update requires that the following linear system is solved, for each cell center:

$$\frac{\mathbf{J}^{(2)} - \hat{\mathbf{J}}^n}{\Delta t} = \frac{q_e}{m_e}\mu_0 \frac{\mathbf{J}^{(2)} + \hat{\mathbf{J}}^n}{2} \times \hat{\mathbf{H}}^n \tag{127}$$

or, more compactly, $\mathrm{A}\mathbf{J}^{(2)} = \mathrm{A}' \hat{\mathbf{J}}\Big|^n$, with

$$\mathrm{A} = \begin{pmatrix} 1 & -aH_z & aH_y \\ aH_z & 1 & -aH_x \\ -aH_y & aH_x & 1 \end{pmatrix} \tag{128}$$

where the hat notation indicates interpolation. Moreover, A' denotes the transpose of A, $\hat{\mathbf{H}} = [H_x \;\; H_y \;\; H_z]'$, and $a = 0.5\Delta t \mu_0 q_e/m_e$. The explicit solution then takes the form

$$\mathbf{J}^{(2)} =$$

$$= \frac{1}{|\mathrm{A}|}\left.\begin{pmatrix} 1 + a^2(H_x^2 - H_y^2 - H_z^2) & 2a(aH_xH_y + H_z) & 2a(aH_zH_x - H_y) \\ 2a(aH_xH_y - H_z) & 1 + a^2(H_y^2 - H_z^2 - H_x^2) & 2a(aH_yH_z - H_x) \\ 2a(aH_zH_x + H_y) & 2a(aH_yH_z - H_x) & 1 + a^2(H_z^2 - H_x^2 - H_y^2) \end{pmatrix}\right|^n \tag{129}$$

where $|\mathrm{A}| = 1 + a^2\left(H_x^2 + H_y^2 + H_z^2\right)$. Finally, (124) is updated, after it is written in conservation form:

$$\frac{\partial \mathbf{J}}{\partial t} + F_x(\mathbf{J}) + G_y(\mathbf{J}) + K_z(\mathbf{J}) = 0 \tag{130}$$

where $F(\mathbf{J}) = \mathbf{J}J_x/\rho'$, $G(\mathbf{J}) = \mathbf{J}J_y/\rho'$, $K(\mathbf{J}) = \mathbf{J}J_z/\rho'$ and $\rho' = \rho + \epsilon_0 m_e \omega_p^2/q_e$. In this case, a two-step Lax-Wendroff approach can be implemented. The time-step is chosen to satisfy the CFL conditions as $\Delta t < \Delta x/(2\max\{\upsilon_{\max}, c\})$, where c is the speed of light in vacuum and $\upsilon_{\max} = \max\{|\mathbf{J}/\rho'|\}$ is the maximum wave velocity in the plasma. After the last update has been performed, all J components are interpolated back at the cell face centers, and can be then introduced in the FDTD method for computing the E field.

3.17 3D Modeling of Nonlinear Ferroelectric Materials

The work of [36] investigates the behavior of nonlinear ferroelectric materials (i.e. materials with nonlinear polarization response) and specifically their effects, when they are present inside a rectangular waveguide. Regarding ferroelectric materials, the linear relation $\mathbf{P} = \epsilon_0 \chi \mathbf{E}$ is valid only for small values of the applied electric-field intensity. When the electric field becomes stronger, all the electric domains of the material are aligned, and polarization reaches saturation, without the requirement of an external bias. The nonlinear behavior of the ferroelectric material can be described reliably by a modified hyperbolic tangent function,

$$\mathbf{P} = P_{\text{sat}} \tanh\left(E_{\text{scale}}\mathbf{E}\right) \tag{131}$$

where P_{sat} is the polarization saturation limit, and $E_{\text{scale}} = \frac{\epsilon_0(\epsilon_r - 1)}{P_{\text{sat}}}$ is a scaling factor.

In order to construct a FDTD algorithm without augmented memory requirements, the time derivative of $\mathbf{D}$ is computed according to

$$\frac{\partial \mathbf{D}}{\partial t} = \frac{\partial \mathbf{D}}{\partial \mathbf{E}} \frac{\partial \mathbf{E}}{\partial t} = P_{\text{sat}} E_{\text{scale}} \left(1 - \tanh^2\left(E_{\text{scale}}\mathbf{E}\right)\right) + \epsilon_0 \tag{132}$$

which leads to the definition of the effective permittivity:

$$\epsilon_{\text{eff}} = P_{\text{sat}} E_{\text{scale}} \left(1 - \tanh^2\left(E_{\text{scale}}\left|\mathbf{E}^{n-1}\right|\right)\right) + \epsilon_0 \tag{133}$$

Then, the electric- and magnetic-field intensities are update in a standard manner, following

$$\mathbf{H}^{n+\frac{1}{2}} = \mathbf{H}^{n-\frac{1}{2}} - \frac{\Delta t}{\mu_0} \nabla \times \mathbf{E}^n \tag{134}$$

$$\mathbf{E}^{n+1} = \mathbf{E}^n + \Delta t (\epsilon_{\text{eff}})^{-1} \nabla \times \mathbf{H}^{n+\frac{1}{2}} \tag{135}$$

Compared to the linear case, simulations show that the presence of a ferroelectric material in a waveguide structure leads to an increase of the peak power density and a spatial compression of the pulsewidth. This type of behavior can be useful in various contemporary applications that require high-power electromagnetic pulse generation, sharp/compressed waveforms, etc.

3.18 Analysis of Second-Harmonic Generation in Periodic Structures

An extension of the split-field (SF) FDTD method to 2D periodic configurations without any assumptions regarding the material symmetries is developed in [37]. For a non-magnetic and non-conducting medium, Maxwell's equations are given by

$$\nabla \times \mathbf{E} = -j\omega\mu_0\mathbf{H}, \quad \nabla \times \mathbf{H} = j\omega\epsilon_0\epsilon_r\mathbf{E} + j\omega\mathbf{F}^{NL} \tag{136}$$

where $\mathbf{F}^{NL}$ is the nonlinear polarization. The SF FDTD method transforms the electric and magnetic fields, considering that the new quantities contain the oblique field propagation in an implicit fashion. This produces the new variables

$$\mathbf{P} = \mathbf{E}e^{j(k_x x + k_y y)}, \quad \mathbf{Q} = c\mu_0\mathbf{H}e^{j(k_x x + k_y y)} \tag{137}$$

where $\mathbf{P}$ and $\mathbf{Q}$ are considered in the phasor domain. A similar transformation can be also implemented to the nonlinear polarization term, using $\mathbf{G}^{NL} = c\mu_0\mathbf{F}^{NL}e^{j(k_x x + k_y y)}$. Substituting the new components into Maxwell's equations, the SF-FDTD algorithm is formulated as

$$\frac{j\omega}{c}\mathbf{P} = \kappa\nabla \times \mathbf{Q} + \frac{j\omega}{c}\kappa\mathbf{q}\mathbf{Q} - j\omega\mathbf{G}^{NL} \tag{138}$$

$$\frac{j\omega}{c}\mathbf{Q} = -\nabla \times \mathbf{P} - \frac{j\omega}{c}\kappa\mathbf{q}\mathbf{P} \tag{139}$$

where $\kappa = \epsilon_r^{-1}$ and

$$\mathbf{q} = \frac{\omega}{c}\begin{bmatrix} 0 & 0 & -k_y \\ 0 & 0 & k_x \\ k_y & -k_x & 0 \end{bmatrix} \tag{140}$$

The appearance of time derivatives on both sides hinders the direct approximation via finite differences. To solve this problem, new variables are defined, separating both $\mathbf{P}$, $\mathbf{Q}$ into two parts:

$$\mathbf{P} = \mathbf{P}_a + \kappa\mathbf{q}\mathbf{Q} - c\mathbf{G}^{NL}, \quad \mathbf{Q} = \mathbf{Q}_a - \mathbf{q}\mathbf{P} \tag{141}$$

After proper manipulations, discretizing the equations with respect to time yields

$$\frac{1}{c\Delta t}\left(\mathbf{P}_a|^{n+1}-\mathbf{P}_a|^{n}\right)=\kappa\nabla\times\mathbf{Q}|^{n+1/2} \tag{142}$$

$$\frac{1}{c\Delta t}\left(\mathbf{Q}_a|^{n+1}-\mathbf{Q}_a|^{n}\right)=-\nabla\times\mathbf{P}|^{n+1/2} \tag{143}$$

The stability of the SF-FDTD algorithm is affected by various factors, including the CFL condition, the averaging process, and large incidence angles. The lower bound can ensure stability in most case studies. Consequently, the CFL number is selected low enough, in order to ensure both stability and convergence. Furthermore, this also leads to lower time and spatial resolutions and, hence, larger grids, simulation times, and computational resources. It is reminded that FDTD models display exponentially growing computational costs. Consequently, increasing the grid size has a severe consequence on the necessary simulation times. The proposed SF-FDTD leapfrog algorithm updates the "a" fields from the **P** and **Q** quantities. After that, the current **P** field is calculated from

$$\mathbf{P}=\frac{\mathbf{P}_a+\kappa\mathbf{q}\mathbf{Q}_a-c\mathbf{G}^{NL}}{\mathbf{I}+\kappa\mathbf{q}^2} \tag{144}$$

where **I** is the identity matrix. After updating **P**, it is straightforward to obtain **Q**. Moving to a particular material configuration, let us examine introducing the polarization terms in (144) for the case of a tensorial second-order nonlinear susceptibility. To this objective, polarization $\mathbf{F}^{NL}$ that represents the nonlinear response in non-centrosymmetric materials associates the second-order nonlinear susceptibility with the field within the considered configuration. Actually, the appearance of a fundamental or pump field at ω_f produces an exchange of energy with the (second-harmonic field) at $\omega_s = 2\omega_f$. The nonlinear polarization is described by a third-rank tensor **d** that, in the case of second harmonic generation (in a periodic nanostructure), can be represented in matrix form as

$$\mathbf{d}=\begin{bmatrix} d_{11} & d_{12} & d_{13} & d_{14} & d_{15} & d_{16}\\ d_{21} & d_{22} & d_{23} & d_{24} & d_{25} & d_{26}\\ d_{31} & d_{32} & d_{33} & d_{34} & d_{35} & d_{36}\end{bmatrix} \tag{145}$$

whose elements are defined according to the involved nonlinear medium. Taking into account the transformation, we find that

$$\begin{bmatrix} G_x^{NL,\omega_f}\\ G_y^{NL,\omega_f}\\ G_z^{NL,\omega_f}\end{bmatrix}=\frac{2}{c}\mathbf{d}\begin{bmatrix} P_x^{\omega_f}E_x^{\omega_s}\\ P_y^{\omega_f}E_y^{\omega_s}\\ P_z^{\omega_f}E_z^{\omega_s}\\ P_z^{\omega_f}E_y^{\omega_s}+P_y^{\omega_f}E_z^{\omega_s}\\ P_z^{\omega_f}E_x^{\omega_s}+P_x^{\omega_f}E_z^{\omega_s}\\ P_x^{\omega_f}E_y^{\omega_s}+P_y^{\omega_f}E_x^{\omega_s}\end{bmatrix} \tag{146}$$

$$\begin{bmatrix} G_x^{NL,\omega_s} \\ G_y^{NL,\omega_s} \\ G_z^{NL,\omega_s} \end{bmatrix} = \frac{1}{c}\mathbf{d} \begin{bmatrix} P_x^{\omega_f} P_x^{\omega_f} \\ P_y^{\omega_f} P_y^{\omega_f} \\ P_z^{\omega_f} P_z^{\omega_f} \\ 2P_z^{\omega_f} P_y^{\omega_f} \\ 2P_z^{\omega_f} P_x^{\omega_f} \\ 2P_x^{\omega_f} P_y^{\omega_f} \end{bmatrix} \tag{147}$$

Once the “a” fields are known, the total fields can be computed. It turns out that they can be expressed using only the “a” fields and the interior electric components. For instance,

$$P_x^{\omega_f} = \frac{P_{xa}^{\omega_f} - \kappa\left(k_y Q_{za}^{\omega_f} - k_x k_y P_y^{\omega_f} - c\bar{G}_x^{NL,\omega_f}\right)}{1+\kappa\left[k_y^2 + 2\left(d_{11}E_x^{\omega_s} + d_{15}E_z^{\omega_s} + d_{16}E_y^{\omega_s}\right)\right]} \tag{148}$$

$$P_x^{\omega_s} = \frac{P_{xa}^{\omega_s} - \kappa\left(k_y Q_{za}^{\omega_s} - k_x k_y P_y^{\omega_s} - cG_x^{NL,\omega_s}\right)}{1+\kappa k_y^2} \tag{149}$$

where

$$\begin{aligned} \bar{G}_x^{NL,\omega_f} &= \frac{2}{c}\Big[d_{12}P_y^{\omega_f}E_y^{\omega_s} + d_{13}P_z^{\omega_f}E_z^{\omega_s} \\ &\quad + d_{14}\left(P_z^{\omega_f}E_y^{\omega_s} + P_y^{\omega_f}E_z^{\omega_s}\right) + d_{15}P_z^{\omega_f}E_x^{\omega_s} + d_{16}P_y^{\omega_f}E_x^{\omega_s}\Big] \end{aligned} \tag{150}$$

Similar formulae are obtained for the remaining components. In this manner, a nonlinear equation system of the form $\mathbf{P} = \mathbf{U}(\mathbf{P})$ is composed. For its solution, a fixed-point iterative procedure can be selected. The key point of this approach is to solve the iterative process with the form $\mathbf{P}^{(p+1)} = \mathbf{U}(\mathbf{P}^{(p)})$, with $p = 1, 2, \ldots$ the number of iterations. It is noted that the fixed-point process needs to be performed at every time-step of the FDTD updating procedure. The approach requires an initial guess of $\mathbf{P}$, which corresponds to the fields considering linear media. Then, subsequent iterations are carried out, so that the precision of the outcomes improves with every iteration. Finally, in order to ensure the convergence of the iterative procedure, the amplitude of $\mathbf{E}$ must be limited by an upper bound, which is dictated by the magnitude of the second-order susceptibility.

3.19 Time-Filtered Integration of Maxwell’s Equations

The case of a dielectric medium with Kerr nonlinearity is also examined in [38], in the context of verifying the performance of a novel FDTD approach that employs an unstaggered temporal grid. Specifically, considering a 1D setup and that the time derivatives of the dielectric displacement and the electric-field intensity are connected via

$$\frac{\partial E}{\partial t} = c\,(E)\,\frac{\partial D}{\partial t} \tag{151}$$

where $c\,(E) = \epsilon + 3\chi^{(3)}E^2$, the corresponding electromagnetic phenomena can be described by the following discrete model:

$$\frac{E|_i^{n+1} - \tilde{E}\Big|_i^{n-1}}{2\Delta t} = \frac{1}{\epsilon + 3\chi^{(3)}\big(E|_i^n\big)^2}\,\frac{H|_{i+\frac{1}{2}}^{av} - H|_{i-\frac{1}{2}}^{av}}{\Delta x} + \nu\,\frac{\partial^4 E}{\partial x^4}\bigg|_i^n \tag{152}$$

$$\frac{H|_p^{n+1} - \tilde{H}\Big|_p^{n-1}}{2\Delta t} = \frac{1}{\mu}\,\frac{E|_{p+\frac{1}{2}}^{av} - E|_{p-\frac{1}{2}}^{av}}{\Delta x} + \nu\,\frac{\partial^4 H}{\partial x^4}\bigg|_i^n \tag{153}$$

In the above equations, it is $p = i$ in case of spatial collocation, or $p = i + 1/2$ in case of staggered spatial grids. The $\tilde{E}\Big|_i^{n-1}$, $\tilde{H}\Big|_p^{n-1}$ components denote values obtained after a time filtering process has been applied, according to the implicit scheme

$$\tilde{F}\Big|_i^{n-1} = F|_i^{n-1} + \gamma\left(-\,\tilde{F}\Big|_i^{n-3} + 4\,\tilde{F}\Big|_i^{n-2} - 6\,\tilde{F}\Big|_i^{n-1} + 4\,F|_i^n - F|_i^{n+1}\right), \quad F = E, H \tag{154}$$

Note that these computations can be conducted in an explicit fashion, after proper reformulation. The aim of the aforementioned filtering approach is to weaken high-frequency modes that emerge due to lack of staggering in time. The averaged values appearing in the spatial derivative approximations are computed via

$$F|_i^{av} = \frac{1}{24}\left(-\,F|_i^n + 26\,F|_i^n - F|_i^n\right) \tag{155}$$

$$F|_{i+\frac{1}{2}}^{av} = \frac{7}{12}\left(F|_{i+1}^n + F|_i^n\right) - \frac{1}{12}\left(F|_{i+2}^n + F|_{i-1}^n\right) \tag{156}$$

for collocated and staggered grids, respectively. Approximations (155) and (156) actually result in fourth-order finite-difference formulae. Finally, the fourth-order derivatives appearing in (152) and (153) act as smoothers that combat inadequately resolved high-frequency oscillations, and are computed via

$$\frac{\partial^4 F}{\partial x^4}\bigg|_i^n = \frac{1}{\Delta x^4}\left(-\,F|_{i-2}^n + 4\,F|_{i-1}^n - 6\,F|_i^n + 4\,F|_{i+1}^n - F|_{i+2}^n\right) \tag{157}$$

To validate the proposed computational scheme, the authors of [38] proceeded to the analysis of parametric Four-Wave Mixing, which is a phenomenon directly related to the third-order nonlinearity.

4 Conclusions

In the present work, we have reviewed a number of important contributions regarding the development of finite-difference models for nonlinear EM problems that may emerge in numerous realistic applications. As nonlinearities do not necessarily comply with a specific unique form, pertinent research efforts have produced a number of different computational approaches that adapt the conventional FDTD algorithm to the peculiar requirements of each problem. The majority of the developed techniques preserve most of the attractive properties of the classic FDTD scheme, thus the study of a wide variety of nonlinear EM problems can be conducted reliably, without requiring special conditions to hold. Evidently, the present study is not completely exhaustive. However, it is the authors' belief that many of the available key contributions in this scientific area have been included, and a quite complete description of the latter has been formulated. In any case, the research on computational nonlinear EM models is ongoing and continuously developing, and novel contributions towards more efficient techniques or numerical approaches for less explored phenomena and applications constantly appear, thus broadening further this already vast scientific subject.

References

1. K. Yee, Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. IEEE Trans. Antennas Propag. **14**(3), 302–307 (1966)
2. A. Taflove, S.C. Hagness, *Computational Electrodynamics: The Finite-Difference Time-Domain Method* (Artech House, Norwood, 2005)
3. J.-F. Lee, R. Lee, A. Cangellaris, Time-domain finite-element methods. IEEE Trans. Antennas Propag. **45**(3), 430–442 (1997)
4. J. Chen, Q.H. Liu, Discontinuous Galerkin time-domain methods for multiscale electromagnetic simulations: a review. Proc. IEEE **101**(2), 242–254 (2013)
5. S.M. Rao, D.R. Wilton, Transient scattering by conducting surfaces of arbitrary shape. IEEE Trans. Antennas Propag. **39**(1), 56–61 (1991)
6. N.V. Kantartzis, T.D. Tsiboukis, *Higher Order FDTD Schemes for Waveguide and Antenna Structures* (Morgan & Claypool Publishers, San Rafael, 2006)
7. M. Okoniewski, M. Mrozowski, M.A. Stuchly, Simple treatment of multi-term dispersion in FDTD. IEEE Microw. Guid. Wave Lett. **7**(5), 121–123 (1997)
8. J.-P. Berenger, A perfectly matched layer for the absorption of electromagnetic waves. J. Comput. Phys. **114**(2), 185–200 (1994)
9. A. Monorchio, R. Mittra, Time-domain (FE/FDTD) technique for solving complex electromagnetic problems. IEEE Microw. Guid. Wave Lett. **8**(2), 93–95 (1998)

10. T. Noda, S. Yokoyama, Thin wire representation in finite difference time domain surge simulation. IEEE Trans. Power Delivery **17**(3), 840–847 (2002)
11. J.G. Maloney, G.S. Smith, The use of surface impedance concepts in the finite-difference time-domain method. IEEE Trans. Antennas Propag. **40**(1), 38–48 (1992)
12. D. Anderson, Variational approach to nonlinear pulse propagation in optical fibers. Phys. Rev. A **27**, 3135–3145 (1983)
13. Y. Chung, N. Dagli, An assessment of finite difference beam propagation method. IEEE J. Quantum Electron. **26**(8), 1335–1339 (1990)
14. F.L. Teixeira, Time-domain finite-difference and finite-element methods for Maxwell equations in complex media. IEEE Trans. Antennas Propag. **56**(8), 2150–2166 (2008)
15. R.M. Joseph, A. Taflove, FDTD Maxwell's equations models for nonlinear electrodynamics and optics. IEEE Trans. Antennas Propag. **45**(3), 364–374 (1997)
16. I.S. Maksymov, A.A. Sukhorukov, A.V. Lavrinenko, Y.S. Kivshar, Comparative study of FDTD-adopted numerical algorithms for Kerr nonlinearities. IEEE Antennas Wirel. Propag. Lett. **10**, 143–146 (2011)
17. P.M. Goorjian, A. Taflove, Direct time integration of Maxwell's equations in nonlinear dispersive media for propagation and scattering of femtosecond electromagnetic solitons. Opt. Lett. **17**(3), 180–182 (1992)
18. P.M. Goorjian, A. Taflove, R.M. Joseph, S.C. Hagness, Computational modeling of femtosecond optical solitons from Maxwell's equations. IEEE J. Quantum Electron. **28**(10), 2416–2422 (1992)
19. K.J. Blow, D. Wood, Theoretical description of transient stimulated Raman scattering in optical fibers. IEEE J. Quantum Electron. **25**(12), 2665–2673 (1989)
20. R.W. Ziolkowski, J.B. Judkins, Applications of the nonlinear finite difference time domain (NL-FDTD) method to pulse propagation in nonlinear media: self-focusing and linear-nonlinear interfaces. Radio Sci. **28**(5), 901–911 (1993)
21. R. Luebbers, K. Kumagai, S. Adachi, T. Uno, FDTD calculation of transient pulse propagation through a nonlinear magnetic sheet. IEEE Trans. Electromagn. Compat. **35**(1), 90–94 (1993)
22. B. Toland, B. Houshmand, T. Itoh, Modeling of nonlinear active regions with the FDTD method. IEEE Microw. Guid. Wave Lett. **3**(9), 333–335 (1993)
23. R. Holland, Finite-difference time-domain (FDTD) analysis of magnetic diffusion. IEEE Trans. Electromagn. Compat. **36**(1), 32–39 (1994)
24. P. Tran, Photonic-band-structure calculation of material possessing Kerr nonlinearity. Phys. Rev. B **52**, 10673–10676 (1995)
25. D.M. Sullivan, Nonlinear FDTD formulations using Z transforms. IEEE Trans. Microwave Theory Tech. **43**(3), 676–682 (1995)
26. A.S. Nagra, R.A. York, FDTD analysis of wave propagation in nonlinear absorbing and gain media. IEEE Trans. Antennas Propag. **46**(3), 334–340 (1998)
27. V. Van, S.K. Chaudhuri, A hybrid implicit-explicit FDTD scheme for nonlinear optical waveguide modeling. IEEE Trans. Microwave Theory Tech. **47**(5), 540–545 (1999)
28. D. Sullivan, J. Liu, M. Kuzyk, Three-dimensional optical pulse simulation using the FDTD method. IEEE Trans. Microwave Theory Tech. **48**(7), 1127–1133 (2000)
29. F. Raineri, Y. Dumeige, A. Levenson, X. Letartre, Nonlinear decoupled FDTD code: phase-matching in 2D defective photonic crystal. Electron. Lett. **38**(25), 1704–1706 (2002)
30. M. Fujii, M. Tahara, I. Sakagami, W. Freude, P. Russer, High-order FDTD and auxiliary differential equation formulation of optical pulse propagation in 2-D kerr and raman nonlinear dispersive media. IEEE J. Quantum Electron. **40**(2), 175–182 (2004)
31. M. Fujii, P. Russer, A nonlinear and dispersive APML ABC for the FD-TD methods. IEEE Microw. Wirel. Compon. Lett. **12**(11), 444–446 (2002)
32. J.H. Greene, A. Taflove, General vector auxiliary differential equation finite-difference time-domain method for nonlinear optics. Opt. Express **14**(18), 8305–8310 (2006)
33. M. Pototschnig, J. Niegemann, L. Tkeshelashvili, K. Busch, Time-domain simulations of the nonlinear Maxwell equations using operator-exponential methods. IEEE Trans. Antennas Propag. **57**(2), 475–483 (2009)

34. A. Naqavi, M. Miri, K. Mehrany, S. Khorasani, Extension of unified formulation for the FDTD simulation of nonlinear dispersive media. IEEE Photon. Technol. Lett. **22**(16), 1214–1216 (2010)
35. J. Liu, M. Brio, Y. Zeng, A.R. Zakharian, W. Hoyer, S.W. Koch, J.V. Moloney, Generalization of the FDTD algorithm for simulations of hydrodynamic nonlinear Drude model. J. Comput. Phys. **229**(17), 5921–5932 (2010)
36. B.T. Caudle, M.E. Baginski, H. Kirkici, M.C. Hamilton, Three-dimensional FDTD simulation of nonlinear ferroelectric materials in rectangular waveguide. IEEE Trans. Plasma Sci. **41**(2), 365–370 (2013)
37. J. Francés, J. Tervo, S. Gallego, S. Bleda, C. Neipp, A. Márquez, Split-field finite-difference time-domain method for second-harmonic generation in two-dimensionally periodic structures. J. Opt. Soc. Am. B **32**(4), 664–669 (2015)
38. A. Mahalov, M. Moustaoui, Time-filtered leapfrog integration of Maxwell equations using unstaggered temporal grids. J. Comput. Phys. **325**, 98–115 (2016)