


Springer Optimization and Its Applications 167

Themistocles M. Rassias
Panos M. Pardalos *Editors*

Nonlinear Analysis and Global Optimization

 Springer

Springer Optimization and Its Applications

Volume 167

Series Editors

Panos M. Pardalos , *University of Florida*

My T. Thai , *University of Florida*

Honorary Editor

Ding-Zhu Du, *University of Texas at Dallas*

Advisory Editors

Roman V. Belavkin, *Middlesex University*

John R. Birge, *University of Chicago*

Sergiy Butenko, *Texas A&M University*

Vipin Kumar, *University of Minnesota*

Anna Nagurney, *University of Massachusetts Amherst*

Jun Pei, *Hefei University of Technology*

Oleg Prokopyev, *University of Pittsburgh*

Steffen Rebennack, *Karlsruhe Institute of Technology*

Mauricio Resende, *Amazon*

Tamás Terlaky, *Lehigh University*

Van Vu, *Yale University*

Michael N. Vrahatis, *University of Patras*

Guoliang Xue, *Arizona State University*

Yinyu Ye, *Stanford University*

Aims and Scope

Optimization has continued to expand in all directions at an astonishing rate. New algorithmic and theoretical techniques are continually developing and the diffusion into other disciplines is proceeding at a rapid pace, with a spot light on machine learning, artificial intelligence, and quantum computing. 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 areas not limited to applied mathematics, engineering, medicine, economics, computer science, operations research, and other sciences.

The series **Springer Optimization and Its Applications (SOIA)** aims to publish state-of-the-art expository works (monographs, contributed volumes, textbooks, handbooks) that focus on theory, methods, and applications of optimization. Topics covered include, but are not limited to, nonlinear optimization, combinatorial optimization, continuous optimization, stochastic optimization, Bayesian optimization, optimal control, discrete optimization, multi-objective optimization, and more. New to the series portfolio include Works at the intersection of optimization and machine learning, artificial intelligence, and quantum computing.

Volumes from this series are indexed by Web of Science, zbMATH, Mathematical Reviews, and SCOPUS.

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

Themistocles M. Rassias • Panos M. Pardalos
Editors

Nonlinear Analysis and Global Optimization

 Springer

Editors

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

Panos M. Pardalos
Department of Industrial
and Systems Engineering
University of Florida
Gainesville, FL, USA

ISSN 1931-6828

ISSN 1931-6836 (electronic)

Springer Optimization and Its Applications

ISBN 978-3-030-61731-8

ISBN 978-3-030-61732-5 (eBook)

<https://doi.org/10.1007/978-3-030-61732-5>

Mathematics Subject Classification: 26-XX, 28-XX, 30-XX, 41-XX, 46-XX, 47-XX, 49-XX, 52-XX, 58-XX, 90-XX, 91-XX, 93-XX

© Springer Nature Switzerland AG 2021

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

Preface

Nonlinear Analysis and Global Optimization is devoted to the study of a broad area of research with strong interdisciplinary flavor in which optimization plays an important role, as well as to the study of topics that are applied toward the investigation of optimization problems. More specifically, the present book treats topics on set-valued analysis, mixed concave–convex sub-superlinear Schrödinger equation, Schrödinger equations in nonlinear optics, exponentially convex functions, optimal lot size under the occurrence of imperfect quality items, generalized equilibrium problems, artificial topologies on a relativistic spacetime, equilibrium points in the restricted three-body problem, optimization models for networks of organ transplants, network curvature measures, error analysis through energy minimization and stability problems, Ekeland variational principles in 2-local Branciari metric spaces, frictional dynamic problems, norm estimates for composite operators, operator factorization and solution of second-order nonlinear difference equations, degenerate Kirchhoff-type inclusion problems, and more.

We would like to express our deep thanks to all the authors of contributed book chapters for participating in this collective effort. We would also like to express our thanks to the staff of Springer for their help throughout the preparation of this book.

Athens, Greece

Themistocles M. Rassias

Gainesville, FL, USA

Panos M. Pardalos

Contents

Friction Models in the Framework of Set-Valued and Convex Analysis	1
Samir Adly, Daniel Goeleven, and Rachid Ouja	
A Survey on Markov’s Theorem on Zeros of Orthogonal Polynomials	23
Kenier Castillo, Marisa de Souza Costa, and Fernando Rodrigo Rafaeli	
A Review of Two Network Curvature Measures	51
Tanima Chatterjee, Bhaskar DasGupta, and Réka Albert	
A Frictional Dynamic Thermal Contact Problem with Normal Compliance and Damage	71
Oanh Chau, Adrien Petrov, Arnaud Heibig, and Manuel Monteiro Marques	
Mixed Concave–Convex Sub-Superlinear Schrödinger Equation: Survey and Development of Some New Cases	109
Riadh Chteoui, Anouar Ben Mabrouk, and Carlo Cattani	
An Optimization Model for a Network of Organ Transplants with Uncertain Availability	163
Gabriella Colajanni and Patrizia Daniele	
Algebraic Based Techniques as Decision Making Tools	183
M. Couceiro, G. Meletiou, and K. Skouri	
Norm Estimates for the Composite Operators	193
Shusen Ding, Guannan Shi, and Yuming Xing	
A Variational Inequality Based Stochastic Approximation for Inverse Problems in Stochastic Partial Differential Equations	207
Rachel Hawks, Baasansuren Jadamba, Akhtar A. Khan, Miguel Sama, and Yidan Yang	

An Iterative Method for a Common Solution of Split Generalized Equilibrium Problems and Fixed Points of a Finite Family of Nonexpansive Mapping	227
Ihssane Hay, Abdellah Bnouhachem, and Themistocles M. Rassias	
Periodic Solutions Around the Out-of-Plane Equilibrium Points in the Restricted Three-Body Problem with Radiation and Angular Velocity Variation	251
Vassilis S. Kalantonis, Aguda Ekele Vincent, Jessica Mrumun Gyegwe, and Efstathios A. Perdios	
Optimal Lot Size with Partial Backlogging Under the Occurrence of Imperfect Quality Items	277
G. Karakatsoulis and K. Skouri	
Error Analysis Through Energy Minimization and Stability Properties of Exponential Integrators	295
Odysseas Kosmas and Dimitrios Vlachos	
A Degenerate Kirchhoff-Type Inclusion Problem with Nonlocal Operator	309
Dumitru Motreanu	
Competition for Medical Supplies Under Stochastic Demand in the Covid-19 Pandemic: A Generalized Nash Equilibrium Framework	331
Anna Nagurney, Mojtaba Salarpour, June Dong, and Pritha Dutta	
Relative Strongly Exponentially Convex Functions	357
Muhammad Aslam Noor, Khalida Inayat Noor, and Themistocles M. Rassias	
Properties of Exponentially m-Convex Functions	373
Muhammad Aslam Noor and Khalida Inayat Noor	
Natural vs. Artificial Topologies on a Relativistic Spacetime	389
Kyriakos Papadopoulos	
On the Approximation of Monotone Variational Inequalities in L^p Spaces with Probability Measure	403
Mauro Passacantando and Fabio Raciti	
Operator Factorization and Solution of Second-Order Nonlinear Difference Equations with Variable Coefficients and Multipoint Constraints	427
E. Providas	
An Invitation to the Study of a Uniqueness Problem	445
Biagio Ricceri	

Schrödinger Equations in Nonlinear Optics	449
Martin Schechter	
Ekeland Variational Principles in 2-Local Branciari Metric Spaces	461
Mihai Turinici	

Friction Models in the Framework of Set-Valued and Convex Analysis



Samir Adly, Daniel Goeleven, and Rachid Oujja

Abstract It is well known that modeling friction forces is a complex problem and constitutes an important topic in both mechanical engineering and applied mathematics. In this paper, we show how the approach of Moreau and Panagiotopoulos can be used to develop a suitable methodology for the formulation and the mathematical analysis of various friction models in nonsmooth mechanics. We study 11 widespread engineering friction models in the context of modern set-valued and convex analysis. The stability analysis (in the sense of Lyapunov) of a two-degree-of-freedom mechanical system with dry friction is also discussed.

1 Introduction

The first systematic study of friction is due to the famous Italian scientist and artist Leonardo da Vinci (1452–1519) (see e.g. [14]). He discovered that the friction force is proportional to load, opposes the motion, and is independent of contact area. These fundamental results have been rediscovered by Guillaume Amontons (1663–1705) and developed by Charles-Augustin de Coulomb (1736–1806) [4, 13]. The Coulomb friction force F is a function of the load and direction of the sliding velocity v . Arthur Morin (1795–1880) found that the friction at zero sliding speed (static friction) is larger than the Coulomb friction (dynamic friction) [24]. Osborne Reynolds (1842–1912) introduced the concept of viscous friction in relation to lubricated contact [33]. Richard Stribeck (1861–1950) observed that friction force decreases with the increase of the sliding speed from the static friction to the Coulomb friction [35]. All these fundamental discoveries have since been the

S. Adly

Laboratoire XLIM, Université de Limoges, Limoges, France

e-mail: samir.adly@unilim.fr

D. Goeleven (✉) · R. Oujja

Laboratoire PIMENT, Université de La Réunion, Saint-Denis, France

e-mail: goeleven@univ-reunion.fr; oujja@univ-reunion.fr

© Springer Nature Switzerland AG 2021

T. M. Rassias, P. M. Pardalos (eds.), *Nonlinear Analysis and Global Optimization*,

Springer Optimization and Its Applications 167,

https://doi.org/10.1007/978-3-030-61732-5_1

subject of much research by the engineering community (see e.g. [6, 7, 17, 20, 25, 26, 30]). Most models of friction are nonsmooth in the sense that the function involved in the model $v \mapsto F(v)$ is not continuous at $v = 0$. The pioneering works of Jean-Jacques Moreau (1923–2014) and Panagiotis D. Panagiotopoulos (1950–1998) catalyzed the development of a mathematical framework applicable to the study of nonsmooth mechanical problems in using advanced results of modern convex analysis and set-valued analysis (see e.g. [2, 16, 19, 21–23, 28, 29]). The approach of Moreau and Panagiotopoulos can, in particular, be used to write a precise and rigorous mathematical model describing the friction force and the stick–slip phenomenon. This approach using set-valued functions left aside the complicated transition processes between “stick” and “slip” but led to rigorous mathematical models like differential inclusions and variational inequalities. However, most engineers prefer to leave aside some mathematical difficulties and use another approach that consists of specifying a value of F at 0 that is mechanically consistent. There are thus two approaches to deal with problems that involve friction: the one that makes mathematicians happy and the one that makes engineers happy. In this expository paper, we summarize the two approaches through different models.

2 The Approach of Moreau and Panagiotopoulos

For many discrete mechanical systems with a finite number of degrees of freedom, the understanding of scalar mechanical laws $\mathcal{F} : \mathbb{R} \rightrightarrows \mathbb{R}; v \mapsto \mathcal{F}(v)$ is very helpful. Let us here consider a set-valued map $\mathcal{F} : \mathbb{R} \rightrightarrows \mathbb{R}, v \mapsto \mathcal{F}(v)$ whose graph may present some finite vertical branches.

Our aim in this section is to propose a mathematical relation that describes a general possibly set-valued graph. The approach for doing that exists in the literature and has essentially been developed by researchers from the nonsmooth mechanical community. It has been introduced by J.J. Moreau [21] for the treatment of monotone set-valued graphs and then extended by P.D. Panagiotopoulos [28] for the treatment of general set-valued graphs including both monotone and non-monotone graphs. Note that filling in the graph of a discontinuous function is a methodology that can also be traced back to J. Rauch [31] for PDEs. This approach is now used by most engineers to formulate concrete models for highly nonlinear phenomena in mechanics like adhesion, friction, unilateral contact, and delamination (see e.g. [29]). It is also used to study nonsmooth switches in electrical systems (see e.g. [15]).

We may first write

$$F \in \mathcal{F}(v), \quad (v \in \mathbb{R}) \tag{1}$$

if $\mathcal{F} : \mathbb{R} \rightrightarrows \mathbb{R}$ is some set-valued function. Some abstract model is depicted in Figure 1.

Fig. 1 Graph of \mathcal{F}

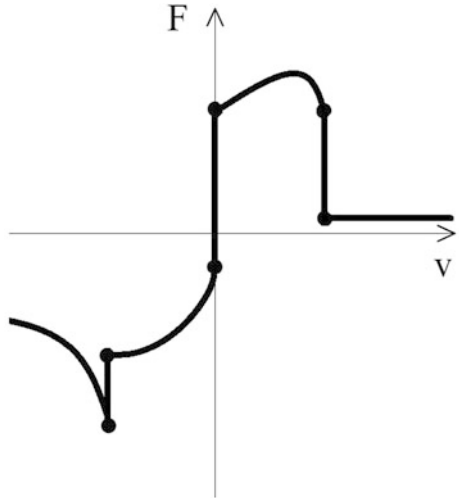
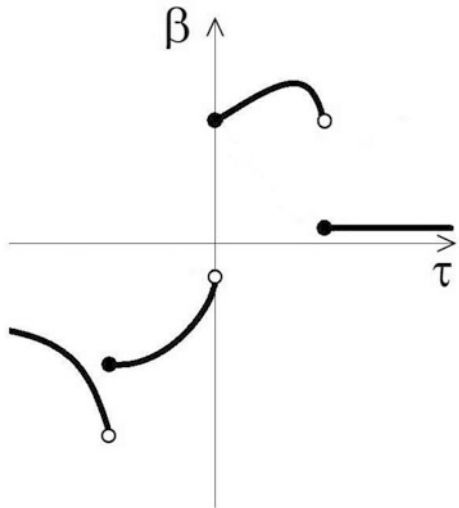


Fig. 2 Graph of the function $\tau \mapsto \beta(\tau)$



In our framework, the approach of Moreau and Panagiotopoulos consists of introducing a possibly discontinuous function $\beta \in L_{loc}^\infty(\mathbb{R}; \mathbb{R})$ such that left limit $\beta(v^-)$ and right limit $\beta(v^+)$ exist for all $v \in \mathbb{R}$, and so that

$$\mathcal{F}(v) = [\min\{\beta(v^-), \beta(v^+)\}, \max\{\beta(v^-), \beta(v^+)\}], \quad (v \in \mathbb{R}). \tag{2}$$

For example, the function β depicted in Figure 2 is deduced from the set-valued graph in Figure 1.

Let us now introduce the function $\Phi : \mathbb{R} \rightarrow \mathbb{R}$ by the formula:

$$\Phi(v) = \int_0^v \beta(\tau) d\tau, \quad (v \in \mathbb{R}). \quad (3)$$

Here, $\beta \in L_{loc}^\infty(\mathbb{R}; \mathbb{R})$, and the function Φ is thus locally Lipschitz. Moreover, a fundamental result in set-valued analysis due to K.C. Chang (see e.g. [11]) ensures that

$$\partial\Phi(v) = [\min\{\beta(v^-), \beta(v^+)\}, \max\{\beta(v^-), \beta(v^+)\}], \quad (v \in \mathbb{R}).$$

Here, for $v \in \mathbb{R}$, $\partial\Phi(v)$ denotes the Clarke's subdifferential (see e.g. [12]) of Φ at v defined by

$$\partial\Phi(v) := \text{co}\left\{ \lim_{n \rightarrow +\infty} \Phi'(v_n) : v_n \rightarrow v, v_n \in D_\Phi \right\},$$

where “co” refers to the convex hull and D_Φ stands for the set of differentiability points of Φ . We have

$$\partial\Phi(v) = \{\xi \in \mathbb{R} : \xi \cdot h \leq \Phi^\circ(v; h), \forall h \in \mathbb{R}\},$$

with

$$\Phi^\circ(v; h) := \limsup_{\substack{\lambda \downarrow 0 \\ w \rightarrow v}} \frac{\Phi(w + \lambda h) - \Phi(w)}{\lambda}.$$

Then,

$$\mathcal{F}(v) = \partial\Phi(v), \quad (v \in \mathbb{R}),$$

and the relation in (1) can be written as

$$F \in \partial\Phi(v), \quad (v \in \mathbb{R}). \quad (4)$$

If, in addition, the function Φ is convex, then

$$\partial\Phi(v) = \{w \in \mathbb{R} : \Phi(h) - \Phi(v) \geq w(h - v), \forall h \in \mathbb{R}\},$$

and Equation (4) reduces to the variational inequality

$$\Phi(h) - \Phi(v) - F(h - v) \geq 0, \quad \forall h \in \mathbb{R},$$

or equivalently

$$\Phi(v + \delta) - \Phi(v) - F\delta \geq 0, \quad \forall \delta \in \mathbb{R}.$$

Roughly speaking, $\partial\Phi$ results from the possibly discontinuous function β by “filling in the gaps.” In other rough words, Φ appears as a “primitive” of \mathcal{F} in the sense that the “derivative” (in the sense of Clarke) of Φ recovers the set-valued function \mathcal{F} .

3 Models of Frictions

We consider a system involving a block of mass $m > 0$ subjected to some external force $\vec{F}_e = F_e \vec{i}$ and which sticks or slips on another fixed body as depicted in Figure 5. Let us denote, respectively, by x , \dot{x} , and \ddot{x} the position, velocity, and acceleration of the sliding body. We have

$$\vec{G} = -mg \vec{j},$$

and we set

$$\vec{R} = -F \vec{i} + \lambda \vec{j}.$$

We have $\lambda = mg$, and $\vec{F} = -F \vec{i}$ denotes the friction force. The constant $g = 9.81 \text{ (m/s}^2\text{)}$ is the acceleration of gravity. The equation of motion for the system is

$$m\ddot{x} = F_e - F. \quad (5)$$

Let us set $v = \dot{x}$ to denote the sliding velocity. As soon as the system slips then $v \neq 0$ and F is a function of the sliding velocity v . At $v = 0$, the system sticks and F takes a value that is determined by other elements of the system. If F exceeds the breakaway force level, then the system switches back to the slip mode (Figure 3).

Example 1 (Set-Valued Coulomb Friction Model) The Coulomb model is a very simple mathematical formulation of the frictional phenomena. It is widely used by engineers to study systems with dry friction. The Coulomb friction model is also called Amontons–Coulomb friction model so as to refer to the work by Guillaume Amontons and Charles-Augustin de Coulomb (see e.g. [5, 20]). The Coulomb model expresses that friction opposes motion and that its magnitude is independent of the sliding velocity $v = \dot{x}$. The model is

$$F(v) = \begin{cases} -F_C & \text{if } v < 0, \\ +F_C & \text{if } v > 0, \end{cases}$$

where F_C is the Coulomb friction force proportional to the normal load $F_N = mg$ in the contact, i.e.

$$F_C = \mu F_N$$

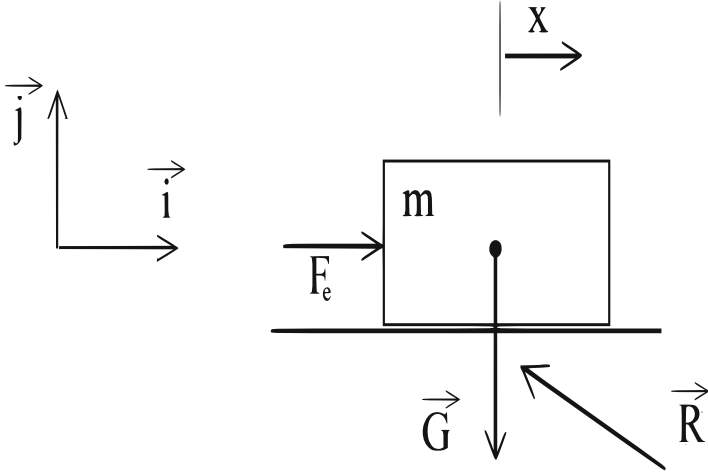


Fig. 3 Basic example

with $\mu > 0$. The coefficient μ is called the Coulomb friction coefficient. It is also called the dynamic friction coefficient. In this model, the value of the friction force is not specified for zero sliding velocity ($v = 0$); it can take any value in the interval $[-F_C, +F_C]$, i.e.

$$v = 0 \implies F \in [-F_C, +F_C].$$

We may thus write

$$F \in \mathcal{F}(v)$$

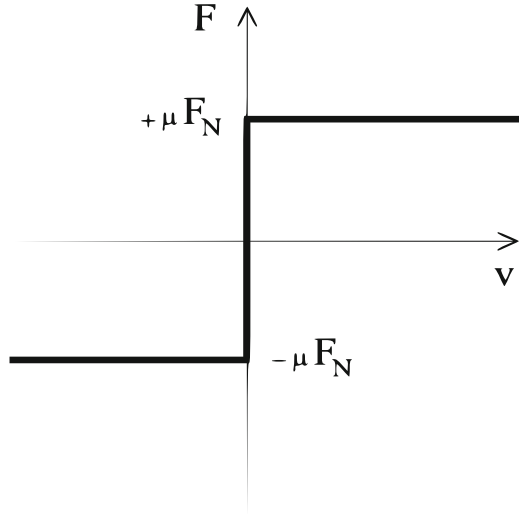
with

$$\mathcal{F}(v) = \begin{cases} -F_C & \text{if } v < 0, \\ [-F_C, +F_C] & \text{if } v = 0, \\ +F_C & \text{if } v > 0. \end{cases}$$

The graph of \mathcal{F} is depicted in Figure 4. We may deduce from this graph the function β defined by

$$\beta(\tau) = \begin{cases} -F_C & \text{if } \tau < 0, \\ +F_C & \text{if } \tau \geq 0. \end{cases}$$

Fig. 4 Coulomb friction model



Using the superpotential

$$\Phi_C(v) = \int_0^v \beta(\tau) d\tau = F_C |v|,$$

we may reduce the Coulomb model to the mathematical formula:

$$F \in \partial \Phi_C(v).$$

By using this model, one leave aside the complicated transition processes between “slip” and “stick.” The function Φ is convex and Equation (5) reduces to the differential inclusion

$$m\ddot{x} - F_e \in -\partial \Phi_C(\dot{x}), \tag{6}$$

which is equivalent to the evolution variational inequality (of second order):

$$(m\ddot{x} - F_e)(h - \dot{x}) + \Phi_C(h) - \Phi_C(\dot{x}) \geq 0, \quad \forall h \in \mathbb{R}.$$

Example 2 (Coulomb Friction Model) The approach described in Example 1 gives a mathematical model that has been the subject of a great number of works in the mathematical literature. Many tools in convex analysis, optimization, and nonlinear analysis have indeed been developed to study differential inclusions and variational inequalities. A value of F at 0 is, however, usually specified in the engineering literature. The approach consists of expressing that the friction opposes motion as

long as the force applied F_e is lesser than the friction force. The resulting model is then given by

$$F = \mathcal{F}(v, F_e)$$

with

$$\mathcal{F}(v, F_e) = \begin{cases} -F_C & \text{if } v < 0, \\ -F_C & \text{if } v = 0 \text{ and } F_e \leq -F_C, \\ F_e & \text{if } v = 0 \text{ and } -F_C < F_e < F_C, \\ +F_C & \text{if } v = 0 \text{ and } F_e \geq +F_C, \\ +F_C & \text{if } v > 0. \end{cases}$$

This model is depicted in the engineering literature as in Figure 4. This leads to a mathematical interpretation of the model as a set-valued function of the variable v as in Example 1. It is confusing since \mathcal{F} is here a singled-valued function of the two variables v and F_e . Moreover, the function is discontinuous at $(0, F_e)$, and the sense of Equation (5) is not obvious. It is however convenient to deduce from Equation (5) a numerical model and simulate it on a computer.

Example 3 (Viscous Friction Model) The viscous friction model is defined by the formula:

$$F = \mathcal{F}(v)$$

with $\mathcal{F} : \mathbb{R} \rightarrow \mathbb{R}; v \mapsto \mathcal{F}(v)$ given by

$$\mathcal{F}(v) = k_v v,$$

where $k_v > 0$ is the viscous coefficient. This simple linear model can be considered as soon as a lubricant is used between the moving body and the fixed one. It can also be used to represent a damping. Here, \mathcal{F} is a single-valued function, and we may set

$$\beta(\tau) = k_v \tau.$$

Then,

$$\Phi(v) = \int_0^v \beta(\tau) d\tau = \frac{1}{2} k_v v^2.$$

Equation (5) reduces here to an ordinary differential equation of the form:

$$m\ddot{x} + k_v\dot{x} = F_e.$$

Example 4 (Set-Valued Viscous Coulomb Friction Model) A model integrating the viscous model and the Coulomb friction model is given by

$$F \in \mathcal{F}(v)$$

with $\mathcal{F} : \mathbb{R} \rightrightarrows \mathbb{R}; v \mapsto \mathcal{F}(v)$ defined by

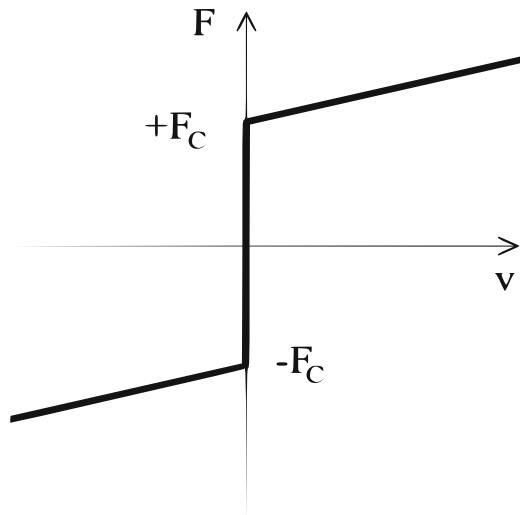
$$\mathcal{F}(v) = \begin{cases} -F_C + k_v v & \text{if } v < 0, \\ [-F_C, F_C] & \text{if } v = 0, \\ +F_C + k_v v & \text{if } v > 0. \end{cases}$$

The graph of \mathcal{F} is depicted in Figure 5. We may thus deduce the function β as

$$\beta(\tau) = k_v \tau + \begin{cases} -F_C & \text{if } \tau < 0, \\ +F_C & \text{if } \tau \geq 0, \end{cases}$$

It results that the viscous Coulomb model can be written as

Fig. 5 Viscous Coulomb friction model



$$F \in \partial\Psi(v),$$

where

$$\Psi(v) = \int_0^v \beta(\tau)d\tau = \frac{1}{2}k_v v^2 + F_C|v|.$$

Note that here

$$\partial\Psi(v) = k_v v + \partial\Phi_C(v),$$

with Φ_C as in Example 1:

$$\Phi_C(v) = F_C|v|.$$

Equation (5) reduces to the following differential inclusion:

$$m\ddot{x} + k_v\dot{x} - F_e \in -\partial\Phi_C(\dot{x}), \tag{7}$$

which is equivalent to the variational inequality:

$$(m\ddot{x} + k_v\dot{x} - F_e)(h - \dot{x}) + \Phi_C(h) - \Phi_C(\dot{x}) \geq 0, \forall h \in \mathbb{R}.$$

Remark 1 We note that the presence of dry friction in Equations (6) and (7) will force the trajectory to converge to an equilibrium in finite time (see Figure 6 and also Remark 2).

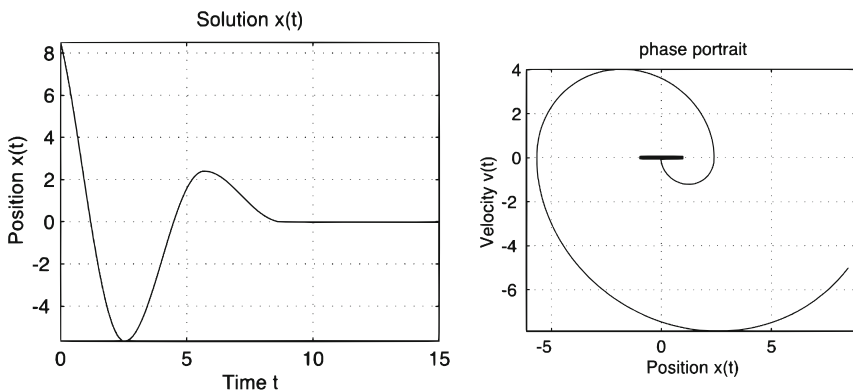


Fig. 6 Finite-time convergence of the trajectory to an equilibrium point of the system (7)

Example 5 (Viscous Coulomb Friction Model) The value of F at 0 in the model of Example 4 is usually specified in the engineering literature. The resulting model is then given by (see e.g. [5])

$$F = \mathcal{F}(v, F_e)$$

with $\mathcal{F} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}; (v, F_e) \mapsto \mathcal{F}(v, F_e)$ defined by

$$\mathcal{F}(v, F_e) = \begin{cases} -F_C + k_v v & \text{if } v < 0, \\ -F_C & \text{if } v = 0 \text{ and } F_e \leq -F_C, \\ F_e & \text{if } v = 0 \text{ and } -F_C < F_e < F_C, \\ +F_C & \text{if } v = 0 \text{ and } F_e \geq +F_C, \\ +F_C + k_v v & \text{if } v > 0. \end{cases}$$

Example 6 (Friction Model of Anderson, Söderberg, and Björklund) Another model combining Coulomb and viscous models is given by[5]

$$F = \mathcal{F}(v),$$

with

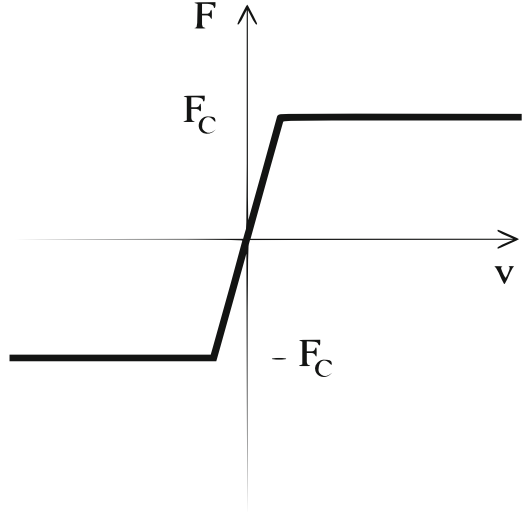
$$\mathcal{F}(v) = \begin{cases} F_C \max\{k_{\text{sat}} v, -1\} & \text{if } v < 0, \\ F_C \min\{k_{\text{sat}} v, +1\} & \text{if } v \geq 0, \end{cases}$$

where $k_{\text{sat}} > 0$ is a coefficient that determines how fast the force changes from $-$ to $+$. Here, \mathcal{F} is a single-valued function (see Figure 7) and is smooth in the sense that it does not present any discontinuities. We may set $\beta = \mathcal{F}$ and

$$\Phi(v) = \int_0^v \beta(\tau) d\tau = \begin{cases} -F_C v & \text{if } v < -\frac{1}{k_{\text{sat}}} \\ F_C v^2 & \text{if } |v| \leq \frac{1}{k_{\text{sat}}} \\ +F_C v & \text{if } v > \frac{1}{k_{\text{sat}}} \end{cases}$$

We have $F = \Phi'(v)$, and Equation (5) reduces to

Fig. 7 Combined Coulomb and viscous friction model



$$m\ddot{x} + \Phi'(\dot{x}) = F_e.$$

Example 7 (Set-Valued Stiction Friction Model) Friction acts like a spring when a small force is applied. This phenomenon is called “stiction.” Note that this term is a linguistic blend of the words “static” and “friction,” which could also be used to promote the principles of “Newspeak” of the famous writer George Orwell [27]. A model of stiction consists of expressing that the transition from stick to slip has to occur via the maximum static friction force $F_S = \mu_S F_N$ that may be higher than the maximum dynamic friction $F_C = \mu F_N$. Here, $\mu_S > 0$ denotes the friction coefficient in the slip phase, and F_S is called the stiction force.

The model is given by (see Figure 8)

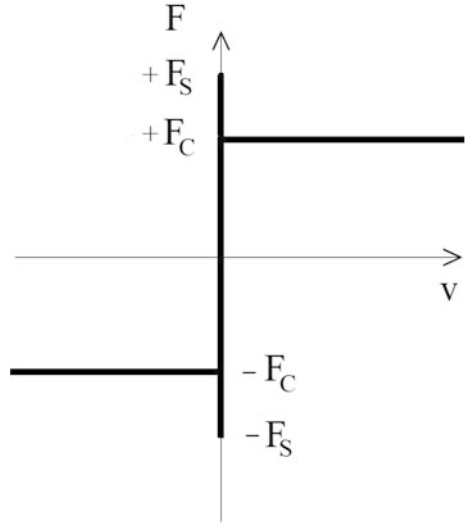
$$F \in \mathcal{F}(v)$$

with $\mathcal{F} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$; $(v, F_e) \mapsto \mathcal{F}(v, F_e)$ is defined by

$$\mathcal{F}(v) = \begin{cases} -F_C & \text{if } v < 0, \\ [-F_S, F_S] & \text{if } v = 0, \\ +F_C & \text{if } v > 0. \end{cases}$$

This set-valued function with $F_S > F_C$ does however not have good mathematical properties. It can, in particular, not be formulated as the subdifferential neither of

Fig. 8 Stiction model



a convex function nor of a locally Lipschitz one. Moreover, a transition from stick to slip is possible for $|F_e| < F_S$, and this is mechanically not consistent [9, 19].

Example 8 (Stiction Friction Model) The stiction friction model is usually used in specifying the value of F at 0 as in Example 2. The resulting model is then given by (see e.g. [7])

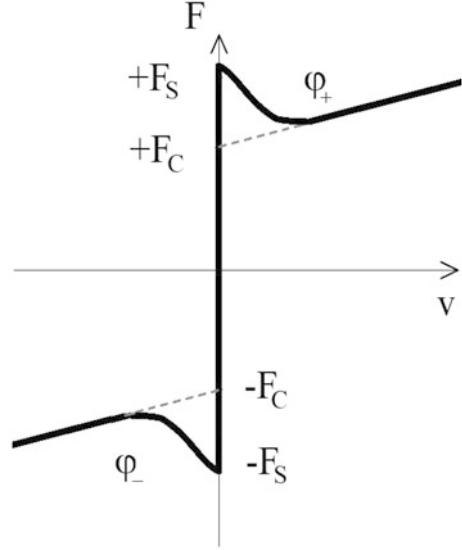
$$F = \mathcal{F}(v, F_e),$$

with

$$\mathcal{F}(v, F_e) = \begin{cases} -F_C & \text{if } v < 0, \\ -F_S & \text{if } v = 0 \text{ and } F_e \leq -F_S, \\ F_e & \text{if } v = 0 \text{ and } -F_C < F_e < F_C, \\ +F_S & \text{if } v = 0 \text{ and } F_e \geq +F_S, \\ +F_C & \text{if } v > 0. \end{cases}$$

Example 9 (Set-Valued Stribeck Friction Model) Most sliding contacts are lubricated, and Stribeck [35] observed that the friction force does not drop suddenly when velocity increases but follows a continuous curve as depicted in Figure 9.

Fig. 9 Set-valued stribeck friction model



The friction decreases with increased sliding speed until a mixed or full film situation is reached. Then, the friction can either be constant, increase, or decrease somewhat with increased sliding speed due to viscous and thermal effects. The velocity at which the friction force is minimal is called the Stribeck velocity. A modern set-valued formulation of the Stribeck friction is

$$F \in \mathcal{F}(v),$$

with

$$\mathcal{F}(v) = \begin{cases} \varphi_-(v) & \text{if } v < 0, \\ [-F_S, F_S] & \text{if } v = 0, \\ \varphi_+(v) & \text{if } v > 0, \end{cases}$$

where $F_S > 0$ is the maximum static force. The functions φ_- and φ_+ are given by the formulae:

$$(\forall v \leq 0) : \varphi_-(v) = k_v v - F_C - (F_S - F_C)e^{-|\frac{v}{v_s}|^\sigma}$$

and

$$(\forall v \geq 0) : \varphi_+(v) = k_v v + F_C + (F_S - F_C)e^{-|\frac{v}{v_s}|^\sigma},$$

where $\sigma > 0$ is an empirical exponent and $v_s > 0$ is an empirical coefficient called the sliding speed coefficient. Different values for σ have been used in the engineering literature [6]. Armstrong-Hélouvy [6] employs $\sigma = 2$. J. Čerkala and A. Jadlovska [10] use $\sigma = 1$ in the study of a two-wheel robot dynamic with differential chassis. Note also that other models for φ_- and φ_+ may be found in the engineering literature [6]. Let us now set

$$\varphi(v) = \begin{cases} \varphi_-(v) + F_S & \text{if } v < 0, \\ \varphi_+(v) - F_S & \text{if } v \geq 0. \end{cases}$$

The function is continuous on \mathbb{R} . It is clear on $] -\infty, 0[\cup] 0, +\infty[$ and it is also true at 0 since

$$\varphi(0^+) = 0 = \varphi(0^-).$$

We have

$$\begin{aligned} \mathcal{F}(v) &= \begin{cases} \varphi_-(v) & \text{if } v < 0, \\ [-F_S, F_S] & \text{if } v = 0, \\ \varphi_+(v) & \text{if } v > 0, \end{cases} = \begin{cases} \varphi(v) - F_S & \text{if } v < 0, \\ [-F_S, F_S] & \text{if } v = 0, \\ \varphi(v) + F_S & \text{if } v > 0, \end{cases} \\ &= \varphi(v) + \begin{cases} -F_S & \text{if } v < 0, \\ [-F_S, F_S] & \text{if } v = 0, \\ +F_S & \text{if } v > 0, \end{cases} \\ &= \varphi(v) + \partial\Phi_S(v), \end{aligned}$$

where

$$\Phi_S(v) = F_S|v|.$$

Equation (5) reduces to the differential inclusion

$$m\ddot{x} - F_e \in -\varphi(\dot{x}) - \partial\Phi_S(\dot{x}),$$

which is equivalent to the evolution variational inequality:

$$(m\ddot{x} - F_e + \varphi(\dot{x}))(h - \dot{x}) + \Phi_S(h) - \Phi_S(\dot{x}) \geq 0, \quad \forall h \in \mathbb{R}.$$

Example 10 (Stribeck Friction Model) The Stribeck friction model is usually used by engineers in specifying the value of F at 0 as in Example 2. The resulting model is then given by (see e.g. [7])

$$F = \mathcal{F}(v, F_e),$$

with

$$\mathcal{F}(v, F_e) = \begin{cases} \varphi_-(v) & \text{if } v < 0, \\ -F_S & \text{if } v = 0 \text{ and } F_e \leq -F_S, \\ F_e & \text{if } v = 0 \text{ and } -F_C < F_e < F_C, \\ +F_S & \text{if } v = 0 \text{ and } F_e \geq +F_S, \\ \varphi_+(v) & \text{if } v > 0. \end{cases}$$

Example 11 (Karnopp Friction Model) The Karnopp friction model [17] is a variant of the stiction friction model which includes a small neighborhood $[-D, +D]$ of zero velocity. More precisely, the Karnopp model is given by

$$F = \mathcal{F}(v, F_e)$$

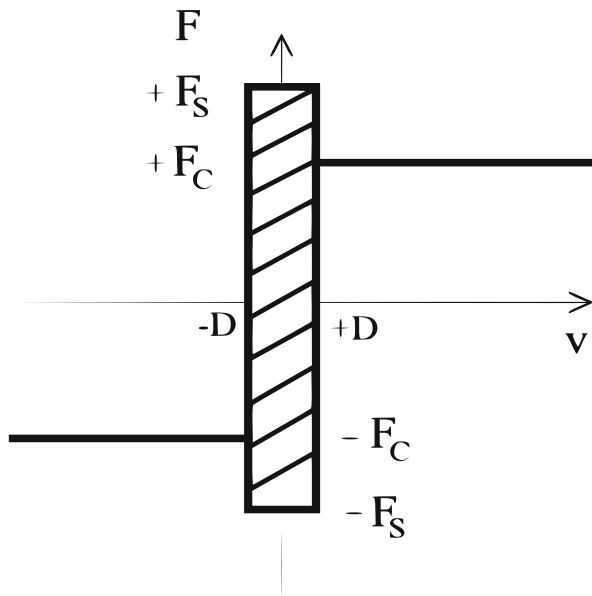
with

$$\mathcal{F}(v, F_e) = \begin{cases} -F_C & \text{if } v < 0, \\ -F_S & \text{if } v = 0 \text{ and } F_e \leq -F_S, \\ F_e & \text{if } v \in [-D, +D] \text{ and } -F_C < F_e < F_C, \\ +F_S & \text{if } v = 0 \text{ and } F_e \geq +F_S, \\ +F_C & \text{if } v > 0. \end{cases}$$

This model is depicted in the engineering literature as in Figure 10. The idea to create a dead zone of zero velocity is a remedy of the numerical problem of detecting when the velocity is equal to zero.

The idea of Karnopp can also be applied to the models described in Examples 2, 5, and 10. It is widely used in the engineering literature (see e.g. [8, 18, 32, 34, 36, 37]).

Fig. 10 Karnopp friction model



4 Stability Analysis of a Two-Degree-of-Freedom Mechanical System

Let us here consider the following two-degree-of-freedom mechanical system where two masses are connected with three springs and three dampers as shown in Figure 11. The equations of motion are given by

$$\begin{cases} m_1 \ddot{x}_1 + c_1 \dot{x}_1 + c_2(\dot{x}_1 - \dot{x}_2) + k_1 x_1 + k_2(x_1 - x_2) = -F_1, \\ m_2 \ddot{x}_2 + c_2(\dot{x}_2 - \dot{x}_1) + c_3 \dot{x}_2 + k_2(x_2 - x_1) + k_3 x_2 = -F_2, \end{cases}$$

where F_1 and F_2 are friction forces. We suppose that the friction forces F_1 and F_2 can be described by the Coulomb friction model, i.e.

$$F_1 \in \partial \Phi_1(\dot{x}_1), \quad F_2 \in \partial \Phi_2(\dot{x}_2),$$

where

$$\Phi_1(v) = \mu_1 m_1 g |v|, \quad \Phi_2(v) = \mu_2 m_2 g |v| \quad (v \in \mathbb{R}).$$

The equations of motion reduce to

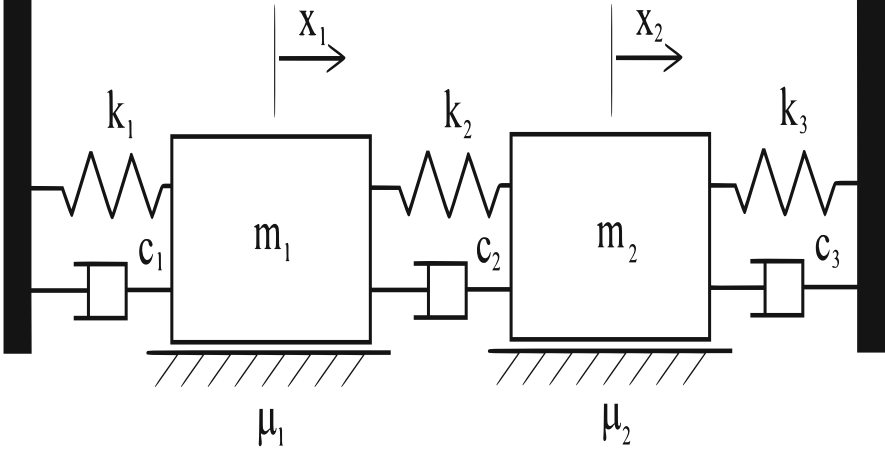


Fig. 11 $m_i > 0$ is the mass of sliding body i ($i = 1, 2$), $k_i > 0$ is the stiffness of spring i ($i = 1, 2, 3$), $c_i > 0$ is the damping coefficient of damper i ($i = 1, 2, 3$), and μ_i is a Coulomb friction coefficient ($i = 1, 2$)

$$\begin{cases} m_1 \ddot{x}_1 + c_1 \dot{x}_1 + c_2 (\dot{x}_1 - \dot{x}_2) + k_1 x_1 + k_2 (x_1 - x_2) \in -\partial \Phi_1(\dot{x}_1), \\ m_2 \ddot{x}_2 + c_2 (\dot{x}_2 - \dot{x}_1) + c_3 \dot{x}_2 + k_2 (x_2 - x_1) + k_3 x_2 \in -\partial \Phi_2(\dot{x}_2). \end{cases} \quad (8)$$

Let us also consider the initial conditions:

$$x(0) = x_0, \quad \dot{x}(0) = v_0, \quad (9)$$

for some $x_0, v_0 \in \mathbb{R}$. The equations of motion can be written as the following system:

$$M \ddot{x}(t) + C \dot{x}(t) + K x(t) = -\partial \Phi(\dot{x}(t)) \quad (t \geq 0), \quad (10)$$

with

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad M = \begin{pmatrix} m_1 & 0 \\ 0 & m_2 \end{pmatrix},$$

$$C = \begin{pmatrix} c_1 + c_2 & -c_2 \\ -c_2 & c_2 + c_3 \end{pmatrix}, \quad K = \begin{pmatrix} k_1 + k_2 & -k_2 \\ -k_2 & k_2 + k_3 \end{pmatrix},$$

and where

$$\Phi(v) = \Phi_1(v_1) + \Phi_2(v_2) \quad (v = (v_1, v_2) \in \mathbb{R}^2).$$

It is easy to transform this system to the following first-order differential inclusion:

$$\dot{X}(t) \in AX(t) - \partial\mathcal{E}(X(t)) \quad (t \geq 0), \quad (11)$$

where

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \dot{x}_1 \\ \dot{x}_2 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\frac{k_1 + k_2}{m_1} & \frac{k_2}{m_1} & -\frac{c_1 + c_2}{m_1} & \frac{c_2}{m_1} \\ \frac{k_2}{m_2} & -\frac{k_2 + k_3}{m_2} & \frac{c_2}{m_2} & -\frac{c_2 + c_3}{m_2} \end{pmatrix},$$

and where $\mathcal{E} : \mathbb{R}^4 \rightarrow \mathbb{R}; v = (v_1, v_2, v_3, v_4) \mapsto \mathcal{E}(v)$ is defined by

$$\mathcal{E}(V) = \frac{1}{m_1} \Phi_1(V_3) + \frac{1}{m_2} \Phi_2(V_4) = \mu_1 g |V_3| + \mu_2 g |V_4| \quad (V \in \mathbb{R}^4).$$

We may thus consider the system:

$$\dot{X}(t) \in AX(t) - \partial\mathcal{E}(X(t)), \quad (t \geq 0), \quad (12)$$

together with the initial condition:

$$X(0) = X_0, \quad (13)$$

where $X_0 \in \mathbb{R}^4$. All assumptions of Theorem 3.2.1 (page 49 in [1]) are satisfied, so we conclude that for each initial condition $X_0 \in \mathbb{R}^4$, there exists a unique trajectory $X \in C^0([0, +\infty[; \mathbb{R}^4)$, right differentiable on $[0, +\infty[$, with $\dot{X} \in L_{\text{loc}}^\infty([0, +\infty[; \mathbb{R}^4)$ and satisfying (12)–(13). Hence, for each initial position $x_0 \in \mathbb{R}^2$ and each initial velocity $v_0 \in \mathbb{R}^2$, there exists a unique trajectory $x \in C^1([0, +\infty[; \mathbb{R}^2)$ such that \dot{x} is right differentiable on $]0, +\infty[$, with $\dot{x} \in L_{\text{loc}}^\infty([0, +\infty[; \mathbb{R}^2)$ and satisfying the system in (8). Let us now denote by \mathcal{W} the set of stationary solutions of (8), i.e.

$$\begin{aligned} \mathcal{W} &= \{x \in \mathbb{R}^2 : Kx \in -\partial\Phi(0, 0)\} \\ &= \{x \in \mathbb{R}^2 : -\mu_1 m_1 g \leq (k_1 + k_2)x_1 - k_2 x_2 \leq \mu_1 m_1 g \text{ and} \\ &\quad -\mu_2 m_2 g \leq -k_2 x_1 + (k_2 + k_3)x_2 \leq \mu_2 m_2 g\} \end{aligned}$$

$$= \{x \in \mathbb{R}^2 : |x_1| \leq R_1 \text{ and } |x_2| \leq R_2\}$$

with

$$R_1 = \frac{(k_2 + k_3)\mu_1 m_1 g + k_2 \mu_2 m_2 g}{k_1 k_2 + k_1 k_3 + k_2 k_3}, \quad R_2 = \frac{(k_1 + k_2)\mu_2 m_2 g + k_2 \mu_1 m_1 g}{k_1 k_2 + k_1 k_3 + k_2 k_3}.$$

We say that a stationary solution $\bar{x} \in \mathscr{W}$ is *stable* provided that for any $\varepsilon > 0$, there exists an $\eta(\varepsilon) > 0$ such that for any $x_0 \in \mathbb{R}^2, v_0 \in \mathbb{R}^2$ with $\sqrt{\|x_0 - \bar{x}\|^2 + \|v_0\|^2} \leq \eta$, the solution $x(\cdot; x_0, v_0)$ of (8)–(9) satisfies

$$(\forall t \geq 0) : \sqrt{\|x(t; x_0, v_0) - \bar{x}\|^2 + \|\dot{x}(t; x_0, v_0)\|^2} \leq \varepsilon.$$

If

$$\lim_{t \rightarrow +\infty} \|x(t; x_0, v_0) - \bar{x}\| = 0$$

and

$$\lim_{t \rightarrow +\infty} \|\dot{x}(t; x_0, v_0)\| = 0,$$

then we say that the stationary point \bar{x} is *attractive*. A stable and attractive stationary point \bar{x} is called *asymptotically stable*. Since the matrices K and C are both symmetric and positive definite, we invoke Theorem 4.2.2, page 67 in [1], to conclude that every stationary solution $\bar{x} \in \mathscr{W}$ is Lyapunov stable. Using Theorem 4.2.4 in [1], we can also prove the following attractivity result:

$$\lim_{t \rightarrow +\infty} \text{dist}(x(t; x_0, v_0), \mathscr{W}) = 0 \text{ and } \lim_{t \rightarrow +\infty} \dot{x}(t; x_0, v_0) = 0.$$

Remark 2 Since $(0, 0) \in \text{Int}(\partial\Phi((0, 0)))$, it is also possible to show that

$$\lim_{t \rightarrow +\infty} x(t; x_0, v_0) = \bar{x},$$

where $\bar{x} \in \mathscr{W}$ is a stationary point. If, in addition, $-K\bar{x}$ is not in the boundary of $\partial\Phi(0, 0) = [-\mu_1 m_1 g, +\mu_1 m_1 g] \times [-\mu_2 m_2 g, +\mu_2 m_2 g]$, then there exists $T \geq 0$ such that

$$(\forall t \geq T) : x(t; x_0, v_0) = \bar{x}.$$

For more details, we refer to Theorem 24.8 in [3].

References

1. S. Adly, *A Variational Approach to Nonsmooth Dynamics: Applications in Unilateral Mechanics and Electronics*. SpringerBriefs (Springer, Berlin, 2017)
2. S. Adly, D. Goeleven, A stability theory for second-order nonsmooth dynamical systems with applications to friction problems. *J. Math. Pures Appl.* **83**, 17–51 (2004)
3. S. Adly, H. Attouch, A. Cabot, Finite time stabilization of nonlinear oscillators subject to dry friction, in *Nonsmooth Mechanics and Analysis*. Adv. Mech. Math., vol. 12 (Springer, New York, 2006), pp. 289–304
4. G. Amontons, On the resistance originating in machines, in *Proceedings of the French Royal Academy of Sciences* (1699), pp. 206–222
5. S. Anderson, A. Söderberg, S. Björklund, Friction models for sliding dry, boundary and mixed lubricated contacts. *Tribol. Int.* **40**, 580–587 (2007)
6. B. Armstrong-Hélouvy, *Control of Machine with Friction*. The Kluwer International Series in Engineering and Computer Science, Robotics (Springer Science+Business Media, New York, 1991)
7. K.J. Åström, Control of systems with friction, in *Proceedings of the Fourth International Conferences on Motion and Vibration Control* (1998), pp. 25–32
8. P. Bingham, S. Theodosiades, T. Saunders, E. Grant, R. Daubney, A study on automotive drivetrain transient response to ‘clutch abuse’ events. *Proc. Inst. Mech. Eng. P D J. Automob. Eng.* **230**(10), 1–14 (2016)
9. P.A. Bliman, M. Sorine, Easy-to-use realistic dry friction models for automatic control, in *Proceedings of the 3rd European Control Conference*, Roma (1995), pp. 3788–3794
10. J. Čerkala, A. Jadovská, Mobile robot dynamics with friction in simulink, in *Conference Paper, 22th Annual Conference Proceedings of the International Scientific Conference*, Technical Computing Bratislava, Bratislava, Slovakia (2014)
11. K.C. Chang, Variational methods for non-differentiable functionals and their applications to partial differential equations. *J. Math. Anal. Appl.* **80**, 102–129 (1981)
12. F.H. Clarke, *Optimization and Nonsmooth Analysis* (Wiley, New York, 1983)
13. C.A. Coulomb, Théorie des machines simples, en ayant égard au frottement de leurs parties, et a la roideur dews cordages. *Mem. Math Phys. Paris X*, 161–342 (1785)
14. L. Da Vinci (1519), *The Notebooks*, ed. by J.P. Richter (Dover, New York, 1970)
15. D. Goeleven, *Complementarity and Variational Inequalities in Electronics* (Academic, Elsevier, London, 2017)
16. M. Jean, J.J. Moreau, Unilaterality and dry friction in the dynamics of rigid body collections, in *Proc. Contact Mechanics Int. Symp.*, ed. by A. Curnier (Presses Polytechniques et Universitaires Romandes, Lausanne, 1992), pp. 31–48
17. D. Karnopp, Computer simulation of stick-slip friction in mechanical dynamic control. *Trans. ASME* **107**, 100–103 (1985)
18. M. Kidouche, R. Riane, On the design of proportional integral observer for a rotary drilling system, in *8th CHAOS Conference Proceedings, Henri Poincaré Institute*, Paris (2015)
19. R.I. Leine, H. Nijmeijer, *Dynamics and Bifurcation of Non-smooth Mechanical Systems*. Lecture Notes in Applied and Computational Mechanics, vol. 18 (Springer, Berlin, 2004)
20. Y.F. Liu, J. Li, Z.M. Zhang, X.H. Hu, W.J. Zhang, Experimental comparison of five friction models on the same test-bed of the micro stick-slip motion system. *Mech. Sci.* **6**, 15–28 (2015)
21. J.J. Moreau, La Notion du Surpotentiel et les Liaisons Unilatérales on Elastostatique. *C.R. Acad. Sci. Paris* **167A**, 954–957 (1968)
22. J.J. Moreau, Dynamique des systèmes à liaisons unilatérales avec frottement sec éventuel; essais numériques. *Tech. Rep. 85-1, LMGC*, Montpellier, 1986
23. J.J. Moreau, Unilateral contact and dry friction in finite freedom dynamics, in *Non-Smooth Mechanics and Applications*, ed. by J.J. Moreau, P.D. Panagiotopoulos. CISM Courses and Lectures, vol. 302 (Springer, Wien, 1988), pp. 1–82

24. A.J. Morin, New friction experiments carried out at Metz in 1831–1833. *Proc. Fr. R. Acad. Sci.* **4**, 1–128 (1833)
25. H. Olsson, Control systems with friction. Department of Automatic Control, Lund Institute of Technology (LTH), Lund, 1996
26. H. Olsson, K.J. Aström, C. Canudas de Wit, M. G öfvert, P. Lischinsky, Friction models and friction compensation. *Eur. J. Control* **4**(3), 176–195 (1998)
27. G. Orwell, *Nineteen Eighty-Four* (Secker and Warburg, London, 1949)
28. P.D. Panagiotopoulos, Non-convex superpotentials in the sense of F.H. Clarke and applications. *Mech. Res. Commun.* **8**, 335–340 (1981)
29. P.D. Panagiotopoulos, *Hemivariational Inequalities. Applications in Mechanics and Engineering* (Springer, Berlin, 1993)
30. V.L. Popov, *Contact Mechanics and Friction, Physical Principles and Applications* (Springer, Berlin, 2010)
31. J. Rauch, Discontinuous semilinear differential equations and multiple valued maps. *Proc. Amer. Math. Soc.* **64**, 277–282 (1977)
32. L. Ravanbod-Shirazi, A. Besançon-Voda, Friction identification using the Karnopp model applied to an electropneumatic actuator. *Proc. Inst. Mech. Eng. P I J. Syst. Control Eng.* **217**(2), 123–138 (2003)
33. O. Reynolds, On the Theory of Lubrication and its application to Mr. Beauchamp Tower's experiments, including an experimental determination of the viscosity of olive oil. *Philos. Trans. R. Soc.* **177**, 157–234 (1886)
34. C. Simone, A.M. Okamura, Modelling of needle insertion for robot-assisted percutaneous therapy, in *Proceedings of the 2002 IEEE International Conference on Robotics & Automation*, Washington, DC (2002)
35. R. Stribeck, Die Wesentlichen Eigenschaften der Gleit- und Rollenlager. *Z. Verein. Deut. Ing.* **46**(38), 1341–1348 (1902)
36. R. Trentini, A. Campos, A. da Silva Silveira, G. Espindola, Identification of friction effects in a linear positioning servopneumatic system. *Sci. Eng. J.* **22**(1), 97–101 (2013)
37. H. Xia, L. Han, C. Pan, H. Jia, L. Yu, Simulation of motion interactions of a 2-DOF linear piezoelectric impact drive mechanism with a single friction interface. *Appl. Sci.* **8**, 1400 (2018). <https://doi.org/10.3390/app8081400>

A Survey on Markov's Theorem on Zeros of Orthogonal Polynomials



Kenier Castillo, Marisa de Souza Costa, and Fernando Rodrigo Rafaeli

Abstract This manuscript is an extended version of the paper by the same authors who appeared in Castillo et al. (Appl Math Comput 339:390–397, 2018). It briefly surveys a Markov's result dating back to the end of the nineteenth century, which is related to zeros of orthogonal polynomials.

1 Introduction

Markov's theorem, dating back to the late nineteenth century, furnishes a method for obtaining information about zeros of orthogonal polynomials from the weight function related to orthogonality. Formally, adopting modern terminology, his result is stated as follows (see [27]):

Theorem 1 (Markov [27]) *Let $\{p_n(x, t)\}$ be a sequence of polynomials that are orthogonal on the interval $A = (a, b)$ with respect to the weight function $\omega(x, t)$ that depends on a parameter $t, t \in B = (c, d)$, i.e.,*

$$\int_a^b p_n(x, t)p_m(x, t)\omega(x, t)dx = 0, \quad m \neq n.$$

Suppose that $\omega(x, t)$ is positive and has a continuous first derivative with respect to t for $x \in A, t \in B$. Furthermore, assume that

K. Castillo
CMUC, Department of Mathematics, University of Coimbra, Coimbra, Portugal
e-mail: kenier@mat.uc.pt

M. S. Costa · F. R. Rafaeli (✉)
FAMAT-UFU, Department of Mathematics, Federal University of Uberlândia, Uberlândia,
Minas Gerais, Brazil
e-mail: marisasc@ufu.br; rafaeli@ufu.br

$$\int_a^b x^k \frac{\partial \omega}{\partial t}(x, t) dx, \quad k = 0, 1, \dots, 2n - 1,$$

converges uniformly for t in every compact subinterval of B . Then, the zeros of $p_n(x, t)$ are increasing (decreasing) functions of t , $t \in B$, provided that

$$\frac{1}{\omega(x, t)} \frac{\partial \omega}{\partial t}(x, t)$$

is an increasing (decreasing) function of x , $x \in A$.

Markov's proof is based on the orthogonality relation (cf. [27, Equation 2]) together with the chain rule (cf. [27, Equation 5]), supposing that the zeros are defined implicitly as differentiable functions of the parameter. In addition, as an application of this result, Markov established that the zeros of Jacobi polynomials, which are orthogonal in $(-1, 1)$ with respect to the weight function $\omega(x, \alpha, \beta) = (1 - x)^\alpha (1 + x)^\beta$, $\alpha, \beta > -1$, are decreasing functions of α and increasing functions of β . Later, in 1939, Szegő, in his classical book [33, Theorem 6.12.1, p. 115], provided a different proof of Markov's theorem. Szegő referred his proof of Theorem 1 in the following way [33, Footnote 31, p. 116]: "This proof does not differ essentially from the original one due to A. Markov, although the present arrangement is somewhat clearer.". Szegő's reasoning (argument, approach) is based on Gauss mechanical quadrature, which was an approach that Stieltjes suggested to handle the problem, see [32, Section 5, p. 391]. In 1971, Freud (see [12, Problem 16, p. 133]) formulated a version of Markov's theorem that is a little more general, considering sequences of polynomials orthogonal with respect to measures in the form $d\alpha(x, t) = \omega(x, t)dv(x)$. A proof of such result appears in Ismail [15, Theorem 3.2, p. 183] (see also in Ismail's book [16, Theorem 7.1.1, p. 204]). Ismail's argument of the proof is also based on Gauss mechanical quadrature. As a consequence, Ismail provided monotonicity properties for the zeros of Hahn and Meixner polynomials (see [16, Theorem 7.1.2, p. 205]). Kroó and Peherstorfer [23, Theorem 1], in a more general context of approximation theory, extended Markov's result to zeros of polynomials that have the minimal L_p -norm. Their approach is based on the implicit function theorem.

The main concern of this work derives from Markov's classic 1886 theorem. This allows the approach to be tailored toward measures with continuous and discrete parts, thus extending the Markov's result. This point at issue was posed by Ismail in his book as an open problem [16, Problem 24.9.1, p. 660] (see also [15, Problem 1, p. 187]). The question is stated as follows:

Problem 1 Let μ be a positive and nontrivial Radon measure on a compact set $A \subset \mathbb{R}$. Assume that $d\mu(x, t)$ has the form

$$d\alpha(x, t) + d\beta(x, t), \tag{1}$$

where $d\alpha(x, t) := \omega(x, t)d\nu(x)$ and $d\beta(x, t) := \sum_{i=0}^{\infty} J_i(t)\delta_{y_i(t)}$,¹ with $t \in B$, B being an open interval on \mathbb{R} . Determine sufficient conditions in order for the zeros of the polynomial $P_n(x, t)$ to be strictly increasing (decreasing) functions of t .

The manuscript is organized in the following way: in Section 2, the main result is stated and proved; in Section 3, some conclusions are drawn from the main result, including Markov's classic theorem, among others; finally, in Section 4, illustrative examples are given; in Sections 4.1 and 4.2, monotonicity properties of zeros of polynomials orthogonal with respect measures with discrete parts are investigated; in Section 4.3, monotonicity properties of zeros of Jacobi, Gegenbauer, and Laguerre orthogonal polynomials are reviewed; in Section 4.4, sharp monotonicity properties involving the zeros of Gegenbauer–Hermite, Jacobi–Laguerre, and Laguerre–Hermite orthogonal polynomials are derived; at last, in Section 4.5, monotonicity properties of zeros of Charlier, Meixner, Kravchuck, and Hahn orthogonal polynomials are revisited.

2 Main Result

The next result extends Markov's theorem to measure with continuous and discrete parts, giving an answer to Problem 1 (see [5]). For a result in the context of polynomials that have minimal L_p -norm, see [4, Theorem 1.1].

Theorem 2 *Assume the notation and conditions of Problem 1. Assume further the existence and continuity for each $x \in A$ and $t \in B$ of $(\partial\omega/\partial t)(x, t)$ and, in addition, suppose that*

$$G(t, x_1, \dots, x_n) := \sum_{i=0}^{\infty} g_i(t, x_1, \dots, x_n)$$

converges at $t = t_0$ and

$$\frac{\partial G}{\partial t}(t, x_1, \dots, x_n) := \sum_{i=0}^{\infty} \frac{\partial g_i}{\partial t}(t, x_1, \dots, x_n),$$

$$\frac{\partial G}{\partial x_j}(t, x_1, \dots, x_n) = \sum_{i=0}^{\infty} \frac{\partial g_i}{\partial x_j}(t, x_1, \dots, x_n),$$

converge uniformly for $t \in B$, where

¹The Dirac measure δ_y is a positive Radon measure whose support is the set $\{y\}$.

$$g_i(t, x_1, \dots, x_n) = J_i(t)(y_i(t) - x_k)^{-1} \prod_{j=1}^n (y_i(t) - x_j)^2,$$

and $(x_1, \dots, x_n) \in \mathbb{R}^n$. Denote by $x_1(t), \dots, x_n(t)$ the zeros of $P_n(x, t)$. Fix $k \in \{1, \dots, n\}$ and set

$$d_{k,i}(t) := \begin{cases} y_i(t) - x_k(t) & \text{if } y_i(t) \neq x_k(t), \\ 1 & \text{if } y_i(t) = x_k(t). \end{cases}$$

Define the function

$$R_{k,i}(t) := \sum_{j=0}^n{}' \frac{2 - \delta_{j,k}}{y_i(t) - x_j(t)},$$

where the prime means that the sum is over all values j and t for which $y_i(t) \neq x_j(t)$. Then $x_k(t)$ is a strictly increasing function for those values of t such that

$$\frac{1}{d_{k,i}(t)} \left\{ \frac{J_i'(t)}{J_i(t)} + y_i'(t)R_{k,i}(t) - \frac{1}{\omega(x_k(t), t)} \frac{\partial \omega}{\partial t}(x_k(t), t) \right\} \geq 0, \quad (2)$$

and

$$\frac{1}{\omega(x, t)} \frac{\partial \omega}{\partial t}(x, t) \quad (3)$$

is an increasing function of $x \in A$, provided that at least the inequality (2) is strict or the function (3) is nonconstant on A .

Proof The proof is based on the implicit function theorem and it is similar to the Markov's one.

Let $P_n(x, t) = (x - x_1(t)) \cdots (x - x_n(t))$ be the n th orthogonal polynomial with respect to (1). In other words, $P_n(x, t)$ satisfies the following orthogonality relations:

$$\int_a^b q(x) P_n(x, t) \omega(x, t) d\nu(x) + \sum_{i=0}^{\infty} J_i(t) q(y_i(t)) P_n(y_i(t), t) = 0 \quad (q \in \mathcal{P}_{n-1}). \quad (4)$$

Since $P_n(x_k(t), t) = 0$, by implicit function theorem,

$$\frac{\partial P_n}{\partial x}(x_k(t), t) \frac{dx_k}{dt}(t) + \frac{\partial P_n}{\partial t}(x_k(t), t) = 0,$$

that is,

$$\frac{dx_k}{dt}(t) = -\frac{\frac{\partial P_n}{\partial t}(x_k(t), t)}{\frac{\partial P_n}{\partial x}(x_k(t), t)}. \quad (5)$$

Now, take

$$q(x) = q(x, v) = \frac{P_n(x, v)}{x - x_k(v)} \in \mathcal{P}_{n-1},$$

and substitute it in the derivative of (4) with respect to t , and then let $v \rightarrow t$. The result is the following:

$$\begin{aligned} & \int \frac{[P_n(x, t)]^2}{x - x_k(t)} \frac{\partial \omega}{\partial t}(x, t) dv(x) + \sum_{i=0}^{\infty} \{J_i'(t) + J_i(t)y_i'(t)R_{k,i}(t)\} \frac{[P_n(y_i(t), t)]^2}{y_i(t) - x_k(t)} \\ & + \int \frac{P_n(x, t)}{x - x_k(t)} \frac{\partial P_n}{\partial t}(x, t) \omega(x, t) dv(x) \\ & + \sum_{i=0}^{\infty} J_i(t) \frac{P_n(y_i(t), t)}{y_i(t) - x_k(t)} \frac{\partial P_n}{\partial t}(y_i(t), t) = 0. \quad (6) \end{aligned}$$

On the other hand, if one takes

$$q(x) = q(x, t) = \left\{ \frac{\partial P_n}{\partial t}(x, t) - \frac{\partial P_n}{\partial t}(x_k(t), t) \right\} \frac{1}{x - x_k(t)} \in \mathcal{P}_{n-1},$$

substitutes it in (4), and subtracts the result from (6), one derives

$$\begin{aligned} & -\frac{\partial P_n}{\partial t}(x_k(t), t) \left\{ \int \frac{P_n(x, t)}{x - x_k(t)} \omega(x, t) dv(x) + \sum_{i=0}^{\infty} J_i(t) \frac{P_n(y_i(t), t)}{y_i(t) - x_k(t)} \right\} = \\ & \int \frac{[P_n(x, t)]^2}{x - x_k(t)} \frac{\partial \omega}{\partial t}(x, t) dv(x) \\ & + \sum_{i=0}^{\infty} \{J_i'(t) + J_i(t)y_i'(t)R_{k,i}(t)\} \frac{[P_n(y_i(t), t)]^2}{y_i(t) - x_k(t)}. \quad (7) \end{aligned}$$

Now, since

$$q(x) = q(x, t) = \frac{\frac{\partial P_n}{\partial x}(x_k(t), t)(x - x_k(t)) - P_n(x, t)}{(x - x_k(t))^2} \in \mathcal{P}_{n-2},$$

by the orthogonality relation (4), one obtains

$$\begin{aligned} \frac{\partial P_n}{\partial x}(x_k(t), t) \left\{ \int \frac{P_n(x, t)}{x - x_k(t)} \omega(x, t) dv(x) + \sum_{i=0}^{\infty} J_i(t) \frac{P_n(y_i(t), t)}{y_i(t) - x_k(t)} \right\} = \\ \int \frac{[P_n(x, t)]^2}{(x - x_k(t))^2} \omega(x, t) dv(x) + \sum_{i=0}^{\infty} J_i(t) \frac{[P_n(y_i(t), t)]^2}{(y_i(t) - x_k(t))^2}. \end{aligned} \quad (8)$$

Therefore, substituting (7) and (8) in (5) yields

$$\begin{aligned} \frac{dx_k}{dt}(t) = \\ \frac{\int \frac{[P_n(x, t)]^2}{x - x_k(t)} \frac{\partial \omega}{\partial t}(x, t) dv(x) + \sum_{i=0}^{\infty} \{J_i'(t) + J_i(t)y_i'(t)R_{k,i}(t)\} \frac{[P_n(y_i(t), t)]^2}{y_i(t) - x_k(t)}}{\int \frac{[P_n(x, t)]^2}{(x - x_k(t))^2} \omega(x, t) dv(x) + \sum_{i=0}^{\infty} J_i(t) \frac{[P_n(y_i(t), t)]^2}{(y_i(t) - x_k(t))^2}}. \end{aligned} \quad (9)$$

Clearly,

$$\frac{1}{\omega(x_k(t), t)} \frac{\partial \omega}{\partial t}(x_k(t), t) \int \frac{P_n(x, t)^2}{x - x_k(t)} d\mu(x, t) = 0. \quad (10)$$

Subtracting (10) from the numerator of the right-hand side of (9) yields

$$\begin{aligned} & \int \frac{[P_n(x, t)]^2}{x - x_k(t)} \frac{\partial \omega}{\partial t}(x, t) dv(x) + \sum_{i=0}^{\infty} \{J_i'(t) + J_i(t)y_i'(t)R_{k,i}(t)\} \frac{[P_n(y_i(t), t)]^2}{y_i(t) - x_k(t)} \\ &= \int \frac{[P_n(x, t)]^2}{x - x_k(t)} \left(\frac{1}{\omega(x, t)} \frac{\partial \omega}{\partial t}(x, t) - \frac{1}{\omega(x_k(t), t)} \frac{\partial \omega}{\partial t}(x_k(t), t) \right) \omega(x, t) dv(x) + \\ &+ \sum_{i=0}^{\infty} \left\{ J_i'(t) + J_i(t)y_i'(t)R_{k,i}(t) - \frac{J_i(t)}{\omega(x_k(t), t)} \frac{\partial \omega}{\partial t}(x_k(t), t) \right\} \frac{[P_n(y_i(t), t)]^2}{y_i(t) - x_k(t)}. \end{aligned} \quad (11)$$

It only remains to note that

$$\frac{1}{x - x_k(t)} \left(\frac{1}{\omega(x, t)} \frac{\partial \omega}{\partial t}(x, t) - \frac{1}{\omega(x_k(t), t)} \frac{\partial \omega}{\partial t}(x_k(t), t) \right) \geq 0.$$

Thus, the sign of $x'_k(t)$ is the same sign as the numerator of (9), and the desired result follows from (11).

3 Markov's Theorem and Its Descendants

In this section, Markov's classic theorem is derived from Theorem 2. In addition, Markov's theorem for even weight function is revisited too, together with other results.

The next result brings us back to Markov's theorem [27] (see also [33, Theorem 6.12.1, p. 115] and [16, Theorem 7.1.1, p. 204]).

Corollary 3.1 *Assume the notation and conditions of Theorem 2 under the constraint that $d\mu(x, t) = \omega(x, t)d\alpha(x)$. In this case, (9) becomes*

$$\begin{aligned} \frac{dx_k}{dt}(t) &= \frac{\int \frac{[P_n(x, t)]^2}{x - x_k(t)} \frac{\partial \omega}{\partial t}(x, t) dv(x)}{\int \frac{[P_n(x, t)]^2}{(x - x_k(t))^2} \omega(x, t) dv(x)} \\ &= \frac{\int \frac{[P_n(x, t)]^2}{x - x_k(t)} \left(\frac{1}{\omega(x, t)} \frac{\partial \omega}{\partial t}(x, t) - \frac{1}{\omega(x_k(t), t)} \frac{\partial \omega}{\partial t}(x_k(t), t) \right) \omega(x, t) dv(x)}{\int \frac{[P_n(x, t)]^2}{(x - x_k(t))^2} \omega(x, t) dv(x)}. \end{aligned} \tag{12}$$

Then, $x_k(t)$ is a strictly increasing (decreasing) function of t if

$$\frac{1}{\omega(x, t)} \frac{\partial \omega}{\partial t}(x, t)$$

is an increasing (decreasing) function of $x \in A$, provided that the last function be nonconstant on A .

Markov's result concerning the zeros of polynomials orthogonal with respect to even weight function was studied by Jordaan, Wang, and Zhou [18, Theorem 2.1]. This case also appears in [23, Corollary 2] in a more general context. For further results about these polynomials, see [6, Chapter 1, Sections 8 and 9].

Corollary 3.2 (Markov's Result for Even Weight Function) *Assume the notation and conditions of Theorem 2 under the constraint that $d\mu(x, t) = \omega(x, t) dx$.*

Suppose, in addition, that $\omega(x, t)$ is an even function of x in $A = (-a, a)$.² Then, the positive zeros $x_k(t)$ are strictly increasing (decreasing) functions of t if

$$\frac{1}{\omega(x, t)} \frac{\partial \omega}{\partial t}(x, t)$$

is an increasing (decreasing) function of $x \in (0, a)$, provided that the last function be nonconstant on $(0, a)$.

Proof Since $\omega(x, t)$ is an even function, then $P_n(-x, t) = (-1)^n P_n(x, t)$ (for further details, see [6, Chapter 1, Section 8]). Therefore, one can write

$$P_{2m}(x, t) = S_m(x^2, t) \quad \text{and} \quad P_{2m+1}(x, t) = x T_m(x^2, t),$$

where S_m and T_m are polynomials of degree m . Let $y_i^{(1)} = y_i^{(1)}(t)$ and $y_i^{(2)} = y_i^{(2)}(t)$, $i = 1, \dots, m$, be the zeros of the polynomials S_m and T_m , respectively. If x_i , $i = 1, \dots, [n/2]$, denote the positive zeros of the polynomial P_n , then

$$x_i = \sqrt{y_i^{(k)}}, \quad i = 1, \dots, [n/2], \quad (13)$$

where $k = 1$ if n is even and $k = 2$ if n is odd. Note that

$$\begin{aligned} \int_{-a}^a P_{2r}(x, t) P_{2l}(x, t) \omega(x, t) dx &= \int_{-a}^a S_r(x^2, t) S_l(x^2, t) \omega(x, t) dx \\ &= 2 \int_0^a S_r(x^2, t) S_l(x^2, t) \omega(x, t) dx = \int_0^{a^2} S_r(y, t) S_l(y, t) \frac{\omega(\sqrt{y}, t)}{\sqrt{y}} dy \end{aligned}$$

and

$$\begin{aligned} \int_{-a}^a P_{2r+1}(x, t) P_{2l+1}(x, t) \omega(x, t) dx &= \int_{-a}^a x T_r(x^2, t) x T_l(x^2, t) \omega(x, t) dx \\ &= 2 \int_0^a T_r(x^2, t) T_l(x^2, t) x^2 \omega(x, t) dx = \int_0^{a^2} T_r(y, t) T_l(y, t) \sqrt{y} \omega(\sqrt{y}, t) dy. \end{aligned}$$

Since $\{P_n(x, t)\}$ is a sequence of orthogonal polynomials with respect to an even weight function $\omega(x, t)$ on $(-a, a)$, it follows that $\{S_n(y, t)\}$ and $\{T_n(y, t)\}$ are sequences of orthogonal polynomials on $(0, a^2)$ with respect to the weight functions

²In the case that $\omega(x, t)$ is an even function in an interval of the form $(-a, a)$, it is well known that the zeros of the orthogonal polynomials are symmetric with respect to the origin, i.e., $x_k(t) = -x_{n-k+1}(t)$, $k = 1, 2, \dots, n$.

$\omega_1(y, t) = \omega(\sqrt{y}, t)/\sqrt{y}$ and $\omega_2(y, t) = \sqrt{y} \omega(\sqrt{y}, t)$, respectively. Now, it is easy to see that

$$\frac{1}{\omega_1(y, t)} \frac{\partial \omega_1(y, t)}{\partial t} = \frac{1}{\omega_2(y, t)} \frac{\partial \omega_2(y, t)}{\partial t} = \frac{1}{\omega(\sqrt{y}, t)} \frac{\partial \omega(\sqrt{y}, t)}{\partial t}.$$

Therefore, since the function $(\omega(x, t))^{-1} \partial \omega(x, t) / \partial t$ increases (decreases) when x increases in $(0, a)$, then the functions $(\omega_1(y, t))^{-1} \partial \omega_1(y, t) / \partial t$ and $(\omega_2(y, t))^{-1} \partial \omega_2(y, t) / \partial t$ increase (decrease) when y increases in $(0, a^2)$. So it follows from Markov's theorem that the zeros $y_i^{(k)} = y_i^{(k)}(t)$, $i = 1, \dots, [n/2]$, $k = 1, 2$, increase (decrease) when t increases in B . Then, the result follows from (13).

In Markov's theorem, one can consider the end points of the interval of the orthogonality as functions of the parameter, i.e., $a = a(t)$ and $b = b(t)$. From this, the following result can be derived:

Corollary 3.3 *Assume the notation and conditions of Theorem 2 under the constraint that $d\mu(x, t) = \omega(x, t)dx$. Furthermore, suppose that $a = a(t)$ and $b = b(t)$ are functions of t with continuous derivatives of the first order. Then, $x_k(t)$ is a strictly increasing (decreasing) function of t if*

$$\frac{1}{\omega(x, t)} \frac{\partial \omega}{\partial t}(x, t)$$

is an increasing (decreasing) function of $x \in A = (a(t), b(t))$, provided that this last function be nonconstant on A , and both $a(t)$ and $b(t)$ increase (decrease) as t increases.

Proof By Leibniz's rule for differentiation under the integral sign, one obtains that the numerator of the right-hand side of (12) becomes

$$\int_{a(t)}^{b(t)} \frac{[P_n(x, t)]^2}{x - x_k(t)} \left(\frac{1}{\omega(x, t)} \frac{\partial \omega}{\partial t}(x, t) - \frac{1}{\omega(x_k(t), t)} \frac{\partial \omega}{\partial t}(x_k(t), t) \right) \omega(x, t) dx + \frac{P_n^2(b(t), t)}{b(t) - x_k(t)} \omega(b(t), t) b'(t) - \frac{P_n^2(a(t), t)}{a(t) - x_k(t)} \omega(a(t), t) a'(t). \quad (14)$$

This establishes the result.

In Corollary 3.3, the hypothesis that the weight function depends on the parameter t may be replaced by the hypothesis that the weight function does not depend on t , that is, $\omega = \omega(x)$. In this case, if both $a(t)$ and $b(t)$ increase (decrease) with t increases (either a or b can be constant), then $x_k(t)$ is an increasing (decreasing) function of t .

Some particular cases of measures of the form

$$d\mu(x, t) = d\alpha(x) + J(t)\delta_y \quad (15)$$

were frequently considered in the literature (see [13, 20–22, 26, 28]). See [13] for general results concerning zeros of polynomials orthogonal with respect to (15). A bit more general case of (15) is presented as follows:

Corollary 3.4 *Assume the notation and conditions of Theorem 2 under the constraint $d\mu(x, t) = d\alpha(x) + \sum_{i=0}^{\infty} J_i(t)\delta_{y_i}$. Furthermore, suppose that y_i , $i = 0, 1, \dots$, are constants, and $J'_i(t) = 0$ for $i \neq l$. Define the sets*

$$C_l^- := \{t \in B \mid J'_l(t) < 0\}, \quad C_l^+ := \{t \in B \mid J'_l(t) > 0\}.$$

If $x_k(t) < y_l$ (respectively, $x_k(t) > y_l$) for each $t \in B$, then $x_k(t)$ is a strictly increasing (respectively, decreasing) function of t on C_l^+ (respectively, on C_l^-). In other words, each zero $x_k(t)$ on the left-hand side of y_l is an increasing (decreasing) function of t on C_l^+ (C_l^-), whereas each zero $x_k(t)$ on the right-hand side of y_l is a decreasing (increasing) function of t on C_l^+ (C_l^-).

Proof In this case, (9) reduces to

$$\frac{dx_k}{dt}(t) = \frac{J'_l(t) \frac{[P_n(y_l, t)]^2}{y_l - x_k(t)}}{\int \frac{[P_n(x, t)]^2}{(x - x_k(t))^2} \omega(x, t) dv(x) + \sum_{i=0}^{\infty} J_i(t) \frac{[P_n(y_i, t)]^2}{(y_i - x_k(t))^2}}.$$

This establishes the result.

The next result was proved firstly in [3, Theorem 2.2]. In order to derive monotonicity properties of zeros, the location of the mass point outside A is required (see Section 4.2 of this manuscript).

Corollary 3.5 *Assume the notation and conditions of Theorem 2 under the constraint $d\mu(x, t) = d\alpha(x) + J\delta_{y(t)}$. Define the sets³*

$$B_- := \{t \in B \mid y(t) \in Co(A)^c \wedge y'(t) < 0\},$$

$$B_+ := \{t \in B \mid y(t) \in Co(A)^c \wedge y'(t) > 0\}.$$

Then all the zeros of $P_n(x, t)$ are strictly decreasing (respectively, increasing) functions of t on B_- (respectively, on B_+).

³ $A^c := \{x \in \mathbb{R} \mid x \notin A\}$ and $Co(A)$ denotes the convex hull of A .

Proof In this case, (9) reduces to

$$\frac{dx_k}{dt}(t) = \frac{J y'(t) \frac{[P_n(y(t), t)]^2}{y(t) - x_k(t)} \sum_{j=0}^n \frac{2 - \delta_{j,k}}{y(t) - x_j(t)}}{\int \frac{[P_n(x, t)]^2}{(x - x_k(t))^2} \omega(x, t) dv(x) + J \frac{[P_n(y(t), t)]^2}{(y(t) - x_k(t))^2}},$$

where the prime means that the sum is over all values j and t for which $y(t) \neq x_j(t)$. This establishes the result.

4 Some Applications

4.1 Sharp Monotonicity Properties of the Zeros of Orthogonal Polynomials Derived from Corollary 3.4

Suppose that $d\mu(x, t) = dx + J_1\delta_{y_1} + J_2\delta_{y_2} + J_3\delta_{y_3}$, where $J_1 = J_1(t) = t$, $J_2 = J_3 = 1$, $y_1 = 2$, $y_2 = 5$, and $y_3 = 7$, with $A = (-1, 1)$ and $B = (0, \infty)$. Let $\{p_n\}$ be the sequence of orthogonal polynomials with respect to $d\mu$, i.e.,

$$\int_{-1}^1 p_n(x)p_m(x)dx + tp_n(2)p_m(2) + p_n(5)p_m(5) + p_n(7)p_m(7) = 0, \quad m \neq n.$$

Then, the zeros of the polynomial p_n located on the left-hand side of $y_1 = 2$ are increasing functions of t , while the zeros of p_n on the right-hand side of $y_1 = 2$ are decreasing functions of t , in view of Corollary 3.4.

Table 1 shows the monotonicity of the zeros of p_4 from this example. Observe that two of them are increasing functions of t , while the other ones are decreasing functions of t .

One can consider an example with an infinite number of mass points. For instance, the Charlier polynomials $\{C_n(x, a)\}$ are orthogonal with respect to a discrete measure whose distribution function has jumps $\omega(x, a) = a^x/x!$ at $x = 0, 1, \dots$, with $a > 0$ (see [16]), that is,

$$\sum_{x=0}^{\infty} C_m(x, a)C_n(x, a)\omega(x, a) = 0, \quad m \neq n.$$

Let $\{C_n^k(x, a, t)\}$ satisfy the orthogonality relation

$$\sum_{x=0}^{\infty} C_m^k(x, a, t)C_n^k(x, a, t)\omega_k(x, a, t) = 0, \quad m \neq n, \quad k \geq 0,$$

Table 1 Zeros of the polynomial p_4 as functions of t

t	$x_1(t)$	$x_2(t)$	$x_3(t)$	$x_4(t)$
0.0	-0.655077	0.46887	4.98364	6.99699
0.5	-0.528752	1.07504	4.83155	6.97651
1.0	-0.502388	1.33983	4.75758	6.96841
1.5	-0.491188	1.48640	4.71348	6.96408
2.0	-0.485007	1.57951	4.68410	6.96138
2.5	-0.481092	1.64394	4.66310	6.95953
3.0	-0.478390	1.69121	4.64733	6.95820
3.5	-0.476414	1.72736	4.63504	6.95718
4.0	-0.474906	1.75592	4.62520	6.95638
4.5	-0.473717	1.77906	4.61713	6.95574
5.0	-0.472756	1.79818	4.61040	6.95521
5.5	-0.471962	1.81425	4.60470	6.95477
6.0	-0.471297	1.82795	4.59981	6.95439
6.5	-0.470730	1.83977	4.59557	6.95407
7.0	-0.470242	1.85006	4.59185	6.95379
7.5	-0.469817	1.85911	4.58857	6.95354
8.0	-0.469444	1.86713	4.58565	6.95332
8.5	-0.469113	1.87428	4.58304	6.95313
9.0	-0.468819	1.88071	4.58069	6.95295
9.5	-0.468555	1.88651	4.57856	6.95279
10.0	-0.468316	1.89177	4.57662	6.95265

Table 2 Zeros of the polynomial $C_4^2(x, 1, t)$ as functions of t

t	$x_1(t)$	$x_2(t)$	$x_3(t)$	$x_4(t)$
0	0.0439673	1.33203	3.07971	5.54429
1	0.0556639	1.53679	2.81524	5.43756
2	0.0594188	1.63309	2.69647	5.40402
3	0.0612692	1.69166	2.62618	5.38764
4	0.0623708	1.73193	2.57873	5.37794
5	0.0631017	1.76169	2.54410	5.37152
6	0.0636219	1.78479	2.51750	5.36697
7	0.0640111	1.80334	2.49630	5.36357
8	0.0643133	1.81864	2.47894	5.36093
9	0.0645547	1.83150	2.46441	5.35883
10	0.0647519	1.84251	2.45204	5.35711

where $\omega_k(x, a, t) = \omega(x, a)$, for $x \neq k$, and $\omega_k(k, a, t) = \omega(k, a) + t$. Note that $C_n^k(x, a, 0) = C_n(x, a)$. For $n = 4$, the zeros of $C_4(x, 1)$ are $x_1 = 0.0439673$, $x_2 = 1.33203$, $x_3 = 3.07971$, and $x_4 = 5.54429$. So, for instance, if one takes $k = 2$, by Corollary 3.4, the zeros $x_1(t)$ and $x_2(t)$ of $C_4^2(x, 1, t)$ are increasing functions of t , while the zeros $x_3(t)$ and $x_4(t)$ of $C_4^2(x, 1, t)$ are decreasing functions of t , see Table 2.

4.2 Sharp Monotonicity Properties of the Zeros of Orthogonal Polynomials Derived from Corollary 3.5

Suppose that $d\mu(x, t) = dx + 10\delta_{y(t)}$, where $y(t) = t$, with $A = (-1, 1)$ and $B = (-2, 2)$. Let $\{p_n\}$ be the sequence of orthogonal polynomials with respect to $d\mu$, i.e.,

$$\int_{-1}^1 p_n(x)p_m(x)dx + 10p_n(t)p_m(t) = 0, \quad m \neq n.$$

Then, by Corollary 3.5, the zeros of the polynomial p_n are increasing functions of t , for $t \in (-2, -1) \cup (1, 2)$. On the other hand, for $t \in (-1, 1)$, one cannot guarantee the monotonicity of these zeros.

Table 3 illustrates the behavior of the zeros $x_1 = x_1(t)$, $x_2 = x_2(t)$, $x_3 = x_3(t)$, and $x_4 = x_4(t)$, of p_4 from this example. Note that they are not monotonic functions of t , when it varies in $(-1, 1)$. In this regard, the statements of Theorem 2 and Corollary 3 in [8] appear to be incorrect.

Table 3 Zeros of the polynomial p_4 as functions of t

t	$x_1(t)$	$x_2(t)$	$x_3(t)$	$x_4(t)$
-2.0	-1.999850	-0.682492	0.142833	0.818103
-1.8	-1.799760	-0.664794	0.158989	0.821979
-1.6	-1.599570	-0.638982	0.179555	0.826752
-1.4	-1.399200	-0.598464	0.206977	0.832897
-1.2	-1.198540	-0.529080	0.246312	0.841409
-1.0	-0.998028	-0.409306	0.306336	0.854084
-0.8	-0.803116	-0.368260	0.329881	0.859152
-0.6	-0.764270	-0.572827	0.299677	0.853545
-0.4	-0.858748	-0.396400	0.335789	0.860400
-0.2	-0.854976	-0.210785	0.301057	0.856301
0.0	-0.846273	-0.098843	0.098843	0.846273
0.2	-0.856301	-0.301057	0.210785	0.854976
0.4	-0.860400	-0.335789	0.396400	0.858748
0.6	-0.853545	-0.299677	0.572827	0.764270
0.8	-0.859152	-0.329881	0.368260	0.803116
1.0	-0.854084	-0.306336	0.409306	0.998028
1.2	-0.841409	-0.246312	0.529080	1.198540
1.4	-0.832897	-0.206977	0.598464	1.399200
1.6	-0.826752	-0.179555	0.638982	1.599570
1.8	-0.821979	-0.158989	0.664794	1.799760
2.0	-0.818103	-0.142833	0.682492	1.999850

4.3 Monotonicity of the Zeros of Classical Continuous Orthogonal Polynomials Derived from Corollary 3.1 (Markov's Theorem)

In this subsection, the classical result established by Markov in 1886 is reviewed. It concerns the monotonicity of zeros of Jacobi orthogonal polynomials. Moreover, the results on zeros of Gegenbauer and Laguerre orthogonal polynomials are revisited too (see Szegő's book [32, Section 5], and Ismail's book [16, Chapter 7]).

Example 4.1 (Zeros of Jacobi Polynomials) Let $P_n^{(\alpha,\beta)}(x)$ be the n th Jacobi polynomial that is orthogonal on $(-1, 1)$ with respect to the weight function $\omega(x, \alpha, \beta) = (1 - x)^\alpha(1 + x)^\beta$, $\alpha, \beta > -1$. Then all its zeros are increasing functions of β and decreasing functions of α , for $\alpha, \beta > -1$.

Proof Since

$$\frac{1}{\omega(x, \alpha, \beta)} \frac{\partial \omega(x, \alpha, \beta)}{\partial \alpha} = \ln(1 - x)$$

is a decreasing function of x and, otherwise,

$$\frac{1}{\omega(x, \alpha, \beta)} \frac{\partial \omega(x, \alpha, \beta)}{\partial \beta} = \ln(1 + x)$$

is an increasing function of x , for $x \in (-1, 1)$, from Markov's theorem, the statements hold.

Figure 1 illustrates the monotonicity of the zeros of $P_n^{(\alpha,\beta)}(x)$ with respect to β , while Figure 2 shows the monotonicity of the zeros of $P_n^{(\alpha,\beta)}(x)$ with respect to α .

Example 4.2 (Zeros of Laguerre Polynomials) Let $L_n^{(\alpha)}(x)$ be the n th Laguerre polynomial that is orthogonal on $(0, \infty)$ with respect to the weight function $\omega(x, \alpha) = x^\alpha e^{-x}$, $\alpha > -1$. Then, all its zeros are increasing functions of α , for $\alpha > -1$.

Proof In this case,

Fig. 1 Zeros of Jacobi polynomials as functions of the parameter β . Graph of the zeros of $P_4^{(1,\beta)}(x)$

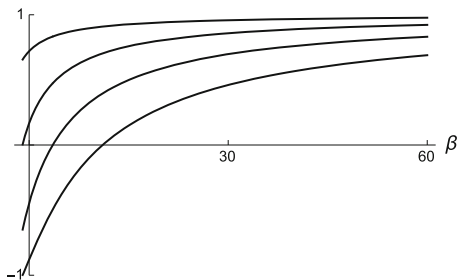


Fig. 2 Zeros of Jacobi polynomials as functions of the parameter α . Graph of the zeros of $P_4^{(\alpha,1)}(x)$

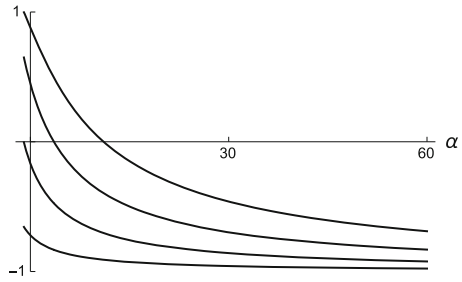
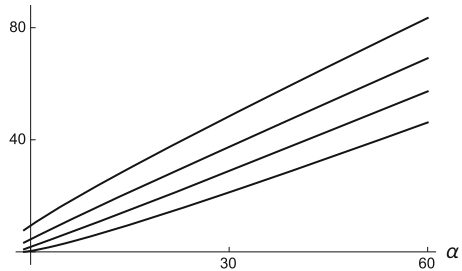


Fig. 3 Zeros of Laguerre polynomials as functions of α . Graph of the zeros of $L_4^{(\alpha)}(x)$



$$\frac{1}{\omega(x, \alpha)} \frac{\partial \omega(x, \alpha)}{\partial \alpha} = \ln x$$

is an increasing function of x , for $x \in (0, \infty)$. So, from Markov’s theorem, the statement holds.

Figure 3 serves to illustrate the monotonicity of the zeros of Laguerre polynomials with respect to the parameter α .

Example 4.3 (Zeros of Gegenbauer Polynomials) Let $P_n^{(\lambda)}(x)$ be the n th Gegenbauer (or ultraspherical) polynomial that is orthogonal on $(-1, 1)$ with respect to the weight function $\omega(x, \lambda) = (1 - x^2)^{\lambda-1/2}$, $\lambda > -1/2$. Then, all its positive zeros are decreasing functions of λ , for $\lambda > -1/2$.⁴

Proof Since $\omega(x, \lambda)$ is an even function and

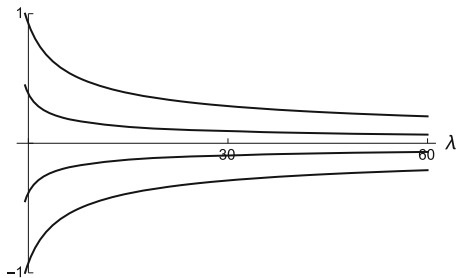
$$\frac{1}{\omega(x, \lambda)} \frac{\partial \omega(x, \lambda)}{\partial \lambda} = \ln(1 - x^2)$$

is a decreasing function of x , for $x \in (0, 1)$, then, from Markov’s theorem for even weight function, the statement holds.

Figure 4 shows the behavior of the zeros of $P_4^{(\lambda)}(x)$ as functions of λ .

⁴Because of the symmetry of the zeros of $P_n^{(\lambda)}(x)$, its negative zeros are increasing functions of λ , for $\lambda > -1/2$.

Fig. 4 Zeros of Gegenbauer polynomials as functions of λ . Graph of the zeros of $P_4^{(\lambda)}(x)$



4.4 Sharp Monotonicity Properties of the Zeros of Classical Continuous Orthogonal Polynomials Derived from Corollary 3.3

The motivation of the next result goes back to a work of Laforgia [24], raised in 1981. He proved that the quantities $\lambda x_{n,k}(\lambda)$ are increasing functions of λ , for $\lambda \in (0, 1)$, where $x_{n,k}(\lambda)$, $k = 1, \dots, [n/2]$, are the positive zeros of Gegenbauer polynomial $P_n^{(\lambda)}(x)$. Later on, Laforgia [25] conjectured that this result remains valid for $\lambda \in (0, \infty)$. In 1988, Ismail and Letessier [17] conjectured that $(\lambda + c)^{\frac{1}{2}} x_{n,k}(\lambda)$, $k = 1, \dots, [n/2]$, increase with $\lambda > 0$ for $c = 0$. Ismail in [15, Conjecture 3, p. 188], following a suggestion of Askey, conjectured this result for $c = 1$, leading up to the ILA conjecture of the title. Ifantis and Siafarikas [14] showed ILA for $k = 1$ and $\lambda > -1/2$, as well as in [7]. Ahmed, Muldoon, and Spigler [1] proved this monotonicity result for $c = (2n^2 + 1)/(4n + 2)$ and $-1/2 < \lambda \leq 3/2$. Elbert and Siafarikas [11] extended the result of Ahmed et al., showing thus ILA for all $\lambda > -1/2$.

Next, using one of the Markov’s descendants’ results, one can prove the following statement related to ILA conjecture.

OBSERVATION 1 (GEGENBAUER–HERMITE) *Let $x_{n,1}(\lambda) > \dots > x_{n,n}(\lambda)$ be the zeros of the Gegenbauer polynomial $P_n^{(\lambda)}(x)$ and let $h_{n,1} > \dots > h_{n,n}$ be the zeros of the Hermite polynomial $H_n(x)$. Then, for all $n \in \mathbb{N}$ and $c \leq -1/2$, the quantities*

$$(\lambda + c)^{\frac{1}{2}} x_{n,k}(\lambda), \quad k = 1, \dots, [n/2],$$

are increasing functions of λ and converge to $h_{n,k}$ when λ goes to infinity. So, for $c = -1/2$, one obtains

$$x_{n,k}(\lambda) \leq (\lambda - 1/2)^{-\frac{1}{2}} h_{n,k}, \quad k = 1, \dots, [n/2].$$

Proof One can assert the asymptotic formula [33, Section 5.6] (see also [19, formula (2.8.3)])

$$\lim_{\lambda \rightarrow \infty} \lambda^{-n/2} P_n^{(\lambda)}(\lambda^{-\frac{1}{2}}x) = \frac{H_n(x)}{n!},$$

where $H_n(x)$ denotes the n th Hermite orthogonal polynomial. Let $h_{n,k}$, $k = 1, \dots, n$, be the zeros of $H_n(x)$ arranged in decreasing order. Thus, that gives

$$\lim_{\lambda \rightarrow \infty} \lambda^{\frac{1}{2}} x_{n,k}(\lambda) = h_{n,k}.$$

Therefore, for $f = f_n(\lambda) = (\lambda + c)^{\frac{1}{2}}$, where c is a constant that may depend on n but does not depend on λ , equivalently that gives

$$\lim_{\lambda \rightarrow \infty} (\lambda + c)^{\frac{1}{2}} x_{n,k}(\lambda) = h_{n,k}.$$

Hence, a natural question arises: is there a value of c such that all the quantities $(\lambda + c)^{\frac{1}{2}} x_{n,k}(\lambda)$, $k = 1, \dots, [n/2]$, are monotonic (increasing or decreasing) functions of λ ? To answer this question, one has to perform the exchange of variables $x = z/f$ to obtain the rescaled Gegenbauer polynomial $P_n^{(\lambda)}(z/f)$ orthogonal on $(-f, f)$ with respect to the weight function $\omega(z, \lambda) = (f^2 - z^2)^{\lambda-1/2}$, $\lambda > -1/2$, and whose zeros are $z_{n,k}(\lambda) = f_n(\lambda)x_{n,k}(\lambda)$. Then, a straightforward calculation yields

$$\frac{\partial f}{\partial \lambda} = \frac{1}{2(\lambda + c)^{\frac{1}{2}}} > 0$$

and

$$\frac{\partial}{\partial z} \left[\frac{1}{\omega(z, \lambda)} \frac{\partial \omega(z, \lambda)}{\partial \lambda} \right] = \frac{2z[(z^2 - f^2) + (2\lambda - 1)f \partial f / \partial \lambda]}{(f^2 - z^2)^2} > 0,$$

for $z \in (0, f)$ and $c \leq -1/2$. Therefore, having Corollaries 3.2 and 3.3 in mind, for $c \leq -1/2$, the quantities $(\lambda + c)^{\frac{1}{2}} x_{n,k}(\lambda)$, $k = 1, \dots, [n/2]$, are increasing functions of λ and converge to $h_{n,k}$ when λ goes to infinity. Therefore, for $c = -1/2$, one obtains

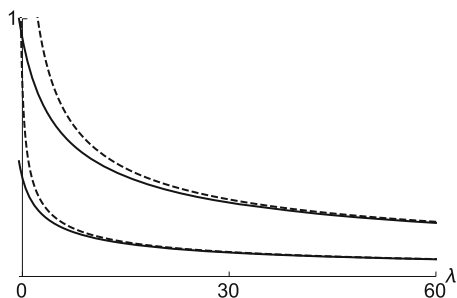
$$(\lambda - 1/2)^{\frac{1}{2}} x_{n,k}(\lambda) \leq h_{n,k}, \quad k = 1, \dots, [n/2],$$

or equivalently

$$x_{n,k}(\lambda) \leq (\lambda - 1/2)^{-\frac{1}{2}} h_{n,k}, \quad k = 1, \dots, [n/2].$$

The right-hand side of the above inequalities are upper bounds for the positive zeros of Gegenbauer polynomials, and they are sharp for large values of λ . See Figure 5.

Fig. 5 Graph of the zeros $x_{4,k}(\lambda)$ (continuous lines) and their upper bounds $(\lambda - 1/2)^{-\frac{1}{2}}h_{4,k}$ (dashed lines), for $k = 1, 2$



The next example describes a connection between the zeros of Jacobi and Laguerre orthogonal polynomials. In [9], using Sturm's comparison theorem on solutions of Sturm–Liouville differential equation, it was shown the monotonicity results for the functions $(\beta + c)(1 - x_{n,k}(\alpha, \beta))$, $k = 1, \dots, n$, where $c = n + (\alpha + 1)/2 + (1 - \alpha^2)/(4n + 2\alpha + 2)$.

OBSERVATION 2 (JACOBI–LAGUERRE) *Let $x_{n,1}(\alpha, \beta) > \dots > x_{n,n}(\alpha, \beta)$ be the zeros of $P_n^{(\alpha, \beta)}(x)$ and let $\ell_{n,1}(\alpha) > \dots > \ell_{n,n}(\alpha)$ be the zeros of $L_n^{(\alpha)}(x)$. Then, for every $n \in \mathbb{N}$, $1 \leq k \leq n$, $\alpha > -1$, and $c \leq 0$, the quantities*

$$(\beta + c)(1 - x_{n,k}(\alpha, \beta))/2$$

are increasing functions of β , for $\beta \in (-1, \infty)$, and converge to $\ell_{n,n-k+1}(\alpha)$ when β goes to infinity. Moreover, the inequalities

$$1 - \frac{2}{\beta}\ell_{n,n-k+1}(\alpha) \leq x_{n,k}(\alpha, \beta)$$

hold.

Proof One can find in the literature [33, Section 5.3] (see also [19, formula (2.8.1)]) the following limit relation between Jacobi and Laguerre polynomials

$$\lim_{\beta \rightarrow \infty} P_n^{(\alpha, \beta)}(1 - 2\beta^{-1}x) = L_n^{(\alpha)}(x).$$

Since the zeros are continuous functions of the coefficients of the polynomials, one derives

$$\lim_{\beta \rightarrow \infty} \beta(1 - x_{n,k}(\alpha, \beta)) = 2\ell_{n,n-k+1}(\alpha), \quad k = 1, \dots, n.$$

Therefore, for $f = f_n(\alpha, \beta) = \beta + c$, where c is a constant that may depend on n and α but does not depend on β , one has equivalently that

$$\lim_{\beta \rightarrow \infty} (\beta + c)(1 - x_{n,k}(\alpha, \beta)) = 2\ell_{n,n-k+1}(\alpha), \quad k = 1, \dots, n.$$

The purpose is to find the function f in such a way that all the quantities $f(1 - x_{n,k}(\alpha, \beta))$ are monotonic (increasing or decreasing) functions of β . For this reason, performing the change of variables $x = 1 - 2z/f$ one obtains that the rescaled Jacobi polynomials $P_n^{(\alpha,\beta)}(1 - 2z/f)$, whose zeros are $z_{n,k}(\alpha, \beta) = f(1 - x_{n,k}(\alpha, \beta))/2$, $k = 1, \dots, n$, are orthogonal on $(0, f)$ with respect to the weight function $\omega(z, \alpha, \beta) = z^\alpha (f - z)^\beta$, for $\alpha, \beta > -1$.

In order to apply the Corollary 3.3 for $z \in (0, f)$, one has to calculate the following derivatives:

$$\frac{\partial f}{\partial \beta} = 1 > 0$$

and

$$\frac{\partial}{\partial z} \left[\frac{1}{\omega(z, \alpha, \beta)} \frac{\partial \omega(z, \alpha, \beta)}{\partial \beta} \right] = \frac{z - f + \beta \partial f / \partial \beta}{(f - z)^2} > 0 \quad \text{if and only if } c \leq 0.$$

Therefore, for $c \leq 0$, the quantities $(\beta + c)(1 - x_{n,k}(\alpha, \beta))/2$ are increasing functions of β and converge to $\ell_{n,n-k+1}(\alpha)$ when β goes to infinity. Thus, for $c = 0$, one obtains

$$\beta(1 - x_{n,k}(\alpha, \beta)) \leq 2\ell_{n,n-k+1}(\alpha), \quad k = 1, \dots, n,$$

or equivalently

$$1 - \frac{2}{\beta} \ell_{n,n-k+1}(\alpha) \leq x_{n,k}(\alpha, \beta), \quad k = 1, \dots, n.$$

This establishes the theorem.

Note that the left-hand sides of the above inequalities are lower bounds for the zeros of Jacobi polynomials, and they are sharp for large values of β . See Figures 6 and 7.

Fig. 6 Graph of $z_{4,k}(\alpha, \beta) = \beta(1 - x_{4,k}(\alpha, \beta))/2$, $1 \leq k \leq 4$ (continuous lines). Observe that each $z_{4,k}(\alpha, \beta)$ is an increasing function of β and goes to $\ell_{4,n-k+1}(\alpha)$ as $\beta \rightarrow \infty$ (dotted lines)

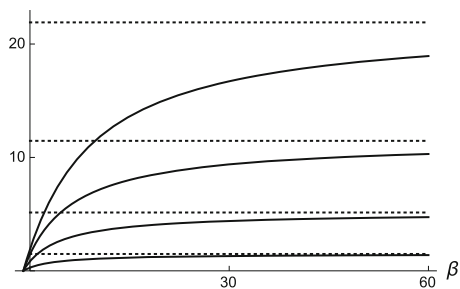
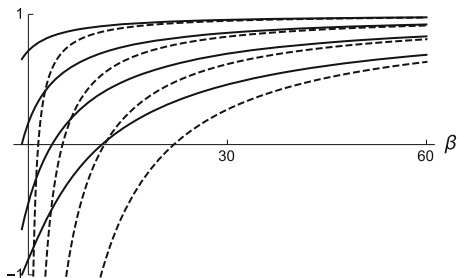


Fig. 7 Graph of the zeros $x_{n,k}(\alpha, \beta)$ (continuous lines) and their lower bounds $1 - 2\ell_{n,n-k+1}(\alpha)/\beta$ (dashed lines), $1 \leq k \leq n$, for the case $n = 4$ and $\alpha = 1$



The last example of this subsection shows the connection between the zeros of Laguerre and Hermite orthogonal polynomials. The first result in this topic was obtained in 1995 by Ifantis and Siafarikas [14]. They showed that $\ell_{n,1}(\alpha)/(\alpha + 1)$ decreases with α , for $\alpha > -1$. In 2003, Natalini and Palumbo [29] proved that $\ell_{n,k}(\alpha)/\sqrt{\alpha + 2n + 1}$ are increasing functions of α , for $\alpha \in (-1, \infty)$. Moreover, they established two additional results on monotonicity of the functions of the form $\ell_{n,k}(\alpha)/\alpha^p$, with fix p , and $2 \leq p \leq 2n + 1$. It was shown in [10] that $[\ell_{n,k}(\alpha) - (2n + \alpha - 1)]/\sqrt{2(n + \alpha - 1)}$ are increasing functions of α , for $\alpha \geq -1/(n - 1)$. In addition, when $k = 1$, it was shown that the last function increases for every $\alpha \in (-1, \infty)$.

OBSERVATION 3 (LAGUERRE–HERMITE) *Let $\ell_{n,1}(\alpha) > \dots > \ell_{n,n}(\alpha)$ be the zeros of $L_n^{(\alpha)}(x)$ and let $h_{n,1} > \dots > h_{n,n}$ be the zeros of $H_n(x)$. Then, for all $n \in \mathbb{N}$ and $1 \leq k \leq n$, the quantities*

$$\frac{\ell_{n,k}(\alpha) - \alpha}{\sqrt{\alpha}}$$

are decreasing functions of α , for $\alpha > 0$, and, moreover, they converge to $\sqrt{2}h_{n,k}$ as $\alpha \rightarrow \infty$. In addition, the inequalities

$$\ell_{n,k}(\alpha) \geq \alpha + \sqrt{2\alpha}h_{n,k}$$

hold for all $\alpha > 0$.

Proof Remember the limit relation between Laguerre and Hermite polynomials (see [19, formula (2.11.1)])

$$\lim_{\alpha \rightarrow \infty} \left(\frac{2}{\alpha}\right)^{n/2} L_n^{(\alpha)}(\alpha + (2\alpha)^{1/2}x) = \frac{(-1)^n}{n!} H_n(x).$$

Whence it follows that

$$\frac{\ell_{n,k}(\alpha) - \alpha}{\sqrt{2\alpha}} \rightarrow h_{n,k} \text{ as } \alpha \rightarrow \infty.$$

Therefore, for any constants c and d that may depend on n but does not depend on α , one can write

$$\frac{\ell_{n,k}(\alpha) - (\alpha + c)}{\sqrt{\alpha + d}} \rightarrow \sqrt{2}h_{n,k} \text{ as } \alpha \rightarrow \infty.$$

To obtain sharp bounds for the zeros of Laguerre polynomials, one needs to determine, if possible, the best constants c and d for which the quantities $[\ell_{n,k}(\alpha) - (\alpha + c)]/\sqrt{\alpha + d}$ are monotonic (increasing or decreasing) functions of α . The best constants mean the infimum or supremum of their values. To go in this direction, one has to perform the change of variables $x = \sqrt{\alpha + d}z + \alpha + c$ to obtain the rescaled Laguerre polynomial $L_n^{(\alpha)}(\sqrt{\alpha + d}z + \alpha + c)$ that is orthogonal on $(-\alpha + c)/\sqrt{\alpha + d}, +\infty)$ with respect to the weight function $\omega(z, \alpha) = (\alpha + c + \sqrt{\alpha + d}z)^\alpha \cdot e^{-(\alpha + c + \sqrt{\alpha + d}z)}$, $\alpha > -1$, and whose zeros are $z_{n,k}(\alpha) = [\ell_{n,k}(\alpha) - (\alpha + c)]/\sqrt{\alpha + d}$, $1 \leq j \leq n$. Now to apply Corollary 3.3 for $z \in (-\alpha + c)/\sqrt{\alpha + d}, +\infty)$, one has to calculate the following derivatives:

$$\frac{\partial}{\partial \alpha} \left[-\frac{\alpha + c}{\sqrt{\alpha + d}} \right] = -\frac{\alpha + 2d - c}{2(\alpha + d)^{\frac{3}{2}}} \tag{16}$$

and

$$\frac{\partial}{\partial z} \left[\frac{1}{\omega(z, \alpha)} \frac{\partial \omega(z, \alpha)}{\partial \alpha} \right] = \frac{c(\alpha + 2d - c) + 2\sqrt{\alpha + d}(d - c)z - (\alpha + d)z^2}{2\sqrt{\alpha + d}(\alpha + c + \sqrt{\alpha + d}z)^2}. \tag{17}$$

Taking $c = d = 0$, (16) and (17) become negative for every $\alpha > 0$. Then, Corollary 3.3 implies that all the quantities

$$z_{n,k}(\alpha) = \frac{\ell_{n,k}(\alpha) - \alpha}{\sqrt{\alpha}}, \quad 1 \leq j \leq n,$$

are decreasing functions of α , for $\alpha > 0$. It was provided that $z_{n,k}(\alpha)$ goes to $\sqrt{2}h_{n,k}$ as $\alpha \rightarrow \infty$, so $\ell_{n,k}(\alpha) \geq \alpha + \sqrt{2\alpha}h_{n,k}$ for all $\alpha > 0$ and $k = 1, \dots, n$.

Figure 8 exemplifies the behavior of the zeros $z_{n,k}(\alpha)$ with respect to the parameter α for $c = d = 0$. Figure 9 shows the lower bounds for the zeros $\ell_{n,k}(\alpha)$ of $L_n^{(\alpha)}(x)$.

Fig. 8 Graph of $z_{n,k}(\alpha) = \frac{\ell_{n,k}(\alpha) - \alpha}{\sqrt{2\alpha}}$, $1 \leq k \leq n$, for $n = 4$ (continuous lines). Observe that each $z_{4,k}(\alpha)$ is a decreasing function of α and goes to $h_{4,k}$ as $\alpha \rightarrow \infty$ (dotted lines)

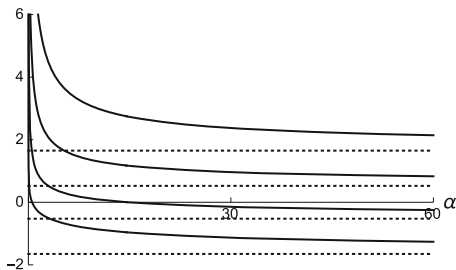
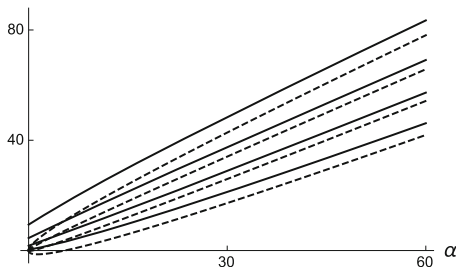


Fig. 9 Graph of the zeros $\ell_{n,k}(\alpha)$ (continuous lines) and their lower bounds $\alpha + \sqrt{2\alpha}h_{n,k}$ (dashed lines), $1 \leq j \leq n$, for the case $n = 4$



4.5 Monotonicity of the Zeros of Classical Discrete Orthogonal Polynomials Derived from Corollary 3.1 (Markov's Theorem)

In this subsection, the monotonicity of the zeros of the families of classical orthogonal polynomials of a discrete variable, Charlier, Meixner, Kravchuk, and Hahn is revisited. Such results can be found in Ismail's book [16, Chapter 7] (see also [2]). For further information to this class of polynomials, see [30].

Example 4.4 Let $C_n^{(a)}(x)$ be the n th Charlier orthogonal polynomial. Then all its zeros are increasing functions of a , for $a \in (0, \infty)$.

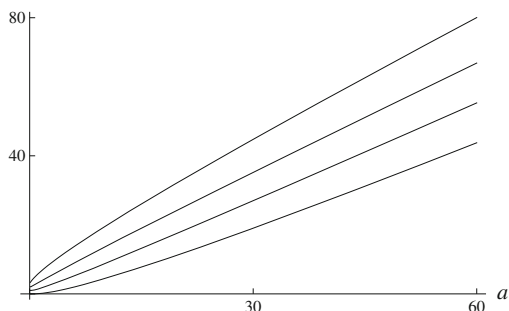
Proof The Charlier polynomials are orthogonal with respect to $\omega(x, a) = a^x / \Gamma(x + 1)$ at $x = 0, 1, 2, \dots$. Let us consider its continuous extension on $(0, \infty)$. Then

$$\frac{1}{\omega(x, a)} \frac{\partial \omega(x, a)}{\partial a} = \frac{\partial}{\partial a} \left[\ln \frac{a^x}{\Gamma(x + 1)} \right] = \frac{\partial}{\partial a} [x \ln(a) + \ln(\Gamma(x + 1))] = \frac{x}{a}$$

is an increasing function of x , for $x \in (0, \infty)$. Thus, from Markov's theorem, one concludes that the zeros of $C_n^{(a)}(x)$ are increasing function of a , for $a > 0$.

Figure 10 presents the graph of the zeros of $C_4^{(a)}(x)$ as functions of the parameter a . Note that its zeros are increasing functions of a , for $a > 0$.

Fig. 10 Zeros of Charlier polynomials as functions of the parameter a . Graph of the zeros of $C_4^{(a)}(x)$



Example 4.5 Let $M_n^{(\beta,c)}(x)$ be the n th Meixner orthogonal polynomial. Then all its zeros are increasing functions of both $\beta \in (0, \infty)$ and $c \in (0, 1)$.

Proof The Meixner polynomials are orthogonal with respect to $\omega(x, \beta, c) = \Gamma(x + \beta)c^x / (\Gamma(\beta)\Gamma(x + 1))$ at $x = 0, 1, 2, \dots$. To prove the monotonicity of the zeros of $M_n^{(\beta,c)}(x)$ with respect to the parameters β and c , one has to consider the analytic extension of $\omega(x, \beta, c) = \Gamma(x + \beta)c^x / (\Gamma(\beta)\Gamma(x + 1))$ on $(0, \infty)$. Therefore,

$$\begin{aligned} \ln \omega(x, \beta, c) &= \ln \frac{\Gamma(x + \beta)c^x}{\Gamma(\beta)\Gamma(x + 1)} \\ &= \ln \Gamma(x + \beta) + x \ln c - \ln \Gamma(\beta) - \ln \Gamma(x + 1). \end{aligned} \tag{18}$$

Computing the derivative of (18) with respect to β , one obtains⁵

$$\begin{aligned} \frac{1}{\omega(x, \beta, c)} \frac{\partial \omega(x, \beta, c)}{\partial \beta} &= \frac{\partial \ln \omega(x, \beta, c)}{\partial \beta} = \frac{\Gamma'(x + \beta)}{\Gamma(x + \beta)} - \frac{\Gamma'(\beta)}{\Gamma(\beta)} \\ &= \frac{x}{\beta(x + \beta)} + \sum_{n=1}^{\infty} \frac{x}{(\beta + n)(x + \beta + n)}, \end{aligned}$$

which is an increasing function of x for $x \in (0, \infty)$, and $\beta > 0$. Thus, from Markov's theorem, it implies that the zeros of $M_n^{(\beta,c)}(x)$ are increasing functions of β , for $\beta > 0$.

On the other hand, differentiating (18) with respect to c , one obtains

⁵One has the identity $\frac{\Gamma'(z)}{\Gamma(z)} = -\gamma - \frac{1}{z} - \sum_{n=1}^{\infty} \left[\frac{1}{z+n} - \frac{1}{n} \right]$, where γ is the Euler constant (see [34, Section 12.3], [31, Chapter 7]).

Fig. 11 Zeros of Meixner polynomials as functions of the parameter β . Graph of the zeros of $M_5^{(\beta, 0.4)}(x)$, $\beta > 0$

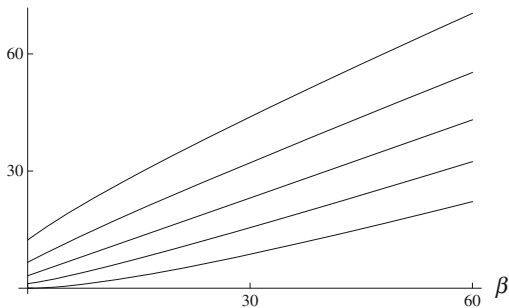
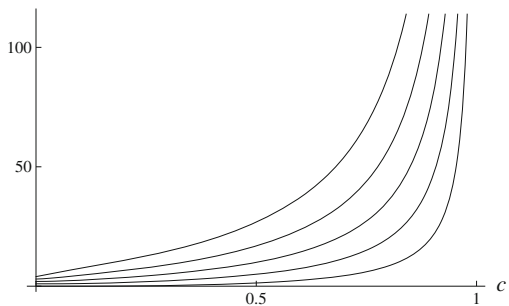


Fig. 12 Zeros of Meixner polynomials as functions of the parameter c . Graph of the zeros of $M_5^{(6, c)}(x)$, $0 < c < 1$



$$\frac{1}{\omega(x, \beta, c)} \frac{\partial \omega(x, \beta, c)}{\partial c} = \frac{\partial \ln \omega(x, \beta, c)}{\partial c} = \frac{x}{c},$$

which is an increasing function of x for $x \in (0, \infty)$, and $c \in (0, 1)$. Hence, from Markov’s theorem, one concludes that the zeros of $M_n^{(\beta, c)}(x)$ are also increasing functions of c , for $c \in (0, 1)$.

To exemplify the monotonicity of the zeros of Meixner polynomials as functions of β and c , one presents two graphs. See Figures 11 and 12.

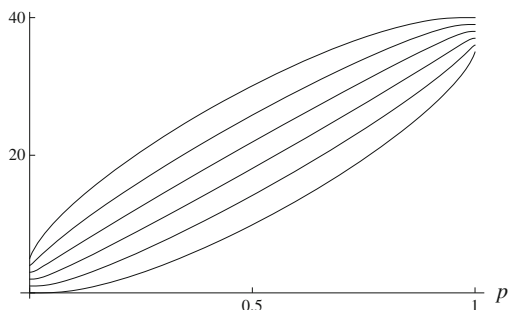
Example 4.6 Let $K_n^{(p, N)}(x)$ be the n th Kravchuk orthogonal polynomial. Then, all its zeros are increasing functions of the parameter p , for $p \in (0, 1)$.

Proof The Kravchuk polynomials are orthogonal with respect to

$$\omega(x, p, N) = \frac{\Gamma(N + 1)p^x(1 - p)^{N-x}}{\Gamma(N + 1 - x)\Gamma(x + 1)}$$

at $x = 0, 1, 2, \dots, N$. Let $\omega(x, p, N)$ be the analytic extension on $(0, N)$ of the Kravchuk weight. Computing the logarithmic derivative of $\omega(x, p, N)$ with respect to p , one obtains

Fig. 13 Zeros of Kravchuk polynomials as functions of the parameter p . Graph of the zeros $K_6^{(p,40)}$, $0 < p < 1$



$$\begin{aligned} \frac{1}{\omega(x, p, N)} \frac{\partial \omega(x, p, N)}{\partial p} &= \frac{\partial}{\partial p} \left[\ln \frac{\Gamma(N+1)p^x(1-p)^{N-x}}{\Gamma(N+1-x)\Gamma(x+1)} \right] \\ &= \frac{\partial}{\partial p} [\ln(\Gamma(N+1)) + x \ln(p) + (N-x) \ln(1-p) - \ln(\Gamma(N+1-x)) \\ &\quad - \ln(\Gamma(x+1))] = \frac{x}{p} - \frac{N-x}{1-p} = \frac{x-Np}{p(1-p)}, \end{aligned}$$

which is obviously an increasing function of x , for $x \in (0, N)$, and $p \in (0, 1)$. Then, from Markov's theorem, all the zeros of $K_n^{(p,N)}(x)$ increase when p increases. See Figure 13.

Example 4.7 Let $P_n^{(\alpha,\beta,N)}(x)$ be the n th Hahn orthogonal polynomial. Then, all its zeros are increasing functions of $\alpha \in (-1, \infty)$ and decreasing functions of $\beta \in (-1, \infty)$.

Proof The Hahn polynomials are orthogonal with respect to

$$\omega(x, \alpha, \beta, N) = \frac{\Gamma(\alpha+x+1)\Gamma(\beta+N-x+1)}{\Gamma(\alpha+1)\Gamma(x+1)\Gamma(\beta+1)\Gamma(N-x+1)}$$

at $x = 0, 1, 2, \dots, N$. Let $\omega(x, \alpha, \beta, N)$ be the analytic extension on $(0, N)$ of the Hahn weight. Then,

$$\begin{aligned} \ln \omega(x, \alpha, \beta, N) &= \ln \frac{\Gamma(\alpha+x+1)\Gamma(\beta+N-x+1)}{\Gamma(\alpha+1)\Gamma(x+1)\Gamma(\beta+1)\Gamma(N-x+1)} \\ &= \ln \Gamma(\alpha+x+1) + \ln \Gamma(\beta+N-x+1) - \ln \Gamma(\alpha+1) \\ &\quad - \ln \Gamma(x+1) - \ln \Gamma(\beta+1) - \ln \Gamma(N-x+1). \end{aligned}$$

Since

$$\begin{aligned} \frac{1}{\omega(x, \alpha, \beta, N)} \frac{\partial \omega(x, \alpha, \beta, N)}{\partial \alpha} &= \frac{\partial \ln \omega(x, \alpha, \beta, N)}{\partial \alpha} = \frac{\Gamma'(\alpha + x + 1)}{\Gamma(\alpha + x + 1)} \frac{\Gamma'(\alpha + 1)}{\Gamma(\alpha + 1)} \\ &= \frac{x}{(\alpha + 1)(x + \alpha + 1)} + \sum_{n=1}^{\infty} \frac{x}{(\alpha + n + 1)(x + \alpha + n + 1)} \end{aligned}$$

is an increasing function of x for $x \in (0, N)$, from Markov's theorem, one derives that the zeros of $P_n^{(\alpha, \beta, N)}(x)$ are increasing functions of α , for $\alpha \in (-1, \infty)$. On the other hand,

$$\begin{aligned} \frac{1}{\omega(x, \alpha, \beta, N)} \frac{\partial \omega(x, \alpha, \beta, N)}{\partial \beta} &= \frac{\partial \ln \omega(x, \alpha, \beta, N)}{\partial \beta} \\ &= \frac{\Gamma'(\beta + N - x + 1)}{\Gamma(\beta + N - x + 1)} - \frac{\Gamma'(\beta + 1)}{\Gamma(\beta + 1)} \\ &= \frac{N - x}{(\beta + 1)(\beta + N - x + 1)} + \sum_{n=1}^{\infty} \frac{N - x}{(\beta + n + 1)(\beta + N - x + 1 + n)} \end{aligned}$$

is a decreasing function of x for $x \in (0, N)$. Thus, from Markov's theorem, it implies that the zeros of $P_n^{(\alpha, \beta, N)}(x)$ are decreasing functions of β , for $\beta \in (-1, \infty)$.

Figures 14 and 15 illustrate these monotonicities.

Fig. 14 Monotonicity of zeros of Hahn polynomials. Graphic of the zeros of $P_4^{(\alpha, 3, 40)}(x)$ as functions of α

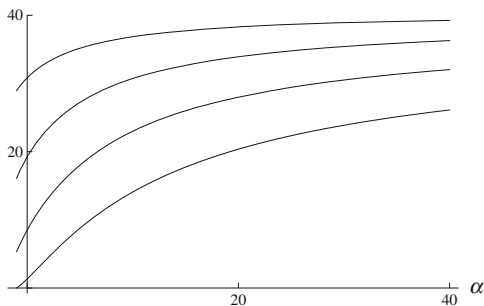
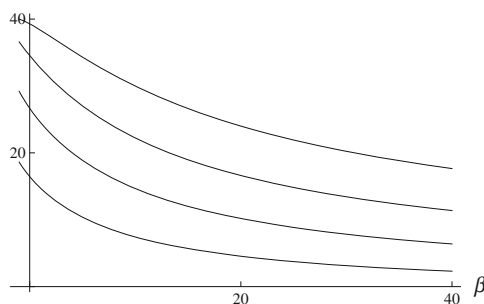


Fig. 15 Monotonicity of zeros of Hahn polynomials. Graphic of the zeros of $P_4^{(8, \beta, 40)}(x)$ as functions of β



Acknowledgments Castillo is supported by the Center for Mathematics of the University of Coimbra, Grant No. UIDB/00324/2020, funded by the Portuguese Government through FCT/MCTES. Rafaeli is supported by the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) Demanda Universal under Grant No. APQ-03782-18 and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

1. S. Ahmed, M.E. Muldoon, R. Spigler, Inequalities and numerical bound for zeros of ultraspherical polynomials. *SIAM J. Math. Anal.* **17**, 1000–1007 (1986)
2. I. Area, D.K. Dimitrov, E. Godoy, V. Paschoa, Zeros of classical orthogonal polynomials of a discrete variable. *Math. Comput.* **82**, 1069–1095 (2013)
3. K. Castillo, F.R. Rafaeli, On the discrete extension of Markov's theorem on monotonicity of zeros. *Electron. Trans. Numer. Anal.* **44**, 271–280 (2015)
4. K. Castillo, M.S. Costa, F.R. Rafaeli, On zeros of polynomials in best L_p -approximation and inserting mass points (2017). arXiv:1706.06295
5. K. Castillo, M.S. Costa, F.R. Rafaeli, On Markov's theorem on zeros of orthogonal polynomials revisited. *Appl. Math. Comput.* **339**, 390–397 (2018)
6. T.S. Chihara, *An Introduction to Orthogonal Polynomials* (Gordon and Breach, New York, 1978)
7. D.K. Dimitrov, On a conjecture concerning monotonicity of zeros of ultraspherical polynomials. *J. Approx. Theory* **85**, 88–97 (1996)
8. D.K. Dimitrov, Monotonicity of zeros of polynomials orthogonal with respect to a discrete measure (2015). arXiv:1501.07235
9. D.K. Dimitrov, F.R. Rafaeli, Monotonicity of zeros of Jacobi polynomials. *J. Approx. Theory* **149**, 15–29 (2007)
10. D.K. Dimitrov, F.R. Rafaeli, Monotonicity of zeros of Laguerre polynomials. *J. Comput. Appl. Math.* **233**, 699–702 (2009)
11. A. Elbert, P.D. Siafarikas, Monotonicity properties of the zeros of ultraspherical polynomials. *J. Approx. Theory* **97**, 31–39 (1999)
12. G. Freud, *Orthogonal Polynomials* (Pergamon Press, Oxford, 1971)
13. E.J. Huertas, F. Marcellán, F.R. Rafaeli, Zeros of orthogonal polynomials generated by canonical perturbations of measures. *Appl. Math. Comput.* **218**, 7109–7127 (2012)
14. E.K. Ifantis, P.D. Siafarikas, Differential inequalities and monotonicity properties of the zeros of associated Laguerre and Hermite polynomials. *Ann. Numer. Math.* **2**, 1–4, 79–91 (1995)
15. M.E.H. Ismail, Monotonicity of zeros of orthogonal polynomials. *q-Series and Partitions*, ed. by D. Stanton (Springer, New York, 1989), pp. 177–190
16. M.E.H. Ismail, *Classical and Quantum Orthogonal Polynomials in One Variable*. *Encyclopedia of Mathematics and Its Applications*, vol. 98 (Cambridge University Press, Cambridge, 2005)
17. M.E.H. Ismail, J. Letessier, Monotonicity of zeros of ultraspherical polynomials, in *Orthogonal Polynomials and Their Applications*, ed. by M. Alfaro, J.S. Dehesa, F.J. Marcellán, J.L. Rubio de Francia, J. Vinuesa. *Lecture Notes in Mathematics*, vol. 1329 (Springer, Berlin, 1988), pp. 329–330
18. K. Jordaan, H. Wang, J. Zhou, Monotonicity of zeros of polynomials orthogonal with respect to an even weight function. *Integral Transform. Spec. Funct.* **25**(9), 721–729 (2014)
19. R. Koekoek, P.A. Lesky, R.F. Swarttouw, *Hypergeometric Orthogonal Polynomials and Their q -Analogues*. *Springer Monographs in Mathematics* (Springer, Berlin, 2010)
20. T.H. Koornwinder, Orthogonal polynomials with weight function $(1-x)^\alpha(1+x)^\beta + M\delta(x+1) + N\delta(x-1)$. *Canad. Math. Bull.* **27**, 205–214 (1984)

21. A.M. Krall, Orthogonal polynomials satisfying fourth order differential equations. Proc. Roy. Soc. Edinb. Sec. A. **87**, 271–288 (1980/1981)
22. H.L. Krall, On orthogonal polynomials satisfying a certain fourth order differential equation. Pa. State Coll. Stud. **6**, 24 (1940)
23. A. Kroó, F. Peherstorfer, On the zeros of polynomials of minimal L_p -norm. Proc. Amer. Math. Soc. **101**, 652–656 (1987)
24. A. Laforgia, A monotonic property for the zeros of ultraspherical polynomials. Proc. Amer. Math. Soc. **83**, 757–758 (1981)
25. A. Laforgia, Monotonicity properties for the zeros of orthogonal polynomials and Bessel function, in *Polynômes Orthogonaux et Applications*, ed. by C. Brezinski, A. Draux, A. P. Magnus, P. Maroni, A. Ronveaux. Lecture Notes in Mathematics, vol. 1171 (Springer, Berlin, 1985), pp. 267–277
26. F. Marcellán, P. Maroni, Sur l’adjonction d’une masse de Dirac à une forme régulière et semi-classique. Ann. Mat. Pura Appl. **162**(4), 1–22 (1992)
27. A. Markov, Sur les racines de certaines équations (second note). Math. Ann. **27**, 177–182 (1886)
28. P. Maroni, Une théorie algébrique des polynômes orthogonaux. Application aux polynômes orthogonaux semi-classiques, in *Orthogonal Polynomials and Their Applications*, ed. by C. Brezinski, L. Gori, A. Ronveaux. Annals Comput. Appl. Math. vol. 9 (Baltzer, Basel, 1991), pp. 95–130
29. P. Natalini, B. Palumbo, Some monotonicity results on the zeros of the generalized Laguerre polynomials. J. Comput. Appl. Math. **153**, 355–360 (2003)
30. A.F. Nikiforov, S.K. Suslov, V.B. Uvarov, *Classical Orthogonal Polynomials of a Discrete Variable* (Springer, Berlin, 1991)
31. E.D. Rainville, *Special Functions* (Chelsea, Bronx, 1971)
32. T.J. Stieltjes, Sur les racines de l’équation $X_n = 0$. Acta Math. **9**, 385–400 (1887)
33. G. Szegő, *Orthogonal Polynomials*. Amer. Math. Soc. Coll. Publ., vol. 23, 4th edn. (American Mathematical Society, Providence, 1975)
34. E.T. Whittaker, G.N. Watson, *A Course of Modern Analysis*, 4th edn. (Cambridge University Press, Cambridge, 1927)

A Review of Two Network Curvature Measures



Tanima Chatterjee, Bhaskar DasGupta, and Réka Albert

Abstract The curvature of higher-dimensional geometric shapes and topological spaces is a natural and powerful generalization of its simpler counterpart in planes and other two-dimensional spaces. Curvature plays a fundamental role in physics, mathematics, and many other areas. However, graphs are discrete objects that do not necessarily have an associated natural geometric embedding. There are many ways in which curvature definitions of a continuous surface or other similar space can be adapted to graphs depending on what kind of local or global properties the measure is desired to reflect. In this chapter, we review two such measures, namely the Gromov-hyperbolic curvature measure and a geometric measure based on topological associations to higher-dimensional complexes.

1 Introduction

Useful insights for many complex systems are often obtained by representing them as graphs¹ and analyzing them using graph-theoretic and combinatorial tools [50]. For analyzing graphs, researchers have proposed and evaluated a number of established graph-theoretic measures such as the *degree-based measures*, (e.g., degree distributions), *connectivity-based measures*, (e.g., clustering coefficients), *geodesic-based measures* (e.g., betweenness centralities), and other more novel graph-theoretic measures such as in [2, 6, 42]. To simplify exposition for this

¹In this chapter, we use the two terms “graph” and “network” interchangeably.

T. Chatterjee · B. DasGupta (✉)

Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA
e-mail: tchatt2@uic.edu; bdasgup@uic.edu

R. Albert

Department of Physics, Pennsylvania State University, University Park, PA, USA
e-mail: ralbert@phys.psu.edu

chapter, our input is an undirected weighted or unweighted graph $G = (V, E)$ of n nodes v_1, \dots, v_n ; the adjacency matrix for G is denoted by $A(G) = [a(G)_{i,j}]$ where $a(G)_{i,j} = 1$ (resp., $a(G)_{i,j} = 0$) if $\{v_i, v_j\} \in E$ (resp., if $\{v_i, v_j\} \notin E$). The notations $\overline{u, v}$ and $\text{dist}_H(u, v)$ denote a *shortest path* between nodes u and v , and the distance between nodes u and v in graph H , respectively.

The graph-theoretic measure discussed in this chapter is an appropriate notion of “network curvature.” A curvature for a graph G for this chapter is a scalar function $\mathcal{C} : G \mapsto \mathbb{R}$. Curvatures are very natural measures of the anomaly of higher-dimensional objects used in mainstream physics and mathematics [10, 13]. However, graphs are discrete objects that do *not* necessarily have an associated natural geometric embedding. There are many ways in which curvature definitions of a continuous surface or other similar space can be adapted to graphs depending on what kind of local or global properties the measure is desired to reflect. More specifically, we discuss *Gromov-hyperbolic* curvature (based on the properties of geodesics and higher-order connectivities) and *geometric curvatures* (based on identifying network motifs with geometric complexes), both of which encode non-trivial *higher-order* correlation among nodes. Some important characteristics of these two curvature measures are as follows.

- ▶ They depend on *non-trivial global* network properties, as opposed to measures such as *degree distributions* or *clustering coefficients* that are *local* in nature or *dense subnetworks* that use *only* pairwise correlations.
- ▶ They can mostly be computed efficiently in polynomial time, as opposed to NP-complete measures such as *cliques* [29], *densest-k-sub-graphs* [29], or some types of community decompositions such as *modularity maximization* [20].
- ▶ When applied to real-world networks, these curvature measures can explain many phenomena one frequently encounters in real graph-theoretic applications that are *not* easily explained by other measures.

2 Gromov-Hyperbolic Curvature

This type of measure for a metric space was originally suggested by Gromov in the context of group theory [33] by observing that many results concerning the fundamental group of a Riemann surface hold true in a more general context. The measure was first defined for *infinite* continuous metric space via properties of geodesics (e.g., see the textbook [13]), but was later also adopted for *finite* graphs. Usually the measure is defined via *geodesic triangles* in the following manner.

Definition 1 (Gromov Curvature Measure via Geodesic Triangles) A graph G has a Gromov curvature (or Gromov-hyperbolicity) of $\delta \stackrel{\text{def}}{=} \delta(G)$ if and only if for every three ordered triple of shortest paths $(\overline{u, v}, \overline{u, w}, \overline{v, w})$, $\overline{u, v}$ lies in a δ -neighborhood of $\overline{u, w} \cup \overline{v, w}$, i.e., for every node x on $\overline{u, v}$, there exists a node y on $\overline{u, w}$ or $\overline{v, w}$ such that $\text{dist}_G(x, y) \leq \delta$.

Definition 2 (The Class of δ -Gromov-Hyperbolic Graphs) An infinite collection \mathcal{G} of graphs belongs to the class of δ -Gromov-hyperbolic graphs (or, simply δ -hyperbolic graphs) if and only if any graph $G \in \mathcal{G}$ has a Gromov curvature of at most δ .

Informally, any infinite metric space has a finite value of δ if it behaves metrically in the large scale as a *negatively curved* Riemannian manifold, and thus the value of δ can be related to the standard scalar curvature of a hyperbolic manifold. For example, a simply connected complete Riemannian manifold whose sectional curvature is below $\alpha < 0$ has a value of $\delta = O((-\alpha)^{-1/2})$ (see [58]). This is a justification of using the value δ as a notion of curvature of any metric space.

For the purpose of designing computational algorithms, it is often useful to consider another alternate but *equivalent* (“up to a constant multiplicative factor”) way of defining Gromov curvature for a graph G via the following 4-node conditions.

Definition 3 (Equivalent Definition of Gromov Curvature via 4-Node Conditions) For a set $\{u_1, u_2, u_3, u_4\}$ of four nodes, let $(\pi_1, \pi_2, \pi_3, \pi_4)$ be a permutation of $\{1, 2, 3, 4\}$ denoting a rearrangement of the indices of nodes such that

$$\begin{aligned} \text{dist}_G(u_{\pi_1}, u_{\pi_2}) + \text{dist}_G(u_{\pi_3}, u_{\pi_4}) &\leq \text{dist}_G(u_{\pi_1}, u_{\pi_3}) + \text{dist}_G(u_{\pi_2}, u_{\pi_4}) \leq \text{dist}_G(u_{\pi_1}, u_{\pi_4}) + \text{dist}_G(u_{\pi_2}, u_{\pi_3}) \\ = S_{u_1, u_2, u_3, u_4} &= M_{u_1, u_2, u_3, u_4} = L_{u_1, u_2, u_3, u_4} \end{aligned}$$

Let $\widehat{\delta} = \widehat{\delta}(G) = \max_{u_1, u_2, u_3, u_4 \in V} \{L_{u_1, u_2, u_3, u_4} - M_{u_1, u_2, u_3, u_4}\}/2$. Then, if \mathcal{G} is a δ -Gromov-hyperbolic graph, then $\delta/c \leq \widehat{\delta} \leq c \delta$ for some absolute constant $c > 0$.

In order to account for the fact that sometimes the value of $\widehat{\delta}(G)$ may be a rare deviation from typical values of $L_{u_1, u_2, u_3, u_4} - M_{u_1, u_2, u_3, u_4}$ that one would obtain for most combinations of nodes $\{u_1, u_2, u_3, u_4\}$, the authors in [3] defined the *average Gromov curvature* of a graph G as $\delta_{\text{ave}}(G) = \sum_{u_1, u_2, u_3, u_4 \in V} (L_{u_1, u_2, u_3, u_4} - M_{u_1, u_2, u_3, u_4}) / \binom{n}{4}$ such that $\delta_{\text{ave}}(G)$ is the expected value of $L_{u_1, u_2, u_3, u_4} - M_{u_1, u_2, u_3, u_4}$ when the four nodes u_1, u_2, u_3, u_4 are picked independently and uniformly at random from the set of all nodes.

It is easy to see that if G is a tree, then $\delta(G) = \widehat{\delta}(G) = 0$, and $\widehat{\delta}(G) \leq D/2$ where D is the diameter of the given graph. Other examples of graph classes for which $\delta(G)$ and $\widehat{\delta}(G)$ are small constants include *chordal graphs*, *cactus of cliques*, *AT-free graphs*, *link graphs of simple polygons*, and *any class of graphs with a fixed diameter*. On the other hand, theoretical investigations have revealed that *expanders*, *vertex-transitive graphs*, and (for certain parameter ranges) classical *Erdős-Rényi* random graphs are δ -hyperbolic only for $\delta = \omega(1)$ [7–9, 45, 47].

2.1 Topological Characteristics of Gromov-Hyperbolicity Measure

The Gromov-hyperbolicity measure $\delta(G)$ enjoys *many* non-trivial topological characteristics. Some examples are as follows.

- ▷ The “ $\delta = o(n)$ ” property is not hereditary (and thus also not monotone). For example, removing a single node or edge can increase/decrease the value of δ *very* sharply.
- ▷ A small value of δ does *not* necessarily imply that the graph is a tree. For example, *all* bounded-diameter graphs have $\delta = O(1)$ irrespective of whether they are tree or not (however, graphs with $\delta = O(1)$ need *not* be of bounded diameter). In general, even for small δ , the metric induced by a δ -hyperbolic graph may be quite far from a tree metric [17].
- ▷ A similar popular measure used in both the bioinformatics and theoretical computer science literature is the *tree-width* measure first introduced by Robertson and Seymour [57]. However, as observed in [23] and elsewhere, the two measures are *not* correlated.

We end this section with a very important topological consequence of small Gromov-hyperbolicity values of a graph, popularly known as the “divergence of geodesic rays” property. The result appears in several forms in prior works such as [3, 7, 13, 33, 45]; we state two such versions. Let $\mathcal{B}(u, r)$ denote the set of nodes contained in a ball of radius r centered at node u in graph G , i.e., $\mathcal{B}(u, r) = \{v \mid \text{dist}_G(u, v) \leq r\}$

Fact 1 (Cylinder Removal Around a Geodesic) *Assume that G is a δ -hyperbolic graph. Let p and q be two nodes of G such that $\text{dist}_G(p, q) = \beta > 6$, and let p', q' be nodes on a shortest path between p and q such that $\text{dist}_G(p, p') = \text{dist}_G(p', q') = \text{dist}_G(q', q) = \beta/3$. For any $0 < \alpha < 1/4$, let \mathcal{C} be set of nodes at a distance of $\alpha\beta - 1$ of a shortest path p', q' between p' and q' , i.e., let $\mathcal{C} = \{u \mid \exists v \in \overline{p', q'} : \text{dist}_G(u, v) = \alpha\beta - 1\}$. Let $G_{-\mathcal{C}}$ be the graph obtained from G by removing the nodes in \mathcal{C} . Then, $\text{dist}_{G_{-\mathcal{C}}}(p, q) \geq (\beta/60) 2^{\alpha\beta/\delta}$.*

Fact 2 (Exponential Divergence of Geodesic Rays) *Assume that G is a δ -hyperbolic graph. Suppose that we are given the following:*

- three integers $\kappa \geq 4$, $\alpha > 0$, $r > 3\kappa\delta$, and
- five nodes v, u_1, u_2, u_3, u_4 such that $\text{dist}_G(v, u_1) = \text{dist}_G(v, u_2) = r$, $\text{dist}_G(u_1, u_2) \geq 3\kappa\delta$, $\text{dist}_G(v, u_3) = \text{dist}_G(v, u_4) = r + \alpha$, and $\text{dist}_G(u_1, u_4) = \text{dist}_G(u_2, u_3) = \alpha$.

Consider any path \mathcal{Q} between u_3 and u_4 that does not involve a node in

$$\bigcup_{0 \leq j \leq r+\alpha} \mathcal{B}(v, j). \text{ Then, the length } |\mathcal{Q}| \text{ of the path } \mathcal{Q} \text{ satisfies } |\mathcal{Q}| > 2^{\frac{\alpha}{6\delta} + \kappa + 1}.$$

For example, these facts are used by Benjamini in [7] to show that graph classes with a constant value of δ *cannot* be expanders and also by Malyshev in [45] to show

that expander graphs must have Gromov-hyperbolicity *at least* proportional to their diameter. Further works on the effect of the hyperbolicity measure δ on expansion and cut-size bounds on graphs and its algorithmic implications are reported in [21].

2.2 Gromov Curvature of Real-World Networks

Recently, there has been a surge of empirical works measuring and analyzing the Gromov curvature δ of networks, and many real-world networks (e.g., preferential attachment networks, networks of high power transceivers in wireless sensor networks, communication networks at the IP layer and at other levels) were observed to have a small constant value of δ [5, 35, 36, 46, 55]. Moreover, extreme congestion at a small number of nodes in a large traffic network that uses the shortest-path routing was shown in [38] to be caused by a small value of δ of the network. The authors in [3] computed Gromov-hyperbolicity values for 11 biological networks (3 transcriptional regulatory, 5 signaling, 1 metabolic, 1 immune response, and 1 oriented protein-protein-interaction networks) and 9 social networks. They reported that the hyperbolicity values of all except one network are small and statistically significant. They also reported several interesting experimentally validated implications of these hyperbolicity values, such as

- ▷ Independent pathways that connect a signal to the same output node (e.g., transcription factor) are rare, and if multiple pathways exist, then they are interconnected through cross-talks.
- ▷ All the biological networks have central influential small-size node neighborhoods that can be selected to find knock-out nodes to cut off specific up- or down-regulation.

2.3 Efficient Computation of Gromov Curvature

Using Definition 3 directly one can compute $\delta(G)$ in $O(n^4)$ time, but this time complexity is prohibitive for large graphs. For faster computation, one needs to define Gromov curvature via an equivalent but more algorithmically amenable formulation as follows.

Definition 4 (Equivalent Definition of Gromov Curvature via Gromov-Product Nodes [33]) For any three nodes u , v , and r , the Gromov-product of u and v anchored at r is defined by

$$(u|v)_r = \frac{1}{2} \left(\text{dist}(u, r) + \text{dist}(v, r) - \text{dist}(u, v) \right)$$

Define the value of Gromov-hyperbolicity “anchored” at a node r as:

$$\delta_r = \max_{u,v,w} \left\{ \min \{ (u|w)_r, (v|w)_r \} - (u|v)_r \right\}$$

Then, the value of Gromov-hyperbolicity of a graph G is defined as

$$\delta \stackrel{\text{def}}{=} \delta(G) = \max_r \{ \delta_r \}$$

The value of $\delta(G)$ computed via Definition 4 is identical to the one computed via geodesic triangles in Definition 1. It was also shown in [33] that $\delta(G) \leq \delta_r \leq 2\delta(G)$ for any r . Let ω be the value such that two $n \times n$ matrices can be multiplied in $O(n^\omega)$ time; the smallest current value of ω is 2.373 [64]. The (max, min)-matrix multiplication of two $n \times n$ matrices A and B , denoted by $A \otimes B$, is defined as:

$$A \otimes B[i, j] = \max_k \min \{ A[i, k], B[k, j] \}$$

Duan and Pettie in [24] showed that $A \otimes B$ can be computed in $O(n^{(3+\omega)/2}) = O(n^{2.688})$ time. Subsequently, Fournier, Ismail, and Vigneron [27] showed that computation of δ_r can be cast as computing a (max, min)-matrix multiplication problem; as a result, one can compute $\delta(G)$ and a 2-approximation of $\delta(G)$ in $O(n^{(5+\omega)/2}) = O(n^{3.69})$ and in $O(n^{(3+\omega)/2}) = O(n^{2.69})$ time, respectively. Faster less accurate approximation is also known, e.g., Chalopin et al. [14] showed that a 8-approximation of $\delta(G)$ can be computed in $O(n^2)$ time. On the other hand, an exact computation of $\delta(G)$ involves computing the “all-pairs-shortest-path” problem which is widely conjectured to take at least $\Omega(n^3)$ time (and, can be done in $O(n^3)$ time [19]).

2.4 Algorithmic Implications of Small Gromov Curvature

A small value of Gromov curvature δ is often crucial for algorithmic designs; for example, several routing-related problems or the diameter estimation problem become easier for graphs with small δ values [16–18, 30]. DasGupta et al. in [21] discussed further implications of small values of δ for several graph-theoretic problems. In particular, they showed that a large family of s - t cuts having at most $d^{O(\delta)}$ cut-edges can be found in polynomial time in δ -hyperbolic graphs of n nodes when d is the maximum degree of any node except s , t and any node within a distance of 35δ of s and the distance between s and t is at least $\Omega(\delta \log n)$, and used such a result to design an approximation algorithm for minimizing bottleneck edges in a graph.

2.5 Statistical Validation of Gromov Curvature via “Scaled” Version

Suppose that $\delta(G)$ has been computed for a given graph G of n nodes and it is indeed a small value compared to the size of the graph. One major task for empirical researchers is then to determine more precisely if $\delta(H)$ is indeed a small number independent of the size of H for every graph H in the class of graphs \mathcal{G} to which G belongs (as opposed to $\delta(H)$ being small specifically only for the particular graph $H = G$ in \mathcal{G}). For this purpose, we can make use of a “scaled” version of Gromov curvature [36, 37, 46]. The basic idea is to “scale” the values of $L_{u_1,u_2,u_3,u_4} - M_{u_1,u_2,u_3,u_4}$ in Definition 3 by a suitable scaling factor μ_{u_1,u_2,u_3,u_4} such that there exists a constant $0 < \varepsilon < 1$ with the following property:

the maximum achievable value of $(L_{u_1,u_2,u_3,u_4} - M_{u_1,u_2,u_3,u_4})/\mu_{u_1,u_2,u_3,u_4}$ is ε in the standard hyperbolic space or in the Euclidean space, and $(L_{u_1,u_2,u_3,u_4} - M_{u_1,u_2,u_3,u_4})/\mu_{u_1,u_2,u_3,u_4}$ goes beyond ε in positively curved spaces.

By using theoretical or empirical calculations, the authors in [37] provide the bounds shown in Figure 1. Following the ideas espoused in [3, 37], assuming G is a connected graph we can use the following criterion to determine if $\delta(H)$ is indeed a small number independent of the size of H for every graph $H \in \mathcal{G}$:

Let $0 < \eta < 1$ be a value indicating the confidence level of our criterion. Then, $\delta(H)$ is a small number independent of the size of H for every graph $H \in \mathcal{G}$ if and only if

$$\forall Y \in \{\emptyset, L, L + M + S\} : \Delta^Y(G) = \frac{\text{number of subset of four nodes } \{u_i, u_j, u_k, u_\ell\} \text{ such that } \delta_{u_i, u_j, u_k, u_\ell}^Y > \varepsilon}{\binom{n}{4}} < 1 - \eta$$

In the above criterion, larger values of η indicate better confidence levels. An alternative method would be to use the procedure outlines in Section 3.7.

Name	Notation	μ_{u_1,u_2,u_3,u_4}	ε
diameter-scaled curvature	$\delta^{\mathcal{D}}$	$\max_{i,j \in \{1,2,3,4\}} \{\text{dist}_G(u_i, u_j)\}$	0.2929
L -scaled curvature	δ^L	L_{u_1,u_2,u_3,u_4}	$\frac{\sqrt{2}-1}{2\sqrt{2}}$
$(L + M + S)$ -scaled curvature	δ^{L+M+S}	$L_{u_1,u_2,u_3,u_4} + M_{u_1,u_2,u_3,u_4} + S_{u_1,u_2,u_3,u_4}$	0.0607

Fig. 1 [37] Various scaled Gromov curvatures

3 Geometric Curvature

There are many well-known measures of curvature of a continuous surface or other similar spaces (e.g., curvature of a manifold) that are widely used in many branches of physics and mathematics. In section 2 we discussed how to relate Gromov curvature to such other curvature notions indirectly via introduction of its scaled version. In this section, we describe a notion of geometric curvatures of graphs by using a correspondence with topological objects in *higher dimension*.

3.1 Basic Topological Concepts

In this section we review some basic concepts from topology; see introductory textbooks such as [28, 34] for further information. For concreteness of exposition, let the underlying metric space be the r -dimensional real space \mathbb{R}^r be for some integer $r > 1$. See Figure 2 for some illustrations of these concepts in \mathbb{R}^3 .

- ▷ A subset $S \subseteq \mathbb{R}^r$ is *convex* if and only if for any pair $x, y \in S$, the *convex combination* of x and y is also in S (i.e., $\lambda x + (1 - \lambda)y \in S$ for any real $0 \leq \lambda \leq 1$).
- ▷ A set of $k + 1$ points $x_0, \dots, x_k \in \mathbb{R}^r$ are called *affinely independent* if and only if for all $\alpha_0, \dots, \alpha_k \in \mathbb{R}$ $\sum_{j=0}^k \alpha_j x_j = 0$ and $\sum_{j=0}^k \alpha_j = 0$ implies $\alpha_0 = \dots = \alpha_k = 0$.
- ▷ The k -*simplex* generated by a set of $k + 1$ affinely independent points $x_0, \dots, x_k \in \mathbb{R}^r$ is the subset of \mathbb{R}^r $\mathcal{A}(x_0, \dots, x_k) = \{ \sum_{j=0}^k \alpha_j x_j \mid \forall j : \alpha_j \geq 0 \text{ and } \sum_{j=0}^k \alpha_j = 1 \}$ generated by *all* convex combinations of x_0, \dots, x_k . For example, the equation of a k -simplex with *unit intercepts* is given by $\sum_{j=0}^k x_j = 1$ with $x_j \geq 0$ for all $0 \leq j \leq k$.
 - ▶ Each $(\ell + 1)$ -subset $\{x_{i_0}, \dots, x_{i_\ell}\} \subseteq \{x_0, \dots, x_k\}$ defines the ℓ -simplex

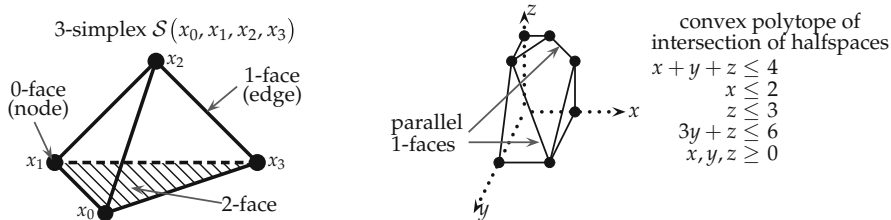


Fig. 2 (Modified from [54]) Illustrations of some topological concepts discussed in Section 3.1 over \mathbb{R}^3

$\mathcal{A}(x_{i_0}, \dots, x_{i_\ell})$ that is called a *face* of dimension ℓ (or a ℓ -*face*) of $\mathcal{A}(x_0, \dots, x_k)$. A $(k - 1)$ -face, 1-face, and 0-face is called a *facet*, an *edge*, and a *node*, respectively.

- ▷ A (closed) *halfspace* is a set of points satisfying $\sum_{j=1}^r a_j x_j \leq b$ for some $a_1, \dots, a_r, b \in \mathbb{R}$. The convex set obtained by a bounded non-empty intersection of a finite number of halfspaces is called a *convex polytope* (called a *convex polygon* in two dimensions).
 - ▶ If the intersection of a halfspace and a convex polytope is a subset of the halfspace, then it is called a *face* of the polytope. Of particular interests are faces of dimensions $r - 1$, 1, and 0, which are called *facets*, *edges*, and *nodes* of the polytope, respectively.
- ▷ We can define a partial order relation $<_{\mathfrak{f}}$ between faces of various dimensions of a simplex or a convex polytope in the usual manner: a ℓ -face \mathfrak{f}^ℓ is a parent of a ℓ' -face $\widehat{\mathfrak{f}}^{\ell'}$ (denoted by $\widehat{\mathfrak{f}}^{\ell'} <_{\mathfrak{f}} \mathfrak{f}^\ell$) if $\widehat{\mathfrak{f}}^{\ell'}$ is contained in \mathfrak{f}^ℓ . Similarly, two ℓ -faces \mathfrak{f}^ℓ and $\widehat{\mathfrak{f}}^\ell$ are *parallel* (denoted by $\mathfrak{f}^\ell \parallel_{\mathfrak{f}} \widehat{\mathfrak{f}}^\ell$) if they have either at least one common *immediate predecessor* or at least one common *immediate successor* (in the partial order $<_{\mathfrak{f}}$) *but not both*.
- ▷ A *simplicial complex* (or just a complex) is a topological space constructed by the union of simplexes via topological associations.

3.2 Topological Association of Networks with a Complex

Informally, a *complex* is “glued” from nodes, edges, cycles, and other sub-graphs of the given graph via topological identification. There are many alternate ways such topological associations can be performed. Here we describe a simple association as used in [22]; for other possible alternative associations the reader is referred to papers such as [11, 26, 62, 63].

To begin our topological association, we (topologically) associate a q -simplex with a $(q + 1)$ -clique \mathcal{K}_{q+1} ; for example, 0-simplexes, 1-simplexes, 2-simplexes, and 3-simplexes are associated with nodes, edges, 3-cycles (triangles), and 4-cliques, respectively. Next, we would also need the concept of an “order” of a simplex for more non-trivial topological association. Consider a p -face f^p of a q -simplex. An order d association of such a face, which we will denote by the notation f_d^p with the additional subscript d , is associated with a sub-graph of *at most* d nodes that is obtained by starting with \mathcal{K}_{p+1} and then *optionally* replacing each edge by a path between the two nodes. For example,

- ▷ f_d^0 is a node of G for all $d \geq 1$.
- ▷ f_2^1 is an edge, and f_d^1 for $d > 2$ is a path having at most d nodes between two nodes adjacent in G .

- ▷ f_3^2 is a triangle (cycle of 3 nodes or a 3-cycle), and f_d^2 for $d > 3$ is obtained from 3 nodes by connecting every pair of nodes by a path such that the total number of nodes in the sub-graph is at most d .

Naturally, the higher the values of p and q are, the more complex are the topological associations.

3.3 Defining Geometric Curvatures for Elementary Components of Given Graph

By elementary components of a graph, we mean sub-graphs of small size such as edges, triangles, 4-cycles, and so forth. In this section, we discuss the case when the elementary components are edges; the other cases can be found in the previously cited references. As discussed in Section 3.2, geometric curvatures are defined by “extrapolating” graphs to higher-dimensional complexes via topological association. For these associations, it is often useful to assign a positive “weight” from the interval $[0, 1]$ to every pair of nodes (1-simplexes) and to every node (0-simplexes) of the graph $G = (V, E)$. If G comes with its own node or edge weights, we may use them directly after normalizing them such that all weights lie between 0 and 1. Otherwise, some choices for these weights that may be appropriate are the following:

- (a) For every pair of nodes $e_{i,j} = \{v_i, v_j\}$, a natural choice for the weight would be $w_{\text{edge}}(e_{i,j}) = 1$ (resp., $w_{\text{edge}}(e_{i,j}) = 0$) if $\{v_i, v_j\} \in E$ (resp., $\{v_i, v_j\} \notin E$). One may also consider more refined choices, e.g., $w_{\text{edge}}(e_{i,j}) = 1/\text{dist}_G(v_i, v_j)$ or a “distance-thresholded” version of it, which may be useful in the study of social networks of the “small world” type [62].
- (b) A natural choice for the weight of a node v_i would be $w_{\text{node}}(v_i) = 1$. A more sophisticated choice that one may consider is

$$w_{\text{node}}(v_i) = \frac{\sum_{v_j: w_{\text{edge}}(\{v_i, v_j\}) \geq \gamma} w_{\text{edge}}(\{v_i, v_j\})}{|\{w_{\text{edge}}(e) \mid e \in E \text{ and } w_{\text{edge}}(e) \geq \gamma\}|}$$

that provides more weight to nodes with higher weighted-degree [62].

Once we have fixed a weighting scheme for 0-simplexes and 1-simplexes, we can assign weights to *higher-dimensional* objects such as k -faces as follows:

- 2-faces:** For a triangle, say $\mathcal{S}(v_1, v_2, v_3)$ with $e_{i,j} = \{v_i, v_j\}$ for $i, j \in \{1, 2, 3\}$, we may assign its weight based on the area of the triangle [63]:

$$w(\mathcal{A}(v_1, v_2, v_3)) = \left[s \left(\prod_{\substack{i,j \in \{1,2,3\} \\ i \neq j}} (s - w_{\text{edge}}(e_{i,j})) \right) \right]^{1/2} \text{ where } s = \sum_{\substack{i,j \in \{1,2,3\} \\ i \neq j}} \frac{w_{\text{edge}}(e_{i,j})}{2}$$

For a polygon of p sides with $p > 3$, we can first do a triangulation of the polygon and then add the weights of these triangles to get the weight for the entire polygon.

k -faces for $k > 2$: We can compute the weight by adding the weights of the $(k - 1)$ -subfaces of this face (for the degenerate case, we will consider subfaces of dimensions lower than $k - 1$ also). Alternately, for some cases, we may also use direct combinatorial formulae for the volume.

Let $w(f)$ denote the weight of an arbitrary face f .

1-Complex-Based Geometric Curvature for a Pair of Nodes

A graph is naturally defined by 1-simplexes (edges) and 0-simplexes (nodes). Thus, without further topological association, a 1-complex-based Forman’s combinatorial Ricci curvature for a pair of nodes $\{v_i, v_j\}$ is given by [26, 62]:

$$\mathfrak{C}_{i,j}^1 = \begin{cases} 0, & \text{if } w_{\text{edge}}(e_{i,j}) = 0 \\ w_{\text{edge}}(e_{i,j}) \left[\frac{w_{\text{node}}(v_i)}{w_{\text{edge}}(e_{i,j})} + \frac{w_{\text{node}}(v_j)}{w_{\text{edge}}(e_{i,j})} - \sum_{\substack{e_{i,j_1}, e_{i_1,j} \\ w_{\text{edge}}(e_{i,j_1}) \neq 0 \\ w_{\text{edge}}(e_{i_1,j}) \neq 0}} \left(\frac{w_{\text{node}}(v_i)}{\sqrt{(w_{\text{edge}}(e_{i,j})w_{\text{edge}}(e_{i,j_1}))}} + \frac{w_{\text{node}}(v_j)}{\sqrt{(w_{\text{edge}}(e_{i,j})w_{\text{edge}}(e_{i_1,j}))}} \right) \right], & \text{otherwise} \end{cases} \tag{1}$$

2-Complex-Based Geometric Curvature for a Pair of Nodes

For 2-complex-based geometric curvatures, we also include topological associations with 2-simplexes (cycles of 3 nodes). Let $\mathcal{C}(v_i, v_j, v_k)$ denote a cycle of length 3 consisting of the edges $\{v_i, v_j\}$, $\{v_j, v_k\}$, and $\{v_i, v_k\}$. Note that in Equation (1) the edges e_{i,j_1} and $e_{i_1,j}$ in the summation actually satisfy $e_{i,j_1} \parallel_{\mathfrak{f}} e_{i,j}$ and $e_{i_1,j} \parallel_{\mathfrak{f}} e_{i,j}$. This observation helps us to lead to Forman’s combinatorial Ricci curvature for 2-complexes [63]:

$$\mathfrak{C}_{i,j}^2 = \begin{cases} 0, & \text{if } w_{\text{edge}}(e_{i,j}) = 0 \\ w_{\text{edge}}(e_{i,j}) \left[\left(\sum_{v_k \neq v_i, v_j} \frac{w(\mathcal{S}(v_i, v_j, v_k))}{w_{\text{edge}}(e_{i,j})} \right) + \frac{w_{\text{node}}(v_i)}{w_{\text{edge}}(e_{i,j})} + \frac{w_{\text{node}}(v_j)}{w_{\text{edge}}(e_{i,j})} \right. \\ \left. - \sum_{\substack{e_{i_1, j_1}: e_{i_1, j_1} \parallel e_{i,j} \\ w_{\text{edge}}(e_{i_1, j_1}) \neq 0}} \left| \sum_{\substack{v_k \neq v_i, v_j \\ w(\mathcal{S}(v_i, v_j, v_k)) \neq 0}} \frac{\sqrt{w_{\text{edge}}(e_{i,j}) w_{\text{edge}}(e_{i_1, j_1})}}{w(\mathcal{S}(v_i, v_j, v_k))} + \sum_{v \in \{v_i, v_j\} \cap \{v_{i_1}, v_{j_1}\}} \frac{w_{\text{node}}(v)}{\sqrt{(w_{\text{edge}}(e_{i,j}) w_{\text{edge}}(e_{i_1, j_1}))}} \right| \right], & \text{otherwise} \end{cases} \quad (2)$$

Higher-Dimensional Geometric Curvature for a Pair of Nodes

k -complex-based curvature $\mathfrak{C}_{i,j}^k$ for $k > 2$ can be defined in a similar manner (e.g., a clique of $k + 1$ nodes correspond to a k -simplex).

3.4 Overall (Scalar) Curvature Value for a Network

One can compute a single scalar value \mathfrak{C} of geometric curvature based on the values of $\mathfrak{C}_{i,j}^k$ values using curvature functions defined by Bloch [11], by using *Euler characteristics* [22] or similar other methods. We discuss the simplest unweighted Euler characteristics based scalar graph curvature as used by DasGupta et al. in [22]. Let \mathcal{F}_d^k be the set of all f_d^k 's that are topologically associated as described in Section 3.2. With such associations via p -faces of order d , the Euler characteristics of the graph $G = (V, E)$ and consequently the curvature can be defined as

$$\mathfrak{C}_d^p(G) \stackrel{\text{def}}{=} \sum_{k=0}^p (-1)^k |\mathcal{F}_d^k|$$

It is easy to see that both $\mathfrak{C}_d^0(G)$ and $\mathfrak{C}_d^1(G)$ are too simplistic to be of use in practice. Considering the next higher value of p , namely $p = 2$, and letting $\mathcal{C}(G)$ denote the number of cycles of at most $d + 1$ nodes in G , we get the following scalar curvature measure for a given graph $G = (V, E)$:

$$\mathfrak{C}_d^2(G) = |V| - |E| + |\mathcal{C}(G)| \quad (3)$$

3.5 Computation of Geometric Curvatures

Let $G = (V, E)$ be the given connected graph with n nodes and m edges. Using Equations (1) and (2) and appropriate data structures, $\mathfrak{C}_{i,j}^1$ and $\mathfrak{C}_{i,j}^2$ can be computed

roughly in $O(m^2)$ and $O(m^3)$ times, respectively. More generally, $\mathfrak{C}_{i,j}^k$ can be computed in $O(m^{O(k)})$ time and $\mathfrak{C}_d^2(G)$ in Equation (3) can be computed in $O(m^d)$ time.

3.6 Real-World Networks and Geometric Curvatures

The usefulness of geometric curvatures for real-world networks was demonstrated in publications such as [59, 62, 63]. Some of these results are as follows.

- ▷ Samal et al. in [59] *empirically* compared geometric curvatures of the type discussed in this chapter with another notion of network curvature, namely the *Ollivier's discretization* of Ricci curvature [53]. Although the Ollivier-Ricci curvature measures were developed based on quite different properties of the classical smooth notion as compared to the geometric curvatures discussed in this chapter, somewhat surprisingly they found that these two measures *are* correlated for many real networks. However, as the authors themselves cautioned in [53], their results should not be construed as implying that one of these curvature measures can be used as a *universal substitute* for the other measure, but merely that for many real networks using one of these that allow faster implementation may suffice.
- ▷ Weber, Saucan, and Jost in [63] computed a specific version of the geometric curvatures discussed in this chapter (the “Euler characteristics” with only up to 2-faces of degree 3) for several real-world networks, such as Zachary’s karate club, social interactions of dolphins, and *E. coli* transcription networks, and showed that networks with a high number of high-degree faces have positive Euler characteristics, whereas low numbers of high-degree faces might hint on negative Euler characteristics.

3.7 Statistical Validations for all Curvature Measures

We may test the statistical significance of any curvature measure $\mathfrak{C}(G)$ by computing its statistical significance value (commonly called *p-value*) with respect to a *null-hypothesis model* of the network. For this purpose, we may use a method as described below that is similar to that used by many other researchers in the network science literature (e.g., see [2, 60]). For each graph G , we will generate a large number q of random graphs G_1, \dots, G_q of the *same* type as G . There are many methods for generating such random graphs. Two such methods are as follows.

Generative null-hypothesis models: One most frequently reported topological characteristics of graphs is the distribution of degrees of nodes. We may select appropriate degree distributions for our given class of graphs that is consistent

with the findings in prior literature. For example, based on the known topological characterizations for biological *transcriptional* and *signaling* networks we may use the following degree distributions [1, 31, 44]:

- (a) the in-degree distribution is *exponential*, and
- (b) the out-degree distribution is governed by a *power-law*.

Random networks with prescribed degree distributions can be generated using the method by Newman et al. [52].

Non-generative null-hypothesis models: For graphs where a consensus degree distribution may be difficult to ascertain, we can use the following methods:

- ▷ We may generate random networks using a *Markov-chain algorithm* [39] by repeatedly swapping randomly chosen compatible pairs of connections in G .
- ▷ We may generate random networks from the degree distribution of G using the method pioneered by Newman and others in [32, 43, 48, 49, 51] that preserves *in expectation* the degree distribution of each node.

Once the random graphs G_1, \dots, G_q have been generated, we first compute the values of $\mathcal{C}(G_1), \dots, \mathcal{C}(G_q)$, and next use a suitable statistical test to determine the probability that $\mathcal{C}(G)$ belongs to the same distribution as $\mathcal{C}(G_1), \dots, \mathcal{C}(G_q)$.

4 Two Applications of Curvature Analysis of Graphs

In this section, we discuss two applications for curvature measures in graphs, namely in finding *critical* elementary components and in detecting *change points*.

4.1 Detecting Critical Elementary Components of Networks

Often real-world networks may have the so-called *critical* elementary components (or simply critical components) whose absence alter some significant *non-trivial global* property of these networks. For example, there is a rich history in finding various types of critical components of a network dating back to quantifications of *fault-tolerances* or *redundancies* in electronic circuits or routing networks. Recent examples of practical application of determining critical components in the context of systems biology include quantifying redundancies in biological networks [2, 41, 61] and confirming the existence of central influential neighborhoods in biological networks [3]. Network curvatures can be applied to these kinds of problems by using the curvature measure as the non-trivial global property of a network. We discuss below a simple formalization of these types of problems as used in [22] where edges are elementary components and they can only be added or deleted but *not* both. Thus, in this setting, the basic question is to find a subset (optionally among a set of prescribed edges) whose deletion may change the network curvature significantly.

This question was formalized as the *extremal anomaly detection problem* in [22] in the following manner.

Definition 5 (Extremal Anomaly Detection Problem (EADP) [22]) Given a connected graph $G = (V, E)$, a curvature measure $\mathfrak{C} : G \mapsto \mathbb{R}$, an edge subset $\tilde{E} \subseteq E$ such that $G \setminus \tilde{E}$ is connected, and a real number $\gamma < \mathfrak{C}(G)$ (resp., $\gamma > \mathfrak{C}(G)$) find an edge subset $\hat{E} \subseteq \tilde{E}$ of minimum cardinality such that $\mathfrak{C}(G \setminus \hat{E}) \leq \gamma$ (resp., $\mathfrak{C}(G \setminus \hat{E}) \geq \gamma$).

4.2 Detecting Change Points in Dynamic Networks

Another application similar to that in Section 4.1 is related to change-point detection in dynamic (i.e., time-evolving) networks. Dynamic networks are networks whose elementary components (such as nodes or edges) are added or removed as the network evolves over time. Examples of such networks include biological signal transduction networks with node dynamics, biochemical reaction networks, and dynamic social networks. The *anomaly detection* or *change-point detection* problem for such networks involves finding elementary components whose addition and/or removal alters a significant topological property of the network between two *successive* time steps. There is an extensive history of research works dealing with change-point detection problems over the last several decades in the “non-network” context of time series data [4, 40] with applications to areas such as medical condition monitoring [12, 65], weather change detection [25, 56], and speech recognition [15]. Again using edges as elementary components and the assumption that edges can only be added or deleted but *not* both, a simple formalization of these type of problems under the name “Targeted Anomaly Detection Problem” appeared in [22]. The formalization is as follows.

Definition 6 (Targeted Anomaly Detection Problem (TADP) [22]) Given two connected graphs $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ with $E_2 \subset E_1$ and a curvature measure $\mathfrak{C} : G \mapsto \mathbb{R}$, an edge subset $E_3 \subseteq E_1 \setminus E_2$ of minimum cardinality such that $\mathfrak{C}(G_1 \setminus E_3) = \mathfrak{C}(G_2)$.

For both these applications (i.e., for both the problems EADP and TADP stated in the previous two sub-sections), the authors in [22] prove several algorithmic results for both the cases when \mathfrak{C} is the Gromov curvature and when \mathfrak{C} is the geometric curvature given by Equation (3) with fixed d . Informally, some of the results proved in [22] are as follows:

- ▷ When \mathfrak{C} is the Gromov curvature, it is NP-hard to design a polynomial time algorithm to approximate both EADP and TADP within a factor of cn for some constants $c > 0$, where n is the number of nodes (the hardness result for EADP holds only for the case when $\gamma > \mathfrak{C}$).
- ▷ The following results hold when \mathfrak{C} is the geometric curvature:

- ▷ EADP is NP-hard but admits a non-trivial approximation algorithm when either γ is sufficient larger than \mathcal{C} or γ is not too far below \mathcal{C} .
- ▷ Polynomial time approximation of TADP within a factor of 2 is hard.

5 Conclusion

Notions of curvatures play a fundamental role in physics and mathematics for visualizing higher-dimensional geometric shapes and topological spaces. However, usage of curvature measures for networks is *not* yet very common due to several reasons such as lack of preferred geometric interpretation of networks and lack of experimental evidences that may lead to specific desired curvature properties. In this chapter we have reviewed two curvature measures for networks, namely the Gromov-hyperbolic and the geometric curvature measures, and two motivating applications of these curvature measures, and we hope that this review will act as a stimulator and motivator of further theoretical or empirical research on the exciting interplay between notions of curvatures from network and non-network domains.

Acknowledgments Chatterjee and Dasgupta thankfully acknowledges support from NSF grant IIS-1814931. Albert thankfully acknowledges support from NSF grant IIS-1814405.

References

1. R. Albert, A.-L. Barabási, Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**(1), 47–97 (2002)
2. R. Albert, B. DasGupta, A. Gitter, G. Gürsoy, R. Hegde, P. Pal, G.S. Sivanathan, E.D. Sontag, A new computationally efficient measure of topological redundancy of biological and social networks. *Phys. Rev. E* **84**(3), 036117 (2011)
3. R. Albert, B. DasGupta, N. Mobasher, Topological implications of negative curvature for biological and social networks. *Phys. Rev. E* **89**(3), 032811 (2014)
4. S. Aminikhangahi, D.J. Cook, A survey of methods for time series change point detection. *Knowl. Inf. Syst.* **51**(2), 339–367 (2017)
5. F. Ariaei, M. Lou, E. Jonckere, B. Krishnamachari, M. Zuniga, Curvature of sensor network: clustering coefficient. *EURASIP J. Wirel. Commun. Netw.* 213185 (2008)
6. D.S. Bassett, N.F. Wymbs, M.A. Porter, P.J. Mucha, J.M. Carlson, S.T. Grafton, Dynamic reconfiguration of human brain networks during learning. *PNAS* **108**(18), 7641–7646 (2011)
7. I. Benjamini, Expanders are not hyperbolic. *Isr. J. Math.* **108**, 33–36 (1998)
8. I. Benjamini, O. Schramm, Finite transitive graph embedding into a hyperbolic metric space must stretch or squeeze, in *Geometric Aspects of Functional Analysis* (Springer, Berlin, 2012), pp. 123–126
9. I. Benjamini, C. Hoppen, E. Ofek, P. Pralat, N. Wormald, Geodesics and almost geodesic cycles in random regular graphs. *J. Graph Theory* **66**, 115–136 (2011)
10. M. Berger, *A Panoramic View of Riemannian Geometry* (Springer, Berlin, 2012)
11. E. Bloch, Combinatorial Ricci curvature for polyhedral surfaces and posets. arXiv:1406.4598v1 (2014)

12. M. Bosc, F. Heitz, J.P. Armspach, I. Namer, D. Gounot, L. Rumbach, Automatic change detection in multimodal serial MRI: application to multiple sclerosis lesion evolution. *Neuroimage* **20**(2), 643–656 (2003)
13. M.R. Bridson, A. Haefliger. *Metric Spaces of Non-positive Curvature* (Springer, Berlin, 1999)
14. J. Chalopin, V. Chepoi, F.F. Dragan, G. Ducoffe, A. Mohammed, Y. Vaxès, Fast approximation and exact computation of negative curvature parameters of graphs, in *34th International Symposium on Computational Geometry* (2018)
15. M.F.R. Chowdhury, S.A. Selouani, D. O’Shaughnessy, Bayesian on-line spectral change point detection: a soft computing approach for on-line ASR. *Int. J. Speech Technol.* **15**(1), 5–23 (2011)
16. V. Chepoi, B. Estellon, Packing and covering δ -hyperbolic spaces by balls, in *Lecture Notes in Computer Science 4627*, ed. by M. Charikar, K. Jansen, O. Reingold, J.D.P. Rolim (Springer, Berlin, 2007), pp. 59–73
17. V. Chepoi, F.F. Dragan, B. Estellon, M. Habib, Y. Vaxès, Diameters, centers, and approximating trees of δ -hyperbolic geodesic spaces and graphs, in *24th Annual Symposium on Computational Geometry* (2008), pp. 59–68
18. V. Chepoi, F.F. Dragan, B. Estellon, M. Habib, Y. Vaxès, Y. Xiang, Additive spanners and distance and routing labeling schemes for δ -hyperbolic graphs. *Algorithmica* **62**(3–4), 713–732 (2012)
19. T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, *Introduction to Algorithms* (MIT Press, Cambridge, 2001)
20. B. DasGupta, D. Desai, Complexity of Newman’s community finding approach for social networks. *J. Comput. Syst. Sci.* **79**, 50–67 (2013)
21. B. DasGupta, M. Karpinski, N. Mobasher, F. Yahyanejad, Effect of Gromov-hyperbolicity parameter on cuts and expansions in graphs and some algorithmic implications. *Algorithmica* **80**(2), 772–800 (2018)
22. B. DasGupta, M.V. Janardhanan, F. Yahyanejad, How did the shape of your network change? (On detecting network anomalies via non-local curvatures). *Algorithmica* **82**(7), 1741–1783 (2020)
23. F. de Montgolfier, M. Soto, L. Viennot, Treewidth and hyperbolicity of the internet, in *10th IEEE International Symposium on Networking Computing and Applications* (2011), pp. 25–32
24. R. Duan, S. Pettie, Fast algorithms for (max, min)-matrix multiplication and bottleneck shortest paths, in *20th Annual ACM-SIAM Symposium on Discrete Algorithms* (2009), pp. 384–391
25. J.F. Ducre-Robitaille, L.A. Vincent, G. Boulet, *Comparison of techniques for detection of discontinuities in temperature series*. *Int. J. Climatol.* **23**(9), 1087–1101 (2003)
26. R. Forman, Bochner’s method for cell complexes and combinatorial Ricci curvature. *Discret. Comput. Geom.* **29**(3), 323–374 (2003)
27. H. Fournier, A. Ismail, A. Vigneron, Computing the Gromov hyperbolicity of a discrete metric space. *Inf. Process. Lett.* **115**, 6–8, 576–579 (2015)
28. T.W. Gamelin, R.E. Greene, *Introduction to Topology* (Dover Publications, Mineola, 1999)
29. M.R. Garey, D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (W. H. Freeman, New York, 1979)
30. C. Gavoille, O. Ly, Distance labeling in hyperbolic graphs, in *Lecture Notes in Computer Science 3827*, ed. by X. Deng, D.-Z. Du (Springer, Berlin, 2005), pp. 1071–1079
31. L. Giot, J.S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y.L. Hao, C.E. Ooi, B. Godwin, E. Vitols, G. Vijayadmodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C.A. Stanyon, R.L. Finley, K.P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R.A. Shinkets, M.P. McKenna, J. Chant, J.M. Rothberg, A protein interaction map of *Drosophila melanogaster*. *Science* **302**(5651), 1727–1736 (2003)
32. M. Girvan, M.E.J. Newman, Community structure in social and biological networks. *PNAS* **99**, 7821–7826 (2002)

33. M. Gromov, Hyperbolic groups. *Essays Group Theory* **8**, 75–263 (1987)
34. M. Henle, *A Combinatorial Introduction to Topology* (Dover Publications, Mineola, 1994)
35. E.A. Jonckheere, P. Lohsoonthorn, Geometry of network security. *Am. Control Conf.* **2**, 976–981 (2004)
36. E. Jonckheere, P. Lohsoonthorn, F. Bonahon, Scaled Gromov hyperbolic graphs. *J. Graph Theory* **57**(2), 157–180 (2007)
37. E. Jonckheere, P. Lohsoonthorn, F. Ariaei, Scaled Gromov four-point condition for network graph curvature computation. *Int. Math.* **7**(3), 137–177 (2011)
38. E. Jonckheere, M. Lou, F. Bonahon, Y. Baryshnikov, Euclidean versus hyperbolic congestion in idealized versus experimental networks. *Int. Math.* **7**(1), 1–27 (2011)
39. R. Kannan, P. Tetali, S. Vempala, Markov-chain algorithms for generating bipartite graphs and tournaments. *Random Struct. Algorithm.* **14**, 293–308 (1999)
40. Y. Kawahara, M. Sugiyama, Sequential change-point detection based on direct density-ratio estimation, in *SIAM International Conference on Data Mining* (2009), pp. 389–400
41. B. Kolb, I.Q. Whishaw, *Fundamentals of Human Neuropsychology* (Freeman, New York, 1996)
42. V. Latora, M. Marchior, A measure of centrality based on network efficiency. *New J. Phys.* **9**, 188 (2007)
43. E.A. Leicht, M.E.J. Newman, Community structure in directed networks. *Phys. Rev. Lett.* **100**, 118703 (2008)
44. S. Li, C.M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.-O. Vidalain, J.-D.J. Han, A. Chesneau, T. Hao, D.S. Goldberg, N. Li, M. Martinez, J.-F. Rual, P. Lamesch, L. Xu, M. Tewari, S.L. Wong, L.V. Zhang, G.F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H.W. Gabel, A. Elewa, B. Baumgartner, D.J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S.E. Mango, W.M. Saxton, S. Strome, S. van den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K.C. Gunsalus, J.W. Harper, M.E. Cusick, F.P. Roth, D.E. Hill, M. Vidal, A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540–543 (2004)
45. A. Malyshev, Expanders are order diameter non-hyperbolic. arXiv:1501.07904 (2015)
46. D. Narayan, I. Saniee, Large-scale curvature of networks. *Phys. Rev. E* **84**, 066108 (2011)
47. O. Narayan, I. Saniee, G.H. Tucci, Lack of hyperbolicity in asymptotic Erdős-Rényi sparse random graphs. *Int. Math.* **11**(3), 277–288 (2015)
48. M.E.J. Newman, The structure and function of complex networks. *SIAM Rev.* **45**, 167–256 (2003)
49. M.E.J. Newman, Detecting community structure in networks. *Eur. Phys. J. B* **38**, 321–330 (2004)
50. M.E.J. Newman, *Networks: An Introduction* (Oxford University Press, Oxford, 2010)
51. M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 026113 (2004)
52. M.E.J. Newman, S.H. Strogatz, D.J. Watts, Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**(2), 026118–026134 (2001)
53. Y. Ollivier, A visual introduction to Riemannian curvatures and some discrete generalizations, in *Analysis and Geometry of Metric Measure Spaces: Lecture Notes of the 50th Séminaire de Mathématiques Supérieures*, vol. 56, ed. by G. Dafni, R. John McCann, A. Stancu (2011), pp. 197–219
54. C.H. Papadimitriou, K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity* (Prentice-Hall Inc., Upper Saddle River, 1982)
55. F. Papadopoulos, D. Krioukov, M. Boguna, A. Vahdat, Greedy forwarding in dynamic scale-free networks embedded in hyperbolic metric spaces. *IEEE Conf. Comput. Commun.* 1–9 (2010)
56. J. Reeves, J. Chen, X.L. Wang, R. Lund, Q.Q. Lu, A review and comparison of changepoint detection techniques for climate data. *J. Appl. Meteorol. Climatol.* **46**(6), 900–915 (2007)
57. N. Robertson, P.D. Seymour, Graph minors. I. excluding a forest. *J. Combinatorial Theory Ser. B* **35**(1), 39–61 (1983)

58. J. Roe, Index theory, coarse geometry, and topology of manifolds, in *Conference Board of the Mathematical Sciences Regional Conference, Series 90* (American Mathematical Society, Providence, 1996)
59. A. Samal, R.P. Sreejith, J. Gu, S. Liu, E. Saucan, J. Jost, Comparative analysis of two discretizations of Ricci curvature for complex networks. *Sci. Rep.* **8**, Article number: 8650 (2018)
60. S.S. Shen-Orr, R. Milo, S. Mangan, U. Alon, Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Gen.* **31**, 64–68 (2002)
61. G. Tononi, O. Sporns, G.M. Edelman, Measures of degeneracy and redundancy in biological networks. *PNAS* **96**, 3257–3262 (1999)
62. M. Weber, J. Jost, E. Saucan, Forman-Ricci flow for change detection in large dynamic data sets, in *International Conference on Information and Computational Science* (2016)
63. M. Weber, E. Saucan, J. Jost, Can one see the shape of a network? arXiv:1608.07838 (2016)
64. V.V. Williams, Multiplying matrices faster than Coppersmith-Winograd, in *44th ACM Symposium on Theory of Computing* (2012), pp. 887–898
65. P. Yang, G. Dumont, J.M. Ansermino, Adaptive change detection in heart rate trend monitoring in anesthetized children. *IEEE Trans. Biomed. Eng.* **53**(11), 2211–2219 (2006)

A Frictional Dynamic Thermal Contact Problem with Normal Compliance and Damage



Oanh Chau, Adrien Petrov, Arnaud Heibig, and Manuel Monteiro Marques

Abstract We study a class of non-clamped dynamical problems for visco-elastic materials, the contact condition is modeled by a normal compliance, with friction, damage and heat exchange. The weak formulation leads to a general system defined by a second-order quasi-variational evolution inequality on the displacement field coupled with a nonlinear evolutional inequality on temperature field and a parabolic variational inequality on the damage field. We present and establish an existence and uniqueness result of different fields, by using general results on evolution variational inequalities, with monotone operators and fixed point methods. Then, we present a fully discrete numerical scheme of approximation and derive an error estimate. Finally, various numerical computations are developed.

1 Introduction

Problems involving contact between deformable bodies abound in industry and everyday life. For this reason, a considerable engineering and mathematical literature is devoted to dynamic and quasi-static frictional contact problems, including mathematical modeling, mathematical analysis, numerical analysis and numerical simulations. The study of contact problems for elastic–visco-elastic materials within the mathematical analysis framework was introduced in the early reference

O. Chau (✉)

University of La Réunion, PIMENT EA4518, Saint-Denis, France
e-mail: oanh.chau@univ-reunion.fr

A. Petrov · A. Heibig

University of Lyon, CNRS, INSA of Lyon, Institut Camille Jordan UMR 5208, Villeurbanne, France
e-mail: apetrov@math.univ-lyon1.fr; arnaud.heibig@insa-lyon.fr

M. M. Marques

University of Lisbon, Faculty of Science, Lisbon, Portugal
e-mail: mdmarques@fc.ul.pt

© Springer Nature Switzerland AG 2021

T. M. Rassias, P. M. Pardalos (eds.), *Nonlinear Analysis and Global Optimization*, Springer Optimization and Its Applications 167, https://doi.org/10.1007/978-3-030-61732-5_4

works [5, 8–10]. In these works, numerous types of frictional contact models with nonlinear visco-elastic or elasto-plastic materials were widely studied, in the framework of linearized infinitesimal deformations, using abstract variational inequalities, with monotonicity and convexity.

Further extensions to non-convex contact conditions with non-monotone and possible multi-valued constitutive laws led to the active domain of non-smooth mechanics within the framework of the so-called hemivariational inequalities, for a mathematical as well as mechanical treatment, we refer to [11].

This paper is a continuation work of the results obtained in [3], p. 251. In [3], the authors studied a problem for the quasi-static contact between an elastic–viscoplastic body and an obstacle, the contact was clamped on some part of the boundary and was frictionless, and it was defined by a normal compliance condition with damage. An existence and uniqueness result on displacement and damage fields has been established, and also some numerical approximations and simulations have been presented.

In this work, we study a class of dynamic contact problems with normal compliance condition and damage, with Coulomb’s friction and thermal effects, for visco-elastic material. The novelty here is that we investigate a general long memory material law, depending on time, on the temperature and the damage. Moreover, the evolution of the temperature is described by a general nonlinear equation, involving the gradient of temperature and the velocity of deformation, and the associated boundary condition is defined by an inclusion of sub-differential type in a non-convex framework. Also, the usual clamped condition has been deleted, so that Korn’s inequality cannot be applied any more. The problem appears then semi-coercive and strongly nonlinear due to the frictions. Semi-coercive problems were first studied in [5] for Coulomb’s friction models, where the inertial term of the dynamic process has been used in order to compensate the loss of coerciveness in the a priori estimates. The variational formulation of the mechanical problem leads to a new non-standard model of system defined by a second-order quasi-variational inequality on the displacement field, coupled with one nonlinear inequality for the temperature field and with a variational inequality on the damage field. Then, by using classical results on evolution variational inequalities, with monotone operators and adopting fixed point methods frequently used in [2], we prove an existence and uniqueness of solution on the displacement, damage, and temperature fields.

The paper is organized as follows. In Section 2, we describe the mechanical problem and specify the assumptions on the data to derive the variational formulation, and then we state our main existence and uniqueness result. In Section 3, we give the proof of the claimed result. In Section 4, we introduce a fully discrete approximation scheme and derive an order error estimate under solution regularity assumptions. In Section 5, we present some numerical simulations in order to show the evolution of deformation, of the Von Mises’s norm, of the temperature and the damage in the body.

2 The Contact Problem

In this section, we study a class of thermal contact problems with non-clamped frictional normal compliance condition, for visco-elastic materials. We describe the mechanical problems, list the assumptions on the data, and derive the corresponding variational formulations. Then, we state an existence and uniqueness result on displacement and temperature fields, which we will prove in the next section.

The physical setting is as follows. A visco-elastic body occupies a bounded domain $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$) with a Lipschitz boundary Γ that is partitioned into two disjoint measurable parts, Γ_F and Γ_C . Let $[0, T]$ be the time interval of interest, where $T > 0$. We assume that a volume force of density \mathbf{f}_0 acts in $\Omega \times (0, T)$ and that surface tractions of density \mathbf{f}_F apply on $\Gamma_F \times (0, T)$. The body may come in contact with an obstacle, the foundation, over the potential contact surface Γ_C . The model of the contact is specified by a general sub-differential boundary condition, where thermal effects may occur in the frictional contact with the foundation. Our aim is to describe the dynamic evolution of the body.

Let us recall now some classical notations, see e.g. [5] for further details. We denote by S_d the space of second-order symmetric tensors on \mathbb{R}^d , while “ \cdot ” and $|\cdot|$ will represent the inner product and the Euclidean norm on S_d and \mathbb{R}^d . Let \mathbf{v} denote the unit outer normal on Γ . Everywhere in the sequel, the indices i and j run from 1 to d , summation over repeated indices is implied, and the index that follows a comma represents the partial derivative with respect to the corresponding component of the independent variable. We also use the following notation:

$$H = \left(L^2(\Omega) \right)^d, \quad \mathcal{H} = \{ \boldsymbol{\sigma} = (\sigma_{ij}) \mid \sigma_{ij} = \sigma_{ji} \in L^2(\Omega), 1 \leq i, j \leq d \},$$

$$H_1 = \{ \mathbf{u} \in H \mid \boldsymbol{\varepsilon}(\mathbf{u}) \in \mathcal{H} \}, \quad \mathcal{H}_1 = \{ \boldsymbol{\sigma} \in \mathcal{H} \mid \text{Div } \boldsymbol{\sigma} \in H \}.$$

Here, $\boldsymbol{\varepsilon} : H_1 \longrightarrow \mathcal{H}$ and $\text{Div} : \mathcal{H}_1 \longrightarrow H$ are the deformation and the divergence operators, respectively, defined by

$$\boldsymbol{\varepsilon}(\mathbf{u}) = (\varepsilon_{ij}(\mathbf{u})), \quad \varepsilon_{ij}(\mathbf{u}) = \frac{1}{2}(u_{i,j} + u_{j,i}), \quad \text{Div } \boldsymbol{\sigma} = (\sigma_{ij,j}).$$

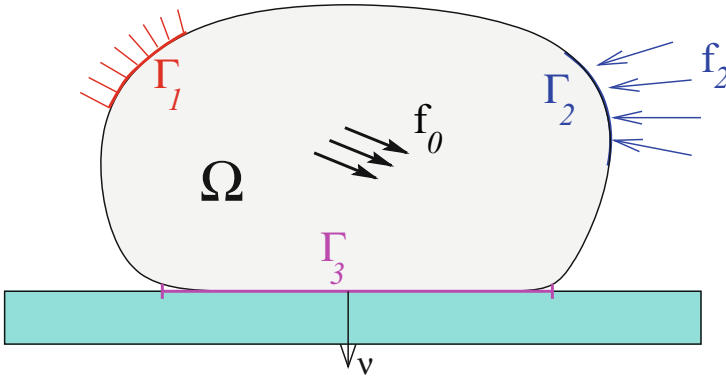
The spaces H , \mathcal{H} , H_1 , and \mathcal{H}_1 are real Hilbert spaces endowed with the canonical inner products given by

$$(\mathbf{u}, \mathbf{v})_H = \int_{\Omega} u_i v_i dx, \quad (\boldsymbol{\sigma}, \boldsymbol{\tau})_{\mathcal{H}} = \int_{\Omega} \sigma_{ij} \tau_{ij} dx,$$

$$(\mathbf{u}, \mathbf{v})_{H_1} = (\mathbf{u}, \mathbf{v})_H + (\boldsymbol{\varepsilon}(\mathbf{u}), \boldsymbol{\varepsilon}(\mathbf{v}))_H, \quad (\boldsymbol{\sigma}, \boldsymbol{\tau})_{\mathcal{H}_1} = (\boldsymbol{\sigma}, \boldsymbol{\tau})_{\mathcal{H}} + (\text{Div } \boldsymbol{\sigma}, \text{Div } \boldsymbol{\tau})_H.$$

We recall that C denotes the class of continuous functions; C^m , $m \in \mathbb{N}^*$ the set of m times continuously differentiable functions; and $W^{m,p}$, $m \in \mathbb{N}$, $1 \leq p \leq +\infty$ the classical Sobolev spaces.

Now, we consider a visco-elastic body which occupies a bounded domain $\Omega \subset \mathbb{R}^d$ ($d = 1, 2, 3$) with a Lipschitz boundary Γ that is partitioned into two disjoint measurable parts, Γ_F and Γ_C . Let $[0, T]$ be the time interval of interest, where $T > 0$. We assume that a volume force of density f_0 acts in $\Omega \times (0, T)$ and that surface tractions of density f_F apply on $\Gamma_F \times (0, T)$. The body may come in contact with an obstacle, the foundation, over the potential contact surface Γ_C , see figure below.



The mathematical contact mechanics
 $\text{meas}(\Gamma_1) = 0; \quad \Gamma_2 = \Gamma_F; \quad \Gamma_3 = \Gamma_C; \quad f_2 = f_F.$

To continue, the mechanical problem is then formulated as follows.

Problem Q: Find a displacement field $\mathbf{u} : (0, T) \times \Omega \rightarrow \mathbb{R}^d$, a stress field $\boldsymbol{\sigma} : (0, T) \times \Omega \rightarrow S_d$, a temperature field $\theta : (0, T) \times \Omega \rightarrow \mathbb{R}_+$, and a damage field $\alpha : (0, T) \times \Omega \rightarrow \mathbb{R}$ such that for a.e. $t \in (0, T)$:

$$\begin{cases} \boldsymbol{\sigma}(t) = \mathcal{A}(t)\boldsymbol{\varepsilon}(\dot{\mathbf{u}}(t)) + \mathcal{G}(t)(\boldsymbol{\varepsilon}(\mathbf{u}(t)), \alpha(t)) + \int_0^t \mathcal{B}(t-s)(\boldsymbol{\varepsilon}(\mathbf{u}(s)), \alpha(s)) ds \\ \quad + C_e(t, \theta(t)) \quad \text{in } \Omega; \end{cases} \tag{1}$$

$$\ddot{\mathbf{u}}(t) = \text{Div } \boldsymbol{\sigma}(t) + \mathbf{f}_0(t) \quad \text{in } \Omega; \tag{2}$$

$$\boldsymbol{\sigma}(t)\nu = \mathbf{f}_F(t) \quad \text{on } \Gamma_F; \tag{3}$$

$$\sigma_\nu(t) = -p_\nu(t, u_\nu(t) - g(t)) \quad \text{on } \Gamma_C; \tag{4}$$

$$\left\{ \begin{array}{l} |\sigma_\tau(t)| \leq p_\tau(t, u_v(t) - g(t)) : \\ |\sigma_\tau(t)| < p_\tau(t, u_v(t) - g(t)) \implies \dot{\mathbf{u}}_\tau(t) = 0; \\ |\sigma_\tau(t)| = p_\tau(t, u_v(t) - g(t)) \implies \dot{\mathbf{u}}_\tau(t) = -\lambda \sigma_\tau(t), \\ \text{for some } \lambda \geq 0; \end{array} \right. \quad \text{on } \Gamma_C; \quad (5)$$

$$[\dot{\alpha}(t) - \gamma \Delta \alpha(t) - \phi_d(\sigma(t), \boldsymbol{\varepsilon}(\mathbf{u}(t)), \alpha(t))](\xi - \alpha(t)) \geq 0 \quad \text{in } \Omega, \quad \forall \xi \in [0, 1]; \quad (6)$$

$$0 \leq \alpha(t) \leq 1 \quad \text{in } \Omega; \quad (7)$$

$$\frac{\partial \alpha}{\partial \nu}(t) = 0 \quad \text{on } \Gamma; \quad (8)$$

$$\dot{\theta}(t) - \operatorname{div}(\mathcal{K}_c(t, \nabla \theta(t))) = D_e(t, \varepsilon(\dot{\mathbf{u}}(t)), \theta(t)) + q(t) \quad \text{in } \Omega; \quad (9)$$

$$- \mathcal{K}_c(t, \mathbf{x}, \nabla \theta(t, \mathbf{x})) \nu := \Xi(t, \mathbf{x}, \theta(t, \mathbf{x})) \in \partial \varphi(t, \mathbf{x}, \theta(t, \mathbf{x})) \quad \text{a.e. } \mathbf{x} \in \Gamma_C; \quad (10)$$

$$\theta(t) = 0 \quad \text{on } \Gamma_F; \quad (11)$$

$$\mathbf{u}(0) = \mathbf{u}_0; \quad \dot{\mathbf{u}}(0) = \mathbf{v}_0; \quad \alpha(0) = \alpha_0; \quad \theta(0) = \theta_0 \quad \text{in } \Omega. \quad (12)$$

Equation (1) is the Kelvin Voigt's long memory thermo-visco-elastic constitutive law of the body including the influence of the damage variable. Here, σ is the stress tensor, \mathcal{A} denotes the viscosity operator with, $\mathcal{A}(t)\boldsymbol{\tau} = \mathcal{A}(t, \cdot, \boldsymbol{\tau})$ is some function defined on Ω , and \mathcal{G} is the elastic operator depending on the linearized strain tensor $\boldsymbol{\varepsilon}(\mathbf{u})$ of infinitesimal deformations and on the damage α , with $\mathcal{G}(t)(\boldsymbol{\tau}, \alpha) = \mathcal{G}(t, \cdot, \boldsymbol{\tau}, \alpha)$ is some function defined on Ω . For example,

$$\mathcal{G}(t)(\boldsymbol{\tau}, \alpha) = \mathcal{G}^0(t)\boldsymbol{\tau} - \alpha C_{da}(t) \quad \text{in } \Omega,$$

where $\mathcal{G}^0(t)\boldsymbol{\tau} = \mathcal{G}^0(t, \cdot, \boldsymbol{\tau})$ is some time-dependent elastic tensor function independent on the damage, defined on Ω , and $C_{da}(t)$ is some time-dependent damage tensor. The term $\mathcal{B}(t)(\boldsymbol{\tau}, \alpha) = \mathcal{B}(t, \cdot, \boldsymbol{\tau}, \alpha)$ represents the relaxation tensor time depending on the linearized strain tensor and the damage, defined on Ω . And the last tensor $C_e(t, \theta) := C_e(t, \cdot, \theta)$ denotes the thermal expansion tensor depending on time and temperature, defined on Ω . For example,

$$C_e(t, \theta) := -\theta C_{exp}(t) \quad \text{in } \Omega,$$

where

$$C_{exp}(t) := (c_{ij}(t, \cdot))$$

is some time-dependent expansion tensor defined on Ω , with $c_{ij} \in L^\infty((0, T) \times \Omega)$.

The model in (2) is the dynamic equation of motion where the mass density $\varrho \equiv 1$. Equation (3) is the traction boundary condition.

On the contact surface, the general relation (4) represents the *normal compliance* contact condition, where σ_v denotes the normal stress, u_v is the normal displacement, g is the gap between the contact surface and the foundation, and p_v is some normal compliance function defined on $(0, T) \times \Gamma_C \times \mathbb{R}$ with the convention that $p_v(t, r) = p_v(t, \cdot, r)$ denotes some function defined on Γ_C , for a.e. $t \in (0, T)$, for all $r \in \mathbb{R}$. The term $u_v - g$ represents, when it is positive, the penetration of the surface asperities into the foundation.

For example, for a.e. $t \in (0, T)$,

$$p_v(t, \cdot, r) = c_v(t, \cdot) r_+ \quad \text{on } \Gamma_C, \quad \forall r \in \mathbb{R}.$$

In this formula, the normal stress is proportional to the penetration, with some positive coefficient c_v defined on $(0, T) \times \Gamma_C$, which is related to the hardness of the foundation.

Equation (5) represents a general version of Coulomb's dry friction law, where σ_τ is the tangential stress, p_τ is the friction bound measuring the maximal frictional resistance defined on $(0, T) \times \Gamma_C \times \mathbb{R}$, and $\dot{\mathbf{u}}_\tau$ is the tangential velocity. Recall that $p_\tau(t, r) = p_\tau(t, \cdot, r)$ is some function defined on Γ_C , for a.e. $t \in (0, T)$, for all $r \in \mathbb{R}$.

For example, for a.e. $t \in (0, T)$,

$$p_\tau(t, \cdot, r) = \mu_\tau(t, \cdot) c_v(t, \cdot) r_+ \quad \text{on } \Gamma_C, \quad \forall r \in \mathbb{R},$$

where the friction bound is proportional to the normal stress with some positive coefficient of friction μ_τ defined on $(0, T) \times \Gamma_C$.

Following Frémond [6, 7], the damage function α represents the percentage of the safe part or undamaged part, $\alpha = 1$ means that the body is undamaged, and $\alpha = 0$ says that the body is completely damaged. The evolution of the microscopic cracks responsible for the damage is described by the parabolic differential inclusion (6) of the damage function α satisfying $0 \leq \alpha \leq 1$, where γ is a positive constant and ϕ_d is a given constitutive function which describes damage source in the system. The inequality (6) means

$$\alpha(t) = 1 \implies \dot{\alpha}(t) - \gamma \Delta \alpha(t) - \phi_d(\boldsymbol{\sigma}(t), \boldsymbol{\varepsilon}(\mathbf{u}(t)), \alpha(t)) \leq 0;$$

and

$$\alpha(t) \in (0, 1) \implies \dot{\alpha}(t) - \gamma \Delta \alpha(t) - \phi_d(\boldsymbol{\sigma}(t), \boldsymbol{\varepsilon}(\mathbf{u}(t)), \alpha(t)) = 0;$$

and

$$\alpha(t) = 0 \implies \dot{\alpha}(t) - \gamma \Delta \alpha(t) - \phi_d(\boldsymbol{\sigma}(t), \boldsymbol{\varepsilon}(\mathbf{u}(t)), \alpha(t)) \geq 0.$$

Equation (8) represents the homogeneous Neumann boundary condition for the damage field, see e.g. [3], p. 241.

The differential equation (9) provides the evolution of the temperature field. There $\mathcal{K}_c(t, \nabla\theta) := \mathcal{K}_c(t, \cdot, \nabla\theta)$ is some nonlinear time-depending function of the temperature gradient $\nabla\theta$, which is defined on Ω . For example, denote by

$$K_c(t, \cdot) := (k_{ij}(t, \cdot))$$

the thermal conductivity tensor defined on Ω , we could consider

$$\mathcal{K}_c(t, \cdot, \nabla\theta) = K_c(t, \cdot) \nabla\theta.$$

In the second member, $q(t)$ denotes the density of volume heat sources, whereas

$$D_e(t, \boldsymbol{\varepsilon}(\dot{\mathbf{u}}(t)), \theta(t)) := D_e(t, \cdot, \boldsymbol{\varepsilon}(\dot{\mathbf{u}}(t)), \theta(t))$$

is the deformation-viscosity heat, which is a nonlinear function defined on Ω and which represents the heat generated by the velocity of deformation (viscosity) and may depend on the temperature.

Example 1

$$D_e(t, \boldsymbol{\varepsilon}(\dot{\mathbf{u}}(t)), \theta(t)) = -C_{exp}(t) : \boldsymbol{\varepsilon}(\dot{\mathbf{u}}(t)) = -c_{ij}(t, \cdot) \varepsilon_{ij}(\dot{\mathbf{u}}(t)). \tag{13}$$

Example 2

$$D_e(t, \boldsymbol{\varepsilon}(\dot{\mathbf{u}}(t)), \theta(t)) = -\theta(t, \cdot) d_e(t, \cdot), \tag{14}$$

with some coefficient $d_e \in L^\infty((0, T) \times \Omega, \mathbb{R}^+)$;

Example 3

$$D_e(t, \boldsymbol{\varepsilon}(\dot{\mathbf{u}}(t)), \theta(t)) = -C_{exp}(t) : \boldsymbol{\varepsilon}(\dot{\mathbf{u}}(t)) - \theta(t, \cdot) d_e(t, \cdot). \tag{15}$$

By assuming the variation of $\theta(t)$ small enough, then the heat function $D_e(t, \boldsymbol{\varepsilon}(\dot{\mathbf{u}}(t)), \theta(t))$ may be considered as a formula which is independent of the temperature.

The associated temperature boundary condition is given by (10) and (11), where \mathcal{E} and φ are some functions defined on $(0, T) \times \Gamma_C \times \mathbb{R}$. Here,

$$\partial\varphi(t, \mathbf{x}, r) := \partial\varphi(t, \mathbf{x}, \cdot)(r), \quad \forall(t, \mathbf{x}, r) \in (0, T) \times \Gamma_C \times \mathbb{R}$$

denotes the sub-differential on the third variable of φ in the locally Lipschitz framework.

We recall that for a locally Lipschitz function $G : \mathbb{R} \rightarrow \mathbb{R}$, at any point $a \in \mathbb{R}$ and for any vector $d \in \mathbb{R}$, we can define the following directional derivative with respect to d :

$$\overline{\lim}_{\tau \rightarrow 0^+} \frac{G(a + \tau d) - G(a)}{\tau} := G^0(a; d). \quad (16)$$

We have for all $a, d \in \mathbb{R}$, for all $\xi \in \partial G(a)$:

$$G^0(a; d) \geq \xi d$$

and

$$|G^0(a; d)| \leq |G^0(a)| \times |d|, \quad |\xi| \leq |G^0(a)|,$$

where

$$\overline{\lim}_{h \rightarrow 0, h \neq 0} \frac{G(a + h) - G(a)}{h} := G^0(a).$$

In the case where G is convex on \mathbb{R} , we have

$$G^0(a; d) = \begin{cases} G'_r(a)d & \text{if } d > 0 \\ G'_l(a)d & \text{if } d < 0 \\ 0 & \text{if } d = 0, \end{cases}$$

and

$$G^0(a) = \max\{G'_r(a), G'_l(a)\},$$

where G'_r and G'_l denote the right side and left side derivatives, respectively.

In the sequel, for a.e. $(t, \mathbf{x}) \in (0, T) \times \Gamma_c$, for all $(r, s) \in \mathbb{R}^2$, we use the notation

$$\varphi^0(t, \mathbf{x}, r; s) := [\varphi(t, \mathbf{x}, \cdot)]^0(r; s),$$

and

$$\varphi^0(t, \mathbf{x}, r) := [\varphi(t, \mathbf{x}, \cdot)]^0(r).$$

Taking the previous example for \mathcal{H}_c , we have

$$\mathcal{H}_c(t, \mathbf{x}, \nabla\theta) v = k_{ij}(t, \mathbf{x}) \frac{\partial \theta}{\partial x_j} v_i.$$

Let us consider, for example,

$$\varphi(t, \mathbf{x}, r) := \frac{1}{2} k_e(t, \mathbf{x})(r - \theta_R(t, \mathbf{x}))^2, \quad \forall (t, \mathbf{x}, r) \in (0, T) \times \Gamma_C \times \mathbb{R}, \quad (17)$$

where θ_R is the temperature of the foundation, and k_e is the heat exchange coefficient between the body and the obstacle. We obtain

$$\mathcal{E}(t, \mathbf{x}, r) = \partial\varphi(t, \mathbf{x}, r) = k_e(t, \mathbf{x})(r - \theta_R(t, \mathbf{x})), \quad (t, \mathbf{x}, r) \in (0, T) \times \Gamma_C \times \mathbb{R}.$$

Finally, the data in \mathbf{u}_0 , \mathbf{v}_0 , α_0 , and θ_0 in (12) represent the initial displacement, velocity, damage, and temperature, respectively.

In view to derive the variational formulation of the mechanical problems (1)–(12), let us first precise the functional framework. Let

$$V = H_1$$

be the admissible displacement space, endowed with the inner product given by

$$(\mathbf{u}, \mathbf{v})_V = (\boldsymbol{\varepsilon}(\mathbf{u}), \boldsymbol{\varepsilon}(\mathbf{v}))_{\mathcal{H}} + (\mathbf{u}, \mathbf{v})_H \quad \forall \mathbf{u}, \mathbf{v} \in V,$$

and let $\|\cdot\|_V$ be the associated norm, i.e.

$$\|\mathbf{v}\|_V^2 = \|\boldsymbol{\varepsilon}(\mathbf{v})\|_{\mathcal{H}}^2 + \|\mathbf{v}\|_H^2 \quad \forall \mathbf{v} \in V.$$

Therefore, $(V, \|\cdot\|_V)$ is a real Hilbert space, where the norm $\|\cdot\|_V$ is equivalent to $\|\cdot\|_{(H^1(\Omega))^d}$.

Let

$$E = \{\eta \in H^1(\Omega), \eta = 0 \text{ on } \Gamma_F\}$$

be the admissible temperature space, endowed with the canonical inner product of $H^1(\Omega)$.

By the Sobolev's trace theorem, there exists a constant $c_0 > 0$ depending only on Ω , and Γ_C such that

$$\|\mathbf{v}\|_{(L^2(\Gamma_C))^d} \leq c_0 \|\mathbf{v}\|_V, \quad \forall \mathbf{v} \in V; \quad \text{and} \quad \|\eta\|_{L^2(\Gamma_C)} \leq c_0 \|\eta\|_E, \quad \forall \eta \in E. \quad (18)$$

Next, we denote the set of admissible damage fields by

$$\mathcal{H}_{da} = \{\xi \in H^1(\Omega), \frac{\partial \xi}{\partial \nu} = 0 \text{ on } \Gamma, 0 \leq \xi \leq 1 \text{ a.e. in } \Omega\}.$$

We use here two Gelfand evolution triples (see e.g. [12], pp. 416) given by

$$V \subset H \equiv H' \subset V', \quad E \subset L^2(\Omega) \equiv (L^2(\Omega))' \subset E',$$

where the inclusions are dense and continuous.

In the study of the mechanical problems (1)–(12), we assume that the viscosity operator $\mathcal{A} : (0, T) \times \Omega \times S_d \longrightarrow S_d$ satisfies

$$\left\{ \begin{array}{l} \text{(i) } \mathcal{A}(\cdot, \cdot, \boldsymbol{\tau}) \text{ is measurable on } (0, T) \times \Omega, \forall \boldsymbol{\tau} \in S_d; \\ \text{(ii) } \mathcal{A}(t, \mathbf{x}, \cdot) \text{ is continuous on } S_d \text{ for a.e. } (t, \mathbf{x}) \in (0, T) \times \Omega; \\ \text{(iii) there exists } m_{\mathcal{A}} > 0 \text{ such that} \\ \quad (\mathcal{A}(t, \mathbf{x}, \boldsymbol{\tau}_1) - \mathcal{A}(t, \mathbf{x}, \boldsymbol{\tau}_2)) \cdot (\boldsymbol{\tau}_1 - \boldsymbol{\tau}_2) \geq m_{\mathcal{A}} |\boldsymbol{\tau}_1 - \boldsymbol{\tau}_2|^2, \\ \quad \forall \boldsymbol{\tau}_1, \boldsymbol{\tau}_2 \in S_d, \text{ for a.e. } (t, \mathbf{x}) \in (0, T) \times \Omega; \\ \text{(iv) there exists } c_0^{\mathcal{A}} \in L^2((0, T) \times \Omega; \mathbb{R}^+), c_1^{\mathcal{A}} > 0 \text{ such that} \\ \quad |\mathcal{A}(t, \mathbf{x}, \boldsymbol{\tau})| \leq c_0^{\mathcal{A}}(t, \mathbf{x}) + c_1^{\mathcal{A}} |\boldsymbol{\tau}|, \\ \quad \forall \boldsymbol{\tau} \in S_d, \text{ for a.e. } (t, \mathbf{x}) \in (0, T) \times \Omega. \end{array} \right. \quad (19)$$

Here, recall that for every $t \in (0, T)$ and $\boldsymbol{\tau} \in S_d$, we write by $\mathcal{A}(t) = \mathcal{A}(t, \cdot, \cdot)$ a functional which is defined on $\Omega \times S_d$ and $\mathcal{A}(t) \boldsymbol{\tau} = \mathcal{A}(t, \cdot, \boldsymbol{\tau})$ some function defined on Ω .

We suppose that the elasticity operator $\mathcal{G} : (0, T) \times \Omega \times S_d \times \mathbb{R} \longrightarrow S_d$ satisfies

$$\left\{ \begin{array}{l} \text{(i) } \mathcal{G}(\cdot, \cdot, \boldsymbol{\tau}, \lambda) \text{ is measurable on } (0, T) \times \Omega, \forall \boldsymbol{\tau} \in S_d, \forall \lambda \in \mathbb{R}; \\ \text{(ii) there exists } L_{\mathcal{G}} > 0 \text{ such that} \\ \quad |\mathcal{G}(t, \mathbf{x}, \boldsymbol{\tau}_1, \lambda_1) - \mathcal{G}(t, \mathbf{x}, \boldsymbol{\tau}_2, \lambda_2)| \leq L_{\mathcal{G}} (|\boldsymbol{\tau}_1 - \boldsymbol{\tau}_2| + |\lambda_1 - \lambda_2|) \\ \quad \forall \boldsymbol{\tau}_1, \boldsymbol{\tau}_2 \in S_d, \forall \lambda_1, \lambda_2 \in \mathbb{R}, \text{ a.e. } (t, \mathbf{x}) \in (0, T) \times \Omega; \\ \text{(iii) there exists } c_0^{\mathcal{G}} \in L^2((0, T) \times \Omega; \mathbb{R}^+), c_1^{\mathcal{G}} \geq 0, c_2^{\mathcal{G}} \geq 0 \text{ such that} \\ \quad |\mathcal{G}(t, \mathbf{x}, \boldsymbol{\tau}, \lambda)| \leq c_0^{\mathcal{G}}(t, \mathbf{x}) + c_1^{\mathcal{G}} |\boldsymbol{\tau}| + c_2^{\mathcal{G}} |\lambda|, \\ \quad \forall \boldsymbol{\tau} \in S_d, \forall \lambda \in \mathbb{R}, \text{ a.e. } (t, \mathbf{x}) \in (0, T) \times \Omega; \\ \text{(iv) the partial derivatives with respect to the first, third, and fourth} \\ \quad \text{variables of } \mathcal{G} \text{ exist and are bounded.} \end{array} \right. \quad (20)$$

We put again $\mathcal{G}(t)(\boldsymbol{\tau}, \lambda) = \mathcal{G}(t, \cdot, \boldsymbol{\tau}, \lambda)$ some function defined on Ω for every $t \in (0, T)$, $\boldsymbol{\tau} \in S_d$, $\lambda \in \mathbb{R}$.

The relaxation tensor $\mathcal{B} : (0, T) \times \Omega \times S_d \times \mathbb{R} \longrightarrow S_d$ satisfies

$$\left\{ \begin{array}{l} \text{(i) } \mathcal{B}(\cdot, \cdot, \boldsymbol{\tau}, \lambda) \in L^\infty((0, T) \times \Omega; S_d), \forall \boldsymbol{\tau} \in S_d, \forall \lambda \in \mathbb{R}; \\ \text{(ii) there exists } L_{\mathcal{B}} > 0 \text{ such that} \\ \quad |\mathcal{B}(t, \mathbf{x}, \boldsymbol{\tau}_1, \lambda_1) - \mathcal{B}(t, \mathbf{x}, \boldsymbol{\tau}_2, \lambda_2)| \leq L_{\mathcal{B}} (|\boldsymbol{\tau}_1 - \boldsymbol{\tau}_2| + |\lambda_1 - \lambda_2|) \\ \quad \forall \boldsymbol{\tau}_1, \boldsymbol{\tau}_2 \in S_d, \forall \lambda_1, \lambda_2 \in \mathbb{R}, \text{ a.e. } (t, \mathbf{x}) \in (0, T) \times \Omega; \\ \text{(iii) the partial derivative with respect to the first variable of} \\ \quad \mathcal{B} \text{ exists and is bounded.} \end{array} \right. \quad (21)$$

The body forces and surface tractions satisfy the regularity conditions:

$$\mathbf{f}_0 \in W^{1,2}(0, T; H), \quad \mathbf{f}_F \in W^{1,2}(0, T; L^2(\Gamma_F)^d). \quad (22)$$

The gap function $g : (0, T) \times \Gamma_C \longrightarrow \mathbf{R}^+$ verifies

$$\left\{ \begin{array}{l} \text{(i) } g \in L^\infty((0, T) \times \Gamma_C; \mathbf{R}^+); \\ \text{(ii) the partial derivative with respect to the first variable of } \\ \quad g \text{ exists and is bounded.} \end{array} \right. \quad (23)$$

The thermal expansion tensor $C_e : (0, T) \times \Omega \times \mathbb{R} \longrightarrow S_d$ verifies

$$\left\{ \begin{array}{l} \text{(i) } C_e(\cdot, \cdot, \vartheta) \text{ is measurable on } (0, T) \times \Omega, \forall \vartheta \in \mathbb{R}; \\ \text{(ii) there exists } L_e > 0 \text{ such that} \\ \quad |C_e(t, \mathbf{x}, \vartheta_1) - C_e(t, \mathbf{x}, \vartheta_2)| \leq L_e |\vartheta_1 - \vartheta_2| \\ \quad \forall \vartheta_1, \vartheta_2 \in \mathbb{R}, \text{ a.e. } (t, \mathbf{x}) \in (0, T) \times \Omega; \\ \text{(iii) there exists } c_0^{C_e} \in L^\infty((0, T) \times \Omega; \mathbf{R}^+), c_1^{C_e} \geq 0 \text{ such that} \\ \quad |C_e(t, \mathbf{x}, \vartheta)| \leq c_0^{C_e}(t, \mathbf{x}) + c_1^{C_e} |\vartheta|, \\ \quad \forall \vartheta \in \mathbb{R}, \text{ for a.e. } (t, \mathbf{x}) \in (0, T) \times \Omega; \\ \text{(iv) the partial derivatives with respect to the first and third variables} \\ \quad \text{of } C_e \text{ exist and are bounded.} \end{array} \right. \quad (24)$$

Here, we use the notation $C_e(t, \vartheta) = C_e(t, \cdot, \vartheta)$ some function defined on Ω , for all $t \in (0, T)$ and $\vartheta \in \mathbb{R}$.

The normal compliance function $p_v : (0, T) \times \Gamma_C \times \mathbb{R} \longrightarrow \mathbb{R}_+$ satisfies

$$\left\{ \begin{array}{l} \text{(i) there exists } L_v > 0 \text{ such that} \\ \quad |p_v(t, \mathbf{x}, r_1) - p_v(t, \mathbf{x}, r_2)| \leq L_v |r_1 - r_2|, \\ \quad \forall r_1, r_2 \in \mathbb{R}, \text{ a.e. } (t, \mathbf{x}) \in (0, T) \times \Gamma_C; \\ \text{(ii) } p_v(\cdot, \cdot, r) \text{ is Lebesgue measurable on } (0, T) \times \Gamma_C, \forall r \in \mathbb{R}; \\ \text{(iii) the mapping } p_v(\cdot, \cdot, r) = 0, \forall r \leq 0; \\ \text{(iv) the partial derivatives with respect to the first and third variables} \\ \quad \text{of } p_v \text{ exist and are bounded.} \end{array} \right. \quad (25)$$

The friction bound function $p_\tau : (0, T) \times \Gamma_C \times \mathbb{R} \longrightarrow \mathbb{R}_+$ satisfies

$$\left\{ \begin{array}{l} \text{(i) there exists } L_\tau > 0 \text{ such that} \\ \quad |p_\tau(t, \mathbf{x}, r_1) - p_\tau(t, \mathbf{x}, r_2)| \leq L_\tau |r_1 - r_2|, \\ \quad \forall r_1, r_2 \in \mathbb{R}, \text{ a.e. } (t, \mathbf{x}) \in (0, T) \times \Gamma_C; \\ \text{(ii) } p_\tau(\cdot, \cdot, r) \text{ is Lebesgue measurable on } (0, T) \times \Gamma_C, \forall r \in \mathbb{R}; \\ \text{(iii) the mapping } p_\tau(\cdot, \cdot, r) = 0, \forall r \leq 0. \end{array} \right. \quad (26)$$

The damage source $\phi_d : \Omega \times S_d \times S_d \times [0, 1] \longrightarrow \mathbb{R}$ verifies

$$\left\{ \begin{array}{l} \text{(i) there exists } L_\phi > 0 \text{ such that} \\ \quad |\phi_d(\mathbf{x}, \boldsymbol{\sigma}_1, \boldsymbol{\varepsilon}_1, \xi_1) - \phi_d(\mathbf{x}, \boldsymbol{\sigma}_2, \boldsymbol{\varepsilon}_2, \xi_2)| \leq L_\phi (|\boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_2| + |\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2| + |\xi_1 - \xi_2|), \\ \quad \forall \boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2, \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2 \in S_d, \forall \xi_1, \xi_2 \in [0, 1], \text{ a.e. } \mathbf{x} \in \Omega; \\ \text{(ii) } \phi_d(\cdot, \boldsymbol{\sigma}, \boldsymbol{\varepsilon}, \xi) \text{ is Lebesgue measurable function on } \Omega, \\ \quad \forall \boldsymbol{\sigma}, \boldsymbol{\varepsilon} \in S_d, \forall \xi \in [0, 1]; \\ \text{(iii) } \phi_d(\cdot, 0, 0, 0) \in L^2(\Omega). \end{array} \right. \quad (27)$$

We assume that the nonlinear function $\mathcal{H}_c : (0, T) \times \Omega \times \mathbb{R}^d \longrightarrow \mathbb{R}^d$ satisfies

$$\left\{ \begin{array}{l} \text{(i) } \mathcal{H}_c(\cdot, \cdot, \xi) \text{ is measurable on } (0, T) \times \Omega, \forall \xi \in \mathbb{R}^d; \\ \text{(ii) } \mathcal{H}_c(t, \mathbf{x}, \cdot) \text{ is continuous on } \mathbb{R}^d, \text{ a.e. } (t, \mathbf{x}) \in (0, T) \times \Omega; \\ \text{(iii) there exists } c_0^{\mathcal{H}_c} \in L^2((0, T) \times \Omega; \mathbb{R}^+), c_1^{\mathcal{H}_c} \geq 0, \text{ such that} \\ \quad |\mathcal{H}_c(t, \mathbf{x}, \xi)| \leq c_0^{\mathcal{H}_c}(t, \mathbf{x}) + c_1^{\mathcal{H}_c} |\xi|, \\ \quad \forall \xi \in \mathbb{R}^d, \text{ a.e. } (t, \mathbf{x}) \in (0, T) \times \Omega; \\ \text{(iv) there exists } m_{\mathcal{H}_c} > 0 \text{ such that} \\ \quad (\mathcal{H}_c(t, \mathbf{x}, \xi_1) - \mathcal{H}_c(t, \mathbf{x}, \xi_2)) \cdot (\xi_1 - \xi_2) \geq m_{\mathcal{H}_c} |\xi_1 - \xi_2|^2, \\ \quad \forall \xi_1, \xi_2 \in \mathbb{R}^d, \text{ a.e. } (t, \mathbf{x}) \in (0, T) \times \Omega; \\ \text{(v) there exists } n_{\mathcal{H}_c} > 0 \text{ such that } \mathcal{H}_c(t, \mathbf{x}, \xi) \cdot \xi \geq n_{\mathcal{H}_c} |\xi|^2, \\ \quad \forall \xi \in \mathbb{R}^d, \text{ a.e. } (t, \mathbf{x}) \in (0, T) \times \Omega. \end{array} \right. \quad (28)$$

We suppose that the deformation-viscosity heat function $D_e : (0, T) \times \Omega \times S_d \times \mathbb{R} \longrightarrow \mathbb{R}$ satisfies

$$\left\{ \begin{array}{l}
\text{(i) } D_e(\cdot, \cdot, \boldsymbol{\tau}, \vartheta) \text{ is measurable on } (0, T) \times \Omega, \forall (\boldsymbol{\tau}, \vartheta) \in S_d \times \mathbb{R}; \\
\text{(ii) the function } D_e(t, \mathbf{x}, \cdot, \cdot) \text{ is Lipschitz continuous on } S_d \times \mathbb{R}, \\
\text{ i.e. } \exists D_V > 0, \exists D_T > 0 : \\
\quad |D_e(t, \mathbf{x}, \boldsymbol{\tau}_1, \vartheta_1) - D_e(t, \mathbf{x}, \boldsymbol{\tau}_2, \vartheta_2)| \leq D_V |\boldsymbol{\tau}_1 - \boldsymbol{\tau}_2| + D_T |\vartheta_1 - \vartheta_2|, \\
\quad \forall (\boldsymbol{\tau}_1, \vartheta_1), (\boldsymbol{\tau}_2, \vartheta_2) \in S_d \times \mathbb{R}, \text{ for a.e. } (t, \mathbf{x}) \in (0, T) \times \Omega; \\
\text{(iii) } D_e(\cdot, \cdot, 0_{S_d}, 0) \in L^\infty((0, T) \times \Omega); \\
\text{(iv) } (D_e(t, \mathbf{x}, \boldsymbol{\tau}, \vartheta_1) - D_e(t, \mathbf{x}, \boldsymbol{\tau}, \vartheta_2)) (\vartheta_1 - \vartheta_2) \leq 0, \\
\quad \forall \boldsymbol{\tau} \in S_d, \forall \vartheta_1, \vartheta_2 \in \mathbb{R}, \text{ a.e. } (t, \mathbf{x}) \in (0, T) \times \Omega.
\end{array} \right. \quad (29)$$

We notice that these conditions are verified in examples (13)–(15).
The heat sources density verifies

$$q \in L^2(0, T; L^2(\Omega)). \quad (30)$$

We suppose that the nonlinear functions $\mathcal{E}, \varphi : (0, T) \times \Gamma_C \times \mathbb{R} \rightarrow \mathbb{R}$ satisfy

$$\left\{ \begin{array}{l}
\text{(i) } \mathcal{E}(\cdot, \cdot, r) \text{ and } \varphi(\cdot, \cdot, r) \text{ are measurable on } (0, T) \times \Gamma_C, \forall r \in \mathbb{R}; \\
\text{(ii) } \varphi(t, \mathbf{x}, \cdot) \text{ is locally Lipschitz on } \mathbb{R} \text{ for a.e. } (t, \mathbf{x}) \in (0, T) \times \Gamma_C; \\
\text{(iii) there exists } c_0^\varphi \in L^2((0, T) \times \Gamma_C; \mathbb{R}^+), c_1^\varphi \geq 0, \text{ such that} \\
\quad |\varphi^0(t, \mathbf{x}, r)| \leq c_0^\varphi(t, \mathbf{x}) + c_1^\varphi |r|, \\
\quad \forall r \in \mathbb{R}, \text{ a.e. } (t, \mathbf{x}) \in (0, T) \times \Gamma_C; \\
\text{(iv) } (\mathcal{E}(t, \mathbf{x}, r_1) - \mathcal{E}(t, \mathbf{x}, r_2)) (r_1 - r_2) \geq 0, \\
\quad \forall r_1, r_2 \in \mathbb{R}, \text{ a.e. } (t, \mathbf{x}) \in (0, T) \times \Gamma_C.
\end{array} \right. \quad (31)$$

These assumptions are clearly satisfied in example (17).

Finally, we assume that the initial data satisfy the conditions

$$\mathbf{u}_0 \in V, \quad \mathbf{v}_0 \in V, \quad \theta_0 \in E, \quad \alpha_0 \in \mathcal{H}_{da}. \quad (32)$$

Using Green's formula, we obtain the following weak formulation of the mechanical problem Q , defined by a system of second-order quasi-variational evolution inequality coupled with a first-order evolution equation.

Problem QV : Find a displacement field $\mathbf{u} : [0, T] \rightarrow V$, a damage field $\alpha : [0, T] \rightarrow \mathcal{H}_{da}$, and a temperature field $\theta : [0, T] \rightarrow E$ satisfying for a.e. $t \in (0, T)$:

$$\left\{ \begin{array}{l} \langle \ddot{\mathbf{u}}(t) + A(t)\dot{\mathbf{u}}(t) + B(t)(\mathbf{u}(t), \alpha(t)) + C(t)\theta(t), \mathbf{w} - \dot{\mathbf{u}}(t) \rangle_{V' \times V}, \\ + \left(\int_0^t \mathcal{B}(t-s)(\boldsymbol{\varepsilon}(\mathbf{u}(s)), \alpha(s)) ds, \boldsymbol{\varepsilon}(\mathbf{w}) - \boldsymbol{\varepsilon}(\dot{\mathbf{u}}(t)) \right)_{\mathcal{H}} \\ + j_\nu(t, \mathbf{u}(t), \mathbf{w} - \dot{\mathbf{u}}(t)) + j_\tau(t, \mathbf{u}(t), \mathbf{w}) - j_\tau(t, \mathbf{u}(t), \dot{\mathbf{u}}(t)) \\ \geq \langle \mathbf{f}(t), \mathbf{w} - \dot{\mathbf{u}}(t) \rangle_{V' \times V}, \quad \forall \mathbf{w} \in V. \end{array} \right. \quad (33)$$

$$\left\{ \begin{array}{l} \langle \dot{\alpha}(t), \xi - \alpha(t) \rangle_{L^2(\Omega)} + \gamma \langle \nabla \alpha(t), \nabla \xi - \nabla \alpha(t) \rangle_{L^2(\Omega)^d} \\ \geq \langle \phi_d(\boldsymbol{\sigma}(t), \boldsymbol{\varepsilon}(\mathbf{u}(t)), \alpha(t)), \xi - \alpha(t) \rangle_{L^2(\Omega)}, \quad \forall \xi \in \mathcal{H}_{da}. \end{array} \right. \quad (34)$$

$$\left\{ \begin{array}{l} \langle \dot{\theta}(t), \eta \rangle_{E' \times E} + \langle K(t)\theta(t), \eta \rangle_{E' \times E} + \psi(t, \theta(t); \eta) \\ \geq \langle R(t), \dot{\mathbf{u}}(t), \theta(t) \rangle, \eta \rangle_{E' \times E} + \langle Q(t), \eta \rangle_{E' \times E}, \quad \forall \eta \in E. \end{array} \right. \quad (35)$$

$$\mathbf{u}(0) = \mathbf{u}_0, \quad \dot{\mathbf{u}}(0) = \mathbf{v}_0, \quad \alpha(0) = \alpha_0, \quad \theta(0) = \theta_0 \quad \text{in } \Omega. \quad (36)$$

Here, the operators and functions $A(t) : V \rightarrow V'$, $B(t) : V \times \mathcal{H}_{da} \rightarrow V'$, $C(t) : E \rightarrow V'$, $j_\nu, j_\tau : (0, T) \times V^2 \rightarrow \mathbb{R}^+$, $K(t) : E \rightarrow E'$, $\psi(t, \cdot; \cdot) : E \times E \rightarrow \mathbb{R}$, $R(t, \cdot, \cdot) : V \times E \rightarrow E'$, $\mathbf{f} : (0, T) \rightarrow V'$, and $Q : (0, T) \rightarrow E'$ are defined by, for all $\mathbf{v} \in V$, $\mathbf{w} \in V$, $\zeta \in E$, $\eta \in E$, $\xi \in \mathcal{H}_{da}$, for a.e. $t \in (0, T)$,

$$\begin{aligned} \langle A(t)\mathbf{v}, \mathbf{w} \rangle_{V' \times V} &= (\mathcal{A}(t)(\boldsymbol{\varepsilon}\mathbf{v}), \boldsymbol{\varepsilon}\mathbf{w})_{\mathcal{H}}, \\ \langle B(t)(\mathbf{v}, \xi), \mathbf{w} \rangle_{V' \times V} &= (\mathcal{G}(t)(\boldsymbol{\varepsilon}\mathbf{v}, \xi), \boldsymbol{\varepsilon}\mathbf{w})_{\mathcal{H}}, \\ \langle C(t)\zeta, \mathbf{w} \rangle_{V' \times V} &= (C_e(t, \zeta(\cdot)), \boldsymbol{\varepsilon}\mathbf{w})_{\mathcal{H}}, \\ j_\nu(t, \mathbf{v}, \mathbf{w}) &= \int_{\Gamma_C} p_\nu(t, v_\nu - g(t)) w_\nu da; \\ j_\tau(t, \mathbf{v}, \mathbf{w}) &= \int_{\Gamma_C} p_\tau(t, v_\nu - g(t)) |\mathbf{w}_\tau| da; \\ \langle \mathbf{f}(t), \mathbf{w} \rangle_{V' \times V} &= (\mathbf{f}_0(t), \mathbf{w})_H + (\mathbf{f}_F(t), \mathbf{w})_{(L^2(\Gamma_F))^d}; \\ \langle K(t)\zeta, \eta \rangle_{E' \times E} &= \int_\Omega \mathcal{K}_c(t, \nabla \zeta) \cdot \nabla \eta dx; \\ \psi(t, \zeta; \eta) &= \int_{\Gamma_C} \varphi^0(t, x, \zeta(x); \eta(x)) da(x); \\ \langle R(t, \mathbf{v}, \zeta), \eta \rangle_{E' \times E} &= \int_\Omega D_e(t, \boldsymbol{\varepsilon}(\mathbf{v}), \zeta) \eta dx; \\ \langle Q(t), \eta \rangle_{E' \times E} &= \int_\Omega q(t) \eta dx. \end{aligned}$$

We notice that from (31), then the formula $\psi(t, \zeta; \eta)$ is well defined for all $\zeta \in E, \eta \in E$, for a.e. $t \in (0, T)$.

The inequality (35) is a consequence of the following equation:

$$\begin{cases} \langle \dot{\theta}(t), \eta \rangle_{E' \times E} + \langle K(t) \theta(t), \eta \rangle_{E' \times E} + \int_{\Gamma_C} \mathcal{E}(t, \theta(t)) \eta \, da \\ = \langle R(t, \dot{\mathbf{u}}(t), \theta(t)), \eta \rangle_{E' \times E} + \langle Q(t), \eta \rangle_{E' \times E}, \quad \forall \eta \in E, \end{cases} \quad (37)$$

where $\mathcal{E}(t, r) := \mathcal{E}(t, \cdot, r)$ for $(t, r) \in (0, T) \times \mathbb{R}$.

In the case when $\varphi(t, \mathbf{x}, \cdot)$ is differentiable for a.e. $(t, \mathbf{x}) \in (0, T) \times \Gamma_C$, we have

$$\mathcal{E}(t, \mathbf{x}, r) = \varphi'(t, \mathbf{x}, r) := [\varphi(t, \mathbf{x}, \cdot)]'(r)$$

for $(t, \mathbf{x}, r) \in (0, T) \times \Gamma_C \times \mathbb{R}$.

Then, for all $\zeta \in E$ and a.e. $t \in (0, T)$, the linear functional

$$\eta \in E \mapsto \psi(t, \zeta; \eta) = \int_{\Gamma_C} \mathcal{E}(t, \zeta) \eta \, da = \int_{\Gamma_C} \varphi'(t, \mathbf{x}, \zeta(\mathbf{x})) \eta(\mathbf{x}) \, da(\mathbf{x})$$

will be denoted by

$$\Phi(t, \zeta) \in E'.$$

The inequality (35) or Equation (37) can be written as

$$\dot{\theta}(t) + K(t) \theta(t) + \Phi(t, \theta(t)) = R(t, \dot{\mathbf{u}}(t), \theta(t)) + Q(t) \quad \text{in } E'.$$

Our main existence and uniqueness result is the following, which we will prove in the next section.

Theorem 1 *Assume that (19)–(32) hold, and under the condition that*

$$L_\tau < \frac{m_{\mathcal{A}}}{\sqrt{2} T c_0^2},$$

then there exists an unique solution $\{\mathbf{u}, \alpha, \theta\}$ to problem QV with the regularity:

$$\begin{cases} \mathbf{u} \in C^1(0, T; H) \cap W^{1,2}(0, T; V) \cap W^{2,2}(0, T; V'); \\ \alpha \in W^{1,2}(0, T; L^2(\Omega)) \cap L^\infty(0, T; \mathcal{K}_{da}); \\ \theta \in C(0, T; L^2(\Omega)) \cap L^2(0, T; E) \cap W^{1,2}(0, T; E'). \end{cases} \quad (38)$$

3 Proof of Theorem 1

The idea is to bring the second-order inequality to a first-order inequality, using monotone operator, convexity, and fixed point arguments, and will be carried out in several steps.

Let us introduce the velocity variable

$$\mathbf{v} = \dot{\mathbf{u}}.$$

The system in problem QV is then written as, for a.e. $t \in (0, T)$,

$$\left\{ \begin{array}{l} \mathbf{u}(t) = \mathbf{u}_0 + \int_0^t \mathbf{v}(s) ds; \\ \langle \dot{\mathbf{v}}(t) + A(t) \mathbf{v}(t) + B(t)(\mathbf{u}(t), \alpha(t)) + C(t) \theta(t), \mathbf{w} - \mathbf{v}(t) \rangle_{V' \times V}, \\ + \left(\int_0^t \mathcal{B}(t-s) (\boldsymbol{\varepsilon}(\mathbf{u}(s)), \alpha(s)) ds, \boldsymbol{\varepsilon}(\mathbf{w}) - \boldsymbol{\varepsilon}(\mathbf{v}(t)) \right)_{\mathcal{H}} \\ + j_v(t, \mathbf{u}(t), \mathbf{w} - \mathbf{v}(t)) + j_\tau(t, \mathbf{u}(t), \mathbf{w}) - j_\tau(t, \mathbf{u}(t), \mathbf{v}(t)) \\ \geq \langle \mathbf{f}(t), \mathbf{w} - \mathbf{v}(t) \rangle_{V' \times V}, \quad \forall \mathbf{w} \in V; \\ \langle \dot{\alpha}(t), \xi - \alpha(t) \rangle_{L^2(\Omega)} + \gamma \langle \nabla \alpha(t), \nabla \xi - \nabla \alpha(t) \rangle_{L^2(\Omega)^d} \\ \geq \langle \phi_d(\boldsymbol{\sigma}(t), \boldsymbol{\varepsilon}(\mathbf{u}(t)), \alpha(t)), \xi - \alpha(t) \rangle_{L^2(\Omega)}, \quad \forall \xi \in \mathcal{K}_{da}; \\ \langle \dot{\theta}(t), \eta \rangle_{E' \times E} + \langle K(t) \theta(t), \eta \rangle_{E' \times E} + \psi(t, \theta(t); \eta) \\ \geq \langle R(t, \mathbf{v}(t), \theta(t)), \eta \rangle_{E' \times E} + \langle Q(t), \eta \rangle_{E' \times E}, \quad \forall \eta \in E; \\ \mathbf{v}(0) = \mathbf{v}_0, \quad \alpha(0) = \alpha_0, \quad \theta(0) = \theta_0 \quad \text{in } \Omega, \end{array} \right.$$

with the regularities:

$$\left\{ \begin{array}{l} \mathbf{v} \in C(0, T; H) \cap L^2(0, T; V) \cap W^{1,2}(0, T; V'); \\ \alpha \in W^{1,2}(0, T; L^2(\Omega)) \cap L^\infty(0, T; K); \\ \theta \in C(0, T; L^2(\Omega)) \cap L^2(0, T; E) \cap W^{1,2}(0, T; E'). \end{array} \right.$$

We begin by the following lemma.

Lemma 1 *For all $\eta \in W^{1,2}(0, T; V')$, there exists an unique*

$$\mathbf{v}_\eta \in C(0, T; H) \cap L^2(0, T; V) \cap W^{1,2}(0, T; V')$$

satisfying

$$\left\{ \begin{array}{l} \langle \dot{\mathbf{v}}_\eta(t) + A(t) \mathbf{v}_\eta(t), \mathbf{w} - \mathbf{v}_\eta(t) \rangle_{V' \times V} + \langle \eta(t), \mathbf{w} - \mathbf{v}_\eta(t) \rangle_{V' \times V} \\ \quad + j_\tau(t, \mathbf{u}_\eta(t), \mathbf{w}) - j_\tau(t, \mathbf{u}_\eta(t), \mathbf{v}_\eta(t)) \geq \langle \mathbf{f}(t), \mathbf{w} - \mathbf{v}_\eta(t) \rangle_{V' \times V}, \\ \quad \forall \mathbf{w} \in V, \quad \text{a.e. } t \in (0, T); \\ \mathbf{v}_\eta(0) = \mathbf{v}_0, \end{array} \right. \quad (39)$$

where

$$\mathbf{u}_\eta(t) = \mathbf{u}_0 + \int_0^t \mathbf{v}_\eta(s) ds, \quad \forall t \in [0, T].$$

Moreover, if $L_\tau < \frac{m_{\mathcal{A}}}{\sqrt{2} T c_0^2}$, then $\exists c > 0$ such that $\forall \eta_1, \eta_2 \in W^{1,2}(0, T; V')$, $\forall t \in [0, T]$:

$$\|\mathbf{v}_{\eta_2}(t) - \mathbf{v}_{\eta_1}(t)\|_H^2 + \int_0^t \|\mathbf{v}_{\eta_2} - \mathbf{v}_{\eta_1}\|_V^2 \leq c \int_0^t \|\eta_1 - \eta_2\|_{V'}^2. \quad (40)$$

Proof Given $\eta \in W^{1,2}(0, T; V')$ and $x \in C(0, T; V)$, by using a general result on parabolic variational inequality (see e.g. [1]), we obtain the existence of a unique $\mathbf{v}_{\eta x} \in C(0, T; H) \cap L^2(0, T; V) \cap W^{1,2}(0, T; V')$ satisfying

$$\left\{ \begin{array}{l} \langle \dot{\mathbf{v}}_{\eta x}(t) + A(t) \mathbf{v}_{\eta x}(t), \mathbf{w} - \mathbf{v}_{\eta x}(t) \rangle_{V' \times V} + \langle \eta(t), \mathbf{w} - \mathbf{v}_{\eta x}(t) \rangle_{V' \times V} \\ \quad + j_\tau(t, x(t), \mathbf{w}) - j_\tau(t, x(t), \mathbf{v}_{\eta x}(t)) \geq \langle \mathbf{f}(t), \mathbf{w} - \mathbf{v}_{\eta x}(t) \rangle_{V' \times V}, \\ \quad \forall \mathbf{w} \in V, \quad \text{a.e. } t \in (0, T); \\ \mathbf{v}_{\eta x}(0) = \mathbf{v}_0. \end{array} \right. \quad (41)$$

Now, let us fix $\eta \in W^{1,2}(0, T; V')$ and consider $A_\eta : C(0, T; V) \rightarrow C(0, T; V)$ defined by

$$\forall x \in C(0, T; V), \quad A_\eta x(t) = \mathbf{u}_0 + \int_0^t \mathbf{v}_{\eta x}(s) ds.$$

We check by algebraic manipulation that for all $\mathbf{u}_1, \mathbf{u}_2, \mathbf{w}_1, \mathbf{w}_2 \in V$, a.e. $t \in (0, T)$, we have

$$j_\tau(t, \mathbf{u}_1, \mathbf{w}_2) - j_\tau(t, \mathbf{u}_1, \mathbf{w}_1) + j_\tau(t, \mathbf{u}_2, \mathbf{w}_1) - j_\tau(t, \mathbf{u}_2, \mathbf{w}_2) \leq c_1 \|\mathbf{u}_2 - \mathbf{u}_1\|_V \|\mathbf{w}_2 - \mathbf{w}_1\|_V,$$

where $c_1 = L_\tau c_0^2$ is involving c_0 , which is defined by (18).

Let $x_1, x_2 \in C(0, T; V)$ be given. Putting in (41) the data $x = x_1$ with $\mathbf{w} = \mathbf{v}_{\eta x_2}$ and $x = x_2$ with $\mathbf{w} = \mathbf{v}_{\eta x_1}$, adding then the two inequalities, and integrating over $(0, T)$, we obtain, $\forall t \in [0, T]$,

$$\begin{aligned} & \| \mathbf{v}_{\eta x_2}(t) - \mathbf{v}_{\eta x_1}(t) \|_H^2 + \int_0^t \| \mathbf{v}_{\eta x_2}(s) - \mathbf{v}_{\eta x_1}(s) \|_V^2 ds \\ & \leq c \int_0^t \| x_2(s) - x_1(s) \|_V^2 ds + c \int_0^t \| \mathbf{v}_{\eta x_2}(s) - \mathbf{v}_{\eta x_1}(s) \|_H^2 ds. \end{aligned}$$

Using Gronwall's inequality (see e.g. [2]), we deduce that

$$\forall x_1, x_2 \in C(0, T; V), \quad \forall t \in [0, T], \quad \| \Lambda_\eta(x_2)(t) - \Lambda_\eta(x_1)(t) \|_V^2 \leq c \int_0^t \| x_2(s) - x_1(s) \|_V^2 ds.$$

Thus, by Banach's fixed point principle, we know that Λ_η has an unique fixed point denoted by x_η . We then verify that

$$\mathbf{v}_\eta = \mathbf{v}_{\eta x_\eta}$$

is the unique solution verifying (39).

Now, let $\eta_1, \eta_2 \in W^{1,2}(0, T; V')$. Putting in (39) the data $\eta = \eta_1$ with $\mathbf{w} = \mathbf{v}_{\eta_2}$ and $\eta = \eta_2$ with $\mathbf{w} = \mathbf{v}_{\eta_1}$, adding then the two inequalities and integrating over $(0, T)$, and using the inequality

$$|a b| \leq \frac{\varepsilon}{4} a^2 + \frac{1}{\varepsilon} b^2$$

for all reals $a, b, \varepsilon > 0$, we obtain for all $\delta > 0$, for all $t \in [0, T]$:

$$\begin{aligned} & \frac{1}{2} \| \mathbf{v}_{\eta_2}(t) - \mathbf{v}_{\eta_1}(t) \|_H^2 + m_{\mathcal{A}} \int_0^t \| \mathbf{v}_{\eta_2}(s) - \mathbf{v}_{\eta_1}(s) \|_V^2 ds \\ & \leq m_{\mathcal{A}} \int_0^t \| \mathbf{v}_{\eta_2}(s) - \mathbf{v}_{\eta_1}(s) \|_H^2 ds + \frac{c_1^2}{4\delta} \int_0^t \| \mathbf{u}_{\eta_2}(s) - \mathbf{u}_{\eta_1}(s) \|_V^2 ds \\ & + \delta \int_0^t \| \mathbf{v}_{\eta_2}(s) - \mathbf{v}_{\eta_1}(s) \|_V^2 ds + \int_0^t \| \mathbf{v}_{\eta_2}(s) - \mathbf{v}_{\eta_1}(s) \|_V \| \eta_2(s) - \eta_1(s) \|_{V'} ds. \\ & \leq m_{\mathcal{A}} \int_0^t \| \mathbf{v}_{\eta_2}(s) - \mathbf{v}_{\eta_1}(s) \|_H^2 ds + \frac{c_1^2}{4\delta} \int_0^t \| \mathbf{u}_{\eta_2}(s) - \mathbf{u}_{\eta_1}(s) \|_V^2 ds \\ & + 2\delta \int_0^t \| \mathbf{v}_{\eta_2}(s) - \mathbf{v}_{\eta_1}(s) \|_V^2 ds + \frac{1}{4\delta} \int_0^t \| \eta_2(s) - \eta_1(s) \|_{V'}^2 ds. \end{aligned}$$

Now, verifying that

$$\int_0^t \| \mathbf{u}_{\eta_2}(s) - \mathbf{u}_{\eta_1}(s) \|_V^2 ds \leq T^2 \int_0^t \| \mathbf{v}_{\eta_2}(s) - \mathbf{v}_{\eta_1}(s) \|_V^2 ds,$$

we have

$$\begin{aligned}
& \frac{1}{2} \|\mathbf{v}_{\eta_2}(t) - \mathbf{v}_{\eta_1}(t)\|_H^2 + (m_{\mathcal{A}} - 2\delta) \int_0^t \|\mathbf{v}_{\eta_2}(s) - \mathbf{v}_{\eta_1}(s)\|_V^2 ds \\
& \leq m_{\mathcal{A}} \int_0^t \|\mathbf{v}_{\eta_2}(s) - \mathbf{v}_{\eta_1}(s)\|_H^2 ds + \frac{c_1^2}{4\delta} T^2 \int_0^t \|\mathbf{v}_{\eta_2}(s) - \mathbf{v}_{\eta_1}(s)\|_V^2 ds \\
& \quad + \frac{1}{4\delta} \int_0^t \|\eta_2(s) - \eta_1(s)\|_V^2 ds.
\end{aligned}$$

We deduce (40) from Gronwall's inequality if

$$\frac{c_1^2}{4\delta} T^2 < m_{\mathcal{A}} - 2\delta,$$

i.e.

$$L_\tau < \frac{m_{\mathcal{A}}}{T c_0^2} \sqrt{2\zeta(1-\zeta)},$$

where

$$\zeta = \frac{2\delta}{m_{\mathcal{A}}} \in]0, 1[.$$

To conclude, we obtain (40) if $\exists \zeta \in]0, 1[$ such that $L_\tau < \frac{m_{\mathcal{A}}}{T c_0^2} \sqrt{2\zeta(1-\zeta)}$.

This last condition is equivalent to

$$L_\tau < \frac{m_{\mathcal{A}}}{\sqrt{2} T c_0^2}.$$

□

Here and below, we denote by $c > 0$ a generic constant, which value may change from lines to lines.

Lemma 2 *For all $\eta \in W^{1,2}(0, T; V')$, there exists a unique*

$$\theta_\eta \in C(0, T; L^2(\Omega)) \cap L^2(0, T; E) \cap W^{1,2}(0, T; E')$$

satisfying

$$\left\{ \begin{array}{l}
\langle \dot{\theta}_\eta(t), \zeta \rangle_{E' \times E} + \langle K(t) \theta_\eta(t), \zeta \rangle_{E' \times E} + \int_{\Gamma_C} \Xi(t, \theta_\eta(t)) \zeta da \\
= \langle R(t, \mathbf{v}_\eta(t), \theta_\eta(t)), \zeta \rangle_{E' \times E} + \langle Q(t), \zeta \rangle_{E' \times E}, \\
\forall \zeta \in E, \text{ a.e. } t \in (0, T); \\
\theta_\eta(0) = \theta_0.
\end{array} \right. \quad (42)$$

Moreover, if $L_\tau < \frac{m_{\mathcal{A}}}{\sqrt{2} T c_0^2}$, then $\exists c > 0$ such that $\forall \eta_1, \eta_2 \in W^{1,2}(0, T; V')$:

$$\|\theta_{\eta_1}(t) - \theta_{\eta_2}(t)\|_{L^2(\Omega)}^2 \leq c \int_0^t \|\eta_1 - \eta_2\|_{V'}^2, \quad \forall t \in [0, T]. \quad (43)$$

Proof Let us fix $\eta \in W^{1,2}(0, T; V')$. We verify that $Q \in L^2(0, T; E')$.

Let us consider the operator $\Psi_\eta(t) : E \longrightarrow E'$ defined for a.e. $t \in (0, T)$ by

$$\begin{cases} \langle \Psi_\eta(t) \xi, \zeta \rangle_{E' \times E} := \langle K(t) \xi, \zeta \rangle_{E' \times E} + \int_{\Gamma_C} \mathcal{E}(t, \xi) \zeta \, da - \langle R(t, \mathbf{v}_\eta(t), \xi), \zeta \rangle_{E' \times E}, \\ \forall \xi, \zeta \in E. \end{cases}$$

Then, the problem is to find $\theta : (0, T) \longrightarrow E$ verifying

$$\begin{cases} \dot{\theta}(t) + \Psi_\eta(t) \theta(t) = Q(t) \text{ in } E', \text{ a.e. } t \in (0, T); \\ \theta(0) = \theta_0. \end{cases}$$

Using the assumptions (28), (29), and (31), $\Psi_\eta(t)$ is strongly monotone for a.e. $t \in (0, T)$. Therefore, the existence and uniqueness result verifying (42) follows from classical result on first-order evolution equation (see e.g. [9], pp. 162–164).

Now, for $\eta_1, \eta_2 \in W^{1,2}(0, T; V')$, we have, for a.e. $t \in (0, T)$,

$$\begin{aligned} & \langle \dot{\theta}_{\eta_1}(t) - \dot{\theta}_{\eta_2}(t), \theta_{\eta_1}(t) - \theta_{\eta_2}(t) \rangle_{E' \times E} + \langle K(t) \theta_{\eta_1}(t) - K(t) \theta_{\eta_2}(t), \theta_{\eta_1}(t) - \theta_{\eta_2}(t) \rangle_{E' \times E} \\ & \leq \langle R(t, \mathbf{v}_{\eta_1}(t), \theta_{\eta_1}(t)) - R(t, \mathbf{v}_{\eta_2}(t), \theta_{\eta_2}(t)), \theta_{\eta_1}(t) - \theta_{\eta_2}(t) \rangle_{E' \times E}. \end{aligned}$$

Then, integrating the last property over $(0, t)$, using the strong monotonicity of $K(t)$ and the Lipschitz continuity of $R(t, \cdot, \cdot) : V \times E \longrightarrow E'$ independently of $t \in (0, T)$, we deduce

$$\|\theta_{\eta_1}(t) - \theta_{\eta_2}(t)\|_{L^2(\Omega)}^2 \leq c \int_0^t \|\mathbf{v}_{\eta_1} - \mathbf{v}_{\eta_2}\|_V^2, \quad \forall t \in [0, T].$$

The inequality (43) follows then from Lemma 1. \square

Lemma 3 For all $\mu \in L^2(0, T; L^2(\Omega))$, there exists an unique

$$\alpha_\mu \in W^{1,2}(0, T; L^2(\Omega)) \cap L^2(0, T; H^1(\Omega))$$

satisfying

$$\begin{cases} \left(\alpha'_\mu(t), \xi - \alpha_\mu(t) \right)_{L^2(\Omega)} + \gamma (\nabla \alpha_\mu(t), \nabla \xi - \nabla \alpha_\mu(t))_{L^2(\Omega)^d} \\ \quad \geq (\mu(t), \xi - \alpha_\mu(t))_{L^2(\Omega)}, \quad \forall \xi \in \mathcal{K}_{da}, \quad \text{a.e. } t \in (0, T); \\ \alpha_\mu(t) \in \mathcal{K}_{da}, \quad \forall t \in [0, T]; \\ \alpha_\mu(0) = \alpha_0. \end{cases} \quad (44)$$

Moreover, $\exists c > 0$ such that $\forall \mu_1, \mu_2 \in L^2(0, T; L^2(\Omega))$:

$$\|\alpha_{\mu_2}(t) - \alpha_{\mu_1}(t)\|_{L^2(\Omega)}^2 \leq c \int_0^t \|\mu_1 - \mu_2\|_{L^2(\Omega)}^2, \quad \forall t \in [0, T]. \quad (45)$$

Proof The inequality (44) follows from classical result on parabolic evolution variational inequalities, see e.g. [1].

Now, for any $\mu_1, \mu_2 \in L^2(0, T; L^2(\Omega))$, putting in (44) the data $\mu = \mu_1$ with $\xi = \alpha_{\mu_2}$, then $\mu = \mu_2$ with $\xi = \alpha_{\mu_1}$, adding then the two inequalities, and integrating over $(0, T)$, we obtain, $\forall t \in [0, T]$,

$$\begin{aligned} & \frac{1}{2} \|\alpha_{\mu_1}(t) - \alpha_{\mu_2}(t)\|_{L^2(\Omega)}^2 + \gamma \int_0^t \|\nabla \alpha_{\mu_1} - \nabla \alpha_{\mu_2}\|_{L^2(\Omega)^d}^2 \\ & \leq \int_0^t \|\mu_1 - \mu_2\|_{L^2(\Omega)} \|\alpha_{\mu_1} - \alpha_{\mu_2}\|_{L^2(\Omega)}. \end{aligned}$$

Thus, the inequality (45) follows from Gronwall's inequality. □

Consider $X := W^{1,2}(0, T; V') \times L^2(0, T; L^2(\Omega))$, and the operator $\Lambda : X \rightarrow X$ is defined by, for all $(\eta, \mu) \in X$,

$$\begin{aligned} \Lambda(\eta, \mu) &= (\Lambda_1(\eta, \mu), \Lambda_2(\eta, \mu)); \\ \Lambda_1(\eta, \mu)(t) &= B(t)(\mathbf{u}_\eta(t), \alpha_\mu(t)) + D(t)(\mathbf{u}_\eta, \alpha_\mu) + j_v(t, \mathbf{u}_\eta(t), \cdot) + C(t)\theta_\eta(t); \\ \Lambda_2(\eta, \mu)(t) &= \phi_d(\boldsymbol{\sigma}_{\eta, \mu}(t), \boldsymbol{\varepsilon}(\mathbf{u}_\eta(t)), \alpha_\mu(t)), \end{aligned}$$

where

$$\langle D(t)(\mathbf{u}_\eta, \alpha_\mu), \mathbf{w} \rangle_{V' \times V} = \left(\int_0^t \mathcal{B}(t-s)(\boldsymbol{\varepsilon}(\mathbf{u}_\eta(s)), \alpha_\mu(s)) ds, \boldsymbol{\varepsilon} \mathbf{w} \right)_{\mathcal{H}}, \quad \forall \mathbf{w} \in V;$$

and

$$\begin{aligned} \boldsymbol{\sigma}_{\eta, \mu}(t) &= \mathcal{A}(t)\boldsymbol{\varepsilon}(\mathbf{v}_\eta(t)) + \mathcal{G}(t)(\boldsymbol{\varepsilon}(\mathbf{u}_\eta(t)), \alpha_\mu(t)) \\ &+ \int_0^t \mathcal{B}(t-s)(\boldsymbol{\varepsilon}(\mathbf{u}_\eta(s)), \alpha_\mu(s)) ds + C_e(t, \theta_\eta(t)). \end{aligned}$$

Lemma 4 Under the condition that $L_\tau < \frac{m_{\mathcal{A}}}{\sqrt{2} T c_0}$, then Λ has a unique fixed point (η^*, μ^*) .

Proof First, we check that from the definition of the operator $C(\cdot)$ and from hypothesis (24), then there exists $c > 0$, such that for a.e. $t \in (0, T)$, for all $\xi_1, \xi_2 \in E$, we have

$$\|C(t) \xi_1 - C(t) \xi_2\|_{V'} \leq c \|\xi_1 - \xi_2\|_{L^2(\Omega)}.$$

Now, let (η_1, μ_1) and (η_2, μ_2) be given in X . We verify that, for a.e. $t \in (0, T)$,

$$\begin{aligned} & \|\Lambda(\eta_1, \mu_1)(t) - \Lambda(\eta_2, \mu_2)(t)\|_{V' \times L^2(\Omega)}^2 \\ & \leq c \|B(t)(\mathbf{u}_{\eta_1}(t), \alpha_{\mu_1}(t)) - B(t)(\mathbf{u}_{\eta_2}(t), \alpha_{\mu_2}(t))\|_{V'}^2 + c \|D(t)(\mathbf{u}_{\eta_1}, \alpha_{\mu_1}) - D(t)(\mathbf{u}_{\eta_2}, \alpha_{\mu_2})\|_{V'}^2 \\ & \quad + c \|j_v(t, \mathbf{u}_{\eta_1}(t), \cdot) - j_v(t, \mathbf{u}_{\eta_2}(t), \cdot)\|_{V'}^2 + c \|C(t)\theta_{\eta_1}(t) - C(t)\theta_{\eta_2}(t)\|_{V'}^2 \\ & \quad + \|\phi_d(\sigma_{\eta_1, \mu_1}(t), \boldsymbol{\varepsilon}(\mathbf{u}_{\eta_1}(t)), \alpha_{\mu_1}(t)) - \phi_d(\sigma_{\eta_2, \mu_2}(t), \boldsymbol{\varepsilon}(\mathbf{u}_{\eta_2}(t)), \alpha_{\mu_2}(t))\|_{L^2(\Omega)}^2. \end{aligned}$$

Thus,

$$\begin{aligned} & \|\Lambda(\eta_1, \mu_1)(t) - \Lambda(\eta_2, \mu_2)(t)\|_{V' \times L^2(\Omega)}^2 \\ & \leq c \|\mathbf{u}_{\eta_1}(t) - \mathbf{u}_{\eta_2}(t)\|_V^2 + c \|\alpha_{\mu_1}(t) - \alpha_{\mu_2}(t)\|_{L^2(\Omega)}^2 + c \|\theta_{\eta_1}(t) - \theta_{\eta_2}(t)\|_{L^2(\Omega)}^2 \\ & \quad + c \|\mathbf{v}_{\eta_1}(t) - \mathbf{v}_{\eta_2}(t)\|_H^2. \end{aligned}$$

We deduce from Lemmas 1–3 that if $L_\tau < \frac{m_{\mathcal{A}}}{\sqrt{2T}c_0^2}$, then $\exists c > 0$ satisfying, for all $(\eta_1, \mu_1), (\eta_2, \mu_2)$ in X and for all $t \in [0, T]$,

$$\|\Lambda(\eta_1, \mu_1)(t) - \Lambda(\eta_2, \mu_2)(t)\|_{V' \times L^2(\Omega)}^2 \leq c \int_0^t \|\eta_2 - \eta_1\|_{V'}^2 + c \int_0^t \|\mu_1 - \mu_2\|_{L^2(\Omega)}^2.$$

Then, using again Banach's fixed point principle, we obtain that Λ has an unique fixed point. \square

Proof of Theorem 1 We have now all the ingredients to prove Theorem 1.

We verify then that the functions

$$\mathbf{u} := \mathbf{u}_{\eta^*}, \quad \alpha := \alpha_{\mu^*}, \quad \theta := \theta_{\eta^*}$$

are solutions to problem QV with the regularities in (38), the uniqueness follows from the uniqueness in Lemmas 1–3. \square

4 Analysis of a Numerical Scheme

In this section, we study a fully discrete numerical approximation scheme of the variational problem QV . For this purpose, let $\{\mathbf{u}, \theta\}$ be the unique solution of the problem QV , and introduce the velocity variable

$$\mathbf{v}(t) = \dot{\mathbf{u}}(t), \quad \forall t \in [0, T].$$

Then,

$$\mathbf{u}(t) = \mathbf{u}_0 + \int_0^t \mathbf{v}(s) ds, \quad \forall t \in [0, T]. \quad (46)$$

Here, we make the following additional assumptions on the different data, operators, and solution fields:

$$\begin{aligned} \mathcal{A}(\cdot, \cdot, \boldsymbol{\tau}) &\in C([0, T] \times \Omega; S_d), \quad \forall \boldsymbol{\tau} \in S_d; \\ \mathcal{G}(\cdot, \cdot, \boldsymbol{\tau}, \lambda) &\in C([0, T] \times \Omega; S_d), \quad \forall (\boldsymbol{\tau}, \lambda) \in S_d \times \mathbb{R}; \\ C_e(\cdot, \cdot, \vartheta) &\in C([0, T] \times \Omega; S_d), \quad \forall \vartheta \in \mathbb{R}; \\ \mathcal{B}(\cdot, \cdot, \boldsymbol{\tau}, \lambda) &\in C([0, T] \times \Omega; S_d), \quad \forall (\boldsymbol{\tau}, \lambda) \in S_d \times \mathbb{R}; \\ \mathbf{f}_0 &\in C([0, T] \times \Omega; \mathbb{R}^d); \quad \mathbf{f}_F \in C([0, T] \times \Gamma_F; \mathbb{R}^d); \\ \mathcal{K}_c(\cdot, \cdot, \xi) &\in C([0, T] \times \Omega; \mathbb{R}^d), \quad \forall \xi \in \mathbb{R}^d; \\ D_e(\cdot, \cdot, \boldsymbol{\tau}, \vartheta) &\in C([0, T] \times \Omega; \mathbb{R}), \quad \forall (\boldsymbol{\tau}, \vartheta) \in S_d \times \mathbb{R}; \\ q &\in C([0, T] \times \Omega; \mathbb{R}^+); \\ \mathbf{v} &\in W^{1,1}(0, T; V) \cap C^1([0, T]; H), \\ \theta &\in C([0, T]; E) \cap H^2(0, T; L^2(\Omega)), \\ \alpha &\in C(0, T; H^2(\Omega)) \cap H^2(0, T; L^2(\Omega)), \end{aligned} \quad (47)$$

and for all $r, r_1, r_2 \in \mathbb{R}$, a.e. $(t, \mathbf{x}) \in (0, T) \times \Gamma_C$:

$$\left\{ \begin{array}{l} \text{(i)} \quad \varphi^0(t, \mathbf{x}, r; r_1 + r_2) \leq \varphi^0(t, \mathbf{x}, r; r_1) + \varphi^0(t, \mathbf{x}, r; r_2); \\ \text{(ii)} \quad \varphi^0(t, \mathbf{x}, r_2; r_1 - r_2) + \varphi^0(t, \mathbf{x}, r_1; r_2 - r_1) \leq 0; \\ \text{(iii)} \quad \text{there exists } c^\varphi \geq 0 \text{ such that} \\ \quad \varphi^0(t, \mathbf{x}, r_1; r) + \varphi^0(t, \mathbf{x}, r_2; -r) \leq c^\varphi |(r_1 - r_2) r|. \end{array} \right. \quad (48)$$

We remark that the example of φ given in (17) satisfies hypothesis (48). From Theorem 1, $\{\mathbf{v}, \theta, \alpha\}$ verify, for all $t \in [0, T]$,

$$\left\{ \begin{array}{l} \langle \dot{\mathbf{v}}(t) + A(t) \mathbf{v}(t) + B(t)(\mathbf{u}(t), \alpha(t)) + C(t) \theta(t), \mathbf{w} - \mathbf{v}(t) \rangle_{V' \times V}, \\ + \left(\int_0^t \mathcal{B}(t-s) (\boldsymbol{\varepsilon}(\mathbf{u}(s)), \alpha(s)) ds, \boldsymbol{\varepsilon}(\mathbf{w}) - \boldsymbol{\varepsilon}(\mathbf{v}(t)) \right)_{\mathcal{H}} \\ + j_v(t, \mathbf{u}(t), \mathbf{w} - \mathbf{v}(t)) + j_\tau(t, \mathbf{u}(t), \mathbf{w}) - j_\tau(t, \mathbf{u}(t), \mathbf{v}(t)) \\ \geq \langle \mathbf{f}(t), \mathbf{w} - \mathbf{v}(t) \rangle_{V' \times V}, \quad \forall \mathbf{w} \in V. \end{array} \right. \quad (49)$$

$$\begin{cases} \langle \dot{\theta}(t), \eta \rangle_{E' \times E} + \langle K(t) \theta(t), \eta \rangle_{E' \times E} + \psi(t, \theta(t); \eta) \\ \geq \langle R(t, \mathbf{v}(t), \theta(t)), \eta \rangle_{E' \times E} + \langle Q(t), \eta \rangle_{E' \times E}, \quad \forall \eta \in E. \end{cases} \quad (50)$$

$$\begin{cases} (\dot{\alpha}(t), \xi - \alpha(t))_{L^2(\Omega)} + \gamma (\nabla \alpha(t), \nabla \xi - \nabla \alpha(t))_{L^2(\Omega)^d} \\ \geq (\phi_d(\boldsymbol{\sigma}(t), \boldsymbol{\varepsilon}(\mathbf{u}(t)), \alpha(t)), \xi - \alpha(t))_{L^2(\Omega)}, \quad \forall \xi \in \mathcal{K}_{da}. \end{cases} \quad (51)$$

$$\mathbf{v}(0) = \mathbf{v}_0, \quad \alpha(0) = \alpha_0, \quad \theta(0) = \theta_0 \quad \text{in } \Omega. \quad (52)$$

Now, let $V^h \subset V$, $E^h \subset E$, and $\mathcal{K}_{da}^h \subset \mathcal{K}_{da}$ be a family of finite dimensional subspaces, with $h > 0$ a discretization parameter. We divide the time interval $[0, T]$ into N equal parts: $t_n = nk$, $n = 0, 1, \dots, N$, with the time step $k = T/N$.

For a continuous operator or function $U \in C([0, T]; X)$ with values in a space X , we use the notation $U_n = U(t_n) \in X$.

Then, from (49)–(52), we introduce the following fully discrete scheme.

Problem P^{hk} Find $\mathbf{v}^{hk} = \{\mathbf{v}_n^{hk}\}_{n=0}^N \subset V^h$, $\theta^{hk} = \{\theta_n^{hk}\}_{n=0}^N \subset E^h$ and $\alpha^{hk} = \{\alpha_n^{hk}\}_{n=0}^N \subset \mathcal{K}_{da}^h$ such that

$$\mathbf{v}_0^{hk} = \mathbf{v}_0^h, \quad \theta_0^{hk} = \theta_0^h, \quad \alpha_0^{hk} = \alpha_0^h \quad (53)$$

and for $n = 1, \dots, N$,

$$\begin{cases} \left(\frac{\mathbf{v}_n^{hk} - \mathbf{v}_{n-1}^{hk}}{k}, \mathbf{w}^h - \mathbf{v}_n^{hk} \right)_H + \langle A_n \mathbf{v}_n^{hk}, \mathbf{w}^h - \mathbf{v}_n^{hk} \rangle_{V' \times V} \\ + \langle B_n \mathbf{u}_{n-1}^{hk}, \mathbf{w}^h - \mathbf{v}_n^{hk} \rangle_{V' \times V} + \langle C_n \theta_{n-1}^{hk}, \mathbf{w}^h - \mathbf{v}_n^{hk} \rangle_{V' \times V} \\ + (k \sum_{m=0}^{n-1} \mathcal{B}(t_n - t_m) (\boldsymbol{\varepsilon}(\mathbf{u}_m^{hk}), \alpha_m^{hk}), \boldsymbol{\varepsilon}(\mathbf{w}^h) - \boldsymbol{\varepsilon}(\mathbf{v}_n^{hk}))_{\mathcal{H}} \\ + j_v(t_n, \mathbf{u}_{n-1}^{hk}, \mathbf{w}^h - \mathbf{v}_n^{hk}) + j_\tau(t_n, \mathbf{u}_{n-1}^{hk}, \mathbf{w}^h) - j_\tau(t_n, \mathbf{u}_{n-1}^{hk}, \mathbf{v}_n^{hk}) \\ \geq \langle \mathbf{f}_n, \mathbf{w}^h - \mathbf{v}_n^{hk} \rangle_{V' \times V}, \quad \forall \mathbf{w}^h \in V^h. \end{cases} \quad (54)$$

$$\begin{cases} \left(\frac{\theta_n^{hk} - \theta_{n-1}^{hk}}{k}, \eta^h \right)_{L^2(\Omega)} + \langle K_n \theta_n^{hk}, \eta^h \rangle_{E' \times E} + \psi(t_n, \theta_n^{hk}; \eta^h) \\ \geq \langle R(t_n, \mathbf{v}_n^{hk}, \theta_n^{hk}), \eta^h \rangle_{E' \times E} + \langle Q_n, \eta^h \rangle_{E' \times E}, \quad \forall \eta^h \in E^h. \end{cases} \quad (55)$$

$$\begin{cases} \left(\frac{\alpha_n^{hk} - \alpha_{n-1}^{hk}}{k}, \xi^h - \alpha_n^{hk} \right)_{L^2(\Omega)} + \gamma (\nabla \alpha_n^{hk}, \nabla (\xi^h - \alpha_n^{hk}))_{L^2(\Omega)^d} \\ \geq (\phi_d(\boldsymbol{\sigma}_{n-1}^{hk}, \boldsymbol{\varepsilon}(\mathbf{u}_{n-1}^{hk}), \alpha_{n-1}^{hk}), \xi^h - \alpha_n^{hk})_{L^2(\Omega)}, \quad \forall \xi^h \in \mathcal{K}_{da}^h, \end{cases} \quad (56)$$

where for $n = 1, \dots, N$,

$$\mathbf{u}_n^{hk} = \mathbf{u}_0^{hk} + k \sum_{j=1}^n \mathbf{v}_j^{hk}; \quad \mathbf{u}_0^{hk} = \mathbf{u}_0^h. \quad (57)$$

$$\begin{cases} \sigma_n^{hk} = A_n \mathbf{v}_n^{hk} + B_n (\mathbf{u}_n^{hk}, \alpha_n^{hk}) + C_n \theta_n^{hk} + k \sum_{m=0}^{n-1} \mathcal{B}(t_n - t_m) (\boldsymbol{\varepsilon}(\mathbf{u}_m^{hk}), \alpha_m^{hk}); \\ \sigma_0^{hk} = \sigma_0^h. \end{cases} \quad (58)$$

Here, $\mathbf{u}_0^h \in V^h$, $\mathbf{v}_0^h \in V^h$, $\theta_0^h \in E^h$, $\alpha_0^h \in \mathcal{K}_{da}^h$, and $\sigma_0^h \in \mathcal{H}$ are suitable approximations of the initial values \mathbf{u}_0 , \mathbf{v}_0 , θ_0 , α_0 , and σ_0 , respectively.

We verify that for $n = 1, \dots, N$, once \mathbf{u}_{n-1}^{hk} , \mathbf{v}_{n-1}^{hk} , θ_{n-1}^{hk} , α_{n-1}^{hk} , and σ_{n-1}^{hk} are known, then we obtain \mathbf{v}_n^{hk} by (54), θ_n^{hk} by (55), α_n^{hk} by (56), \mathbf{u}_n^{hk} by (57) (using $\mathbf{u}_n^{hk} = \mathbf{u}_{n-1}^{hk} + k \mathbf{v}_n^{hk}$), and σ_n^{hk} by (58).

We now turn to an error analysis of the numerical solution. Here, we use and extend the technique developed in [3], p. 241.

Proof We have to estimate the following numerical solution errors, respectively, for the velocity, temperature, and damage:

$$\mathbf{v}_n - \mathbf{v}_n^{hk}, \quad \theta_n - \theta_n^{hk}, \quad \alpha_n - \alpha_n^{hk}, \quad 1 \leq n \leq N.$$

First step. Estimate of $(\alpha_n - \alpha_n^{hk})_{1 \leq n \leq N}$. Let us fix $n = 1, \dots, N$.

Using (51) with $t = t_n$, $\xi = \alpha_n^{hk}$ and (56) with $\xi^h = \xi_n^h \in \mathcal{K}_{da}^h$ and then adding the two inequalities, we obtain after some algebraic manipulation, for some constant $c > 0$,

$$\begin{aligned} & \|\alpha_n - \alpha_n^{hk}\|_{L^2(\Omega)}^2 + k \sum_{j=1}^n \|\nabla(\alpha_j - \alpha_j^{hk})\|_{L^2(\Omega)}^2 \\ & \leq +c \|\mathbf{u}_0 - \mathbf{u}_0^h\|_V^2 + c \|\sigma_0 - \sigma_0^h\|_{\mathcal{H}}^2 + c \|\alpha_0 - \alpha_0^h\|_{L^2(\Omega)}^2 \\ & \quad + c k \sum_{j=1}^n \left\| \frac{\alpha_j - \alpha_{j-1}}{k} - \dot{\alpha}_j \right\|_{L^2(\Omega)}^2 + c k \sum_{j=1}^n \|\alpha_j - \alpha_j^{hk}\|_{L^2(\Omega)}^2 \\ & \quad + c k^2 + c k \sum_{j=1}^{n-1} \|\mathbf{u}_j - \mathbf{u}_j^{hk}\|_V^2 + c \varepsilon k \sum_{j=1}^{n-1} \|\sigma_j - \sigma_j^{hk}\|_{\mathcal{H}}^2 \\ & \quad + c A_0^2 + c k A_1 + c k A_2 + c k A_3 + c k A_4, \end{aligned}$$

where $\varepsilon > 0$ is a small parameter which will be chosen later and

$$\begin{aligned} A_0 & := \max_{1 \leq j \leq N} \|\alpha_j - \xi_j^h\|_{L^2(\Omega)}; \\ \nabla A_1 & := \sum_{j=1}^N \|\nabla(\alpha_j - \xi_j^h)\|_{L^2(\Omega)}^2; \\ A_1 & := \sum_{j=1}^N \|\alpha_j - \xi_j^h\|_{L^2(\Omega)}^2; \\ A_2 & := \sum_{j=1}^{N-1} \|(\alpha_{j+1} - \xi_{j+1}^h) - (\alpha_j - \xi_j^h)\|_{L^2(\Omega)}^2; \\ A_3 & := \sum_{j=1}^N \|\phi_d(\sigma_j, \boldsymbol{\varepsilon}(\mathbf{u}_j), \alpha_j) - \frac{\alpha_j - \alpha_{j-1}}{k} + \gamma_j \Delta \alpha_j\|_{L^2(\Omega)} \times \|\alpha_j - \xi_j^h\|_{L^2(\Omega)}. \end{aligned}$$

From (47), we have

$$k A_3 \leq c A_0$$

and

$$\left\| \frac{\alpha_j - \alpha_{j-1}}{k} - \dot{\alpha}_j \right\|_{L^2(\Omega)} \leq \int_{t_{j-1}}^{t_j} \|\ddot{\alpha}(s)\|_{L^2(\Omega)} ds, \quad 1 \leq j \leq N.$$

We deduce that

$$\sum_{j=1}^n \left\| \frac{\alpha_j - \alpha_{j-1}}{k} - \dot{\alpha}_j \right\|_{L^2(\Omega)}^2 \leq c k.$$

From (46) and (57), we have

$$\begin{aligned} & k \sum_{j=1}^{n-1} \|\mathbf{u}_j - \mathbf{u}_j^{hk}\|_V^2 \\ & \leq c \|\mathbf{u}_0 - \mathbf{u}_0^h\|_V^2 + c k I + c k \sum_{j=1}^{n-1} \left(k \sum_{i=1}^j \|\mathbf{v}_i - \mathbf{v}_i^{hk}\|_V^2 \right), \end{aligned}$$

where by using (47),

$$I := \sum_{j=1}^N \left\| \int_0^{t_j} \mathbf{v} - k \sum_{i=1}^j \mathbf{v}_i \right\|_V^2 \leq c k.$$

From (58), we have for $n = 1, \dots, N$,

$$\begin{aligned} & \|\boldsymbol{\sigma}_n - \boldsymbol{\sigma}_n^{hk}\|_{\mathcal{H}}^2 \\ & \leq c \|\mathbf{v}_n - \mathbf{v}_n^{hk}\|_V^2 + c \|\mathbf{u}_n - \mathbf{u}_n^{hk}\|_V^2 + c \|\theta_n - \theta_n^{hk}\|_{L^2(\Omega)}^2 \\ & + \left\| \int_0^{t_n} \mathcal{B}(t_n - s) (\boldsymbol{\varepsilon}(\mathbf{u}(s)), \alpha(s)) ds - k \sum_{m=0}^{n-1} \mathcal{B}(t_n - t_m) (\boldsymbol{\varepsilon}(\mathbf{u}_m^{hk}), \alpha_m^{hk}) \right\|_{\mathcal{H}}^2 \end{aligned}$$

Therefore, we arrive to the following error estimate for the damage:

For some constant $c > 0$ and for $n = 1, \dots, N$,

$$\begin{aligned}
& \|\alpha_n - \alpha_n^{hk}\|_{L^2(\Omega)}^2 + k \sum_{j=1}^n \|\nabla(\alpha_j - \alpha_j^{hk})\|_{L^2(\Omega)}^2 \\
& \leq +c \|\mathbf{u}_0 - \mathbf{u}_0^h\|_V^2 + c \|\boldsymbol{\sigma}_0 - \boldsymbol{\sigma}_0^h\|_{\mathcal{H}}^2 + c \|\alpha_0 - \alpha_0^h\|_{L^2(\Omega)}^2 \\
& \quad + c k \sum_{j=1}^n \|\alpha_j - \alpha_j^{hk}\|_{L^2(\Omega)}^2 \\
& \quad + c k^2 + c k \sum_{j=1}^{n-1} \left(k \sum_{i=1}^j \|\mathbf{v}_i - \mathbf{v}_i^{hk}\|_V^2 \right) \\
& \quad + c \varepsilon k \sum_{j=1}^{n-1} \|\mathbf{v}_j - \mathbf{v}_j^{hk}\|_V^2 + c \varepsilon k \sum_{j=1}^{n-1} \|\theta_j - \theta_j^{hk}\|_{L^2(\Omega)}^2 \\
& \quad + c A_0 + c A_0^2 + c k \nabla A_1 + c k A_1 + c k A_2.
\end{aligned} \tag{59}$$

Second step. Estimate of $(\varepsilon_n := \theta_n - \theta_n^{hk})_{1 \leq n \leq N}$.

Let us fix $n = 1, \dots, N$ and denote shortly $\varepsilon_j := \theta_j - \theta_j^{hk}$, $1 \leq j \leq N$. We take (50), where $t = t_n$ and $\eta = -\eta^h$, and add to (55), with $\eta^h \in E^h$, we have

$$\begin{aligned}
& \left(\dot{\theta}_n - \frac{\theta_n^{hk} - \theta_n^{hk-1}}{k}, \eta^h \right)_{L^2(\Omega)} + \langle K_n \theta_n - K_n \theta_n^{hk}, \eta^h \rangle_{E' \times E} \\
& \leq \psi(t_n, \theta_n; -\eta^h) + \psi(t_n, \theta_n^{hk}; \eta^h) + \langle R(t_n, \mathbf{v}_n, \theta_n) - R(t_n, \mathbf{v}_n^{hk}, \theta_n^{hk}), \eta^h \rangle_{E' \times E}.
\end{aligned}$$

Taking $\eta^h = \eta_n^h - \theta_n + \varepsilon_n$, then we have

$$\begin{aligned}
& \left(\frac{\varepsilon_n - \varepsilon_{n-1}}{k}, \varepsilon_n \right)_{L^2(\Omega)} + \langle K_n \theta_n - K_n \theta_n^{hk}, \varepsilon_n \rangle_{E' \times E} \\
& \leq \langle K_n \theta_n - K_n \theta_n^{hk}, \theta_n - \eta_n^h \rangle_{E' \times E} \\
& \quad + \langle R(t_n, \mathbf{v}_n, \theta_n) - R(t_n, \mathbf{v}_n^{hk}, \theta_n^{hk}), \eta^h \rangle_{E' \times E} \\
& \quad + \left(\dot{\theta}_n - \frac{\theta_n - \theta_{n-1}}{k} + \frac{\varepsilon_n - \varepsilon_{n-1}}{k}, \theta_n - \eta_n^h \right)_{L^2(\Omega)} - \left(\dot{\theta}_n - \frac{\theta_n - \theta_{n-1}}{k}, \varepsilon_n \right)_{L^2(\Omega)} \\
& \quad + \psi(t_n, \theta_n; -\eta^h) + \psi(t_n, \theta_n^{hk}; \eta^h).
\end{aligned}$$

From (28), we have

$$|\langle K_n \theta_n - K_n \theta_n^{hk}, \theta_n - \eta_n^h \rangle_{E' \times E}| \leq c \|\theta_n - \theta_n^{hk}\|_E \times \|\theta_n - \eta_n^h\|_E.$$

From (29), we have

$$\begin{aligned}
& |\langle R(t_n, \mathbf{v}_n, \theta_n) - R(t_n, \mathbf{v}_n^{hk}, \theta_n^{hk}), \eta^h \rangle_{E' \times E}| \\
& \leq D_V \|\mathbf{v}_n - \mathbf{v}_n^{hk}\|_V \times \|\eta^h\|_{L^2(\Omega)} + D_T \|\theta_n - \theta_n^{hk}\|_{L^2(\Omega)} \times \|\eta^h\|_{L^2(\Omega)}; \\
& \leq D_V \|\mathbf{v}_n - \mathbf{v}_n^{hk}\|_V \times \|\eta_n^h - \theta_n\|_{L^2(\Omega)} + D_T \|\theta_n - \theta_n^{hk}\|_{L^2(\Omega)} \times \|\eta_n^h - \theta_n\|_{L^2(\Omega)} \\
& \quad + D_V \|\mathbf{v}_n - \mathbf{v}_n^{hk}\|_V \times \|\theta_n - \theta_n^{hk}\|_{L^2(\Omega)} + D_T \|\theta_n - \theta_n^{hk}\|_{L^2(\Omega)}^2.
\end{aligned}$$

Then, let us denote

$$B_0 := \max_{1 \leq n \leq N} \|\theta_n - \eta_n^h\|_{L^2(\Omega)}.$$

We have

$$D_V \|\mathbf{v}_n - \mathbf{v}_n^{hk}\|_V \times \|\eta_n^h - \theta_n\|_{L^2(\Omega)} \leq D_V \|\mathbf{v}_n - \mathbf{v}_n^{hk}\|_V B_0 \leq \frac{1}{2} D_V^2 \|\mathbf{v}_n - \mathbf{v}_n^{hk}\|_V^2 + \frac{1}{2} B_0^2;$$

and for $\epsilon_1 > 0$,

$$D_T \|\theta_n - \theta_n^{hk}\|_{L^2(\Omega)} \times \|\eta_n^h - \theta_n\|_{L^2(\Omega)} \leq \epsilon_1 \|\theta_n - \theta_n^{hk}\|_{L^2(\Omega)}^2 + \frac{1}{4\epsilon_1} (D_T B_0)^2;$$

and for $\epsilon > 0$,

$$D_V \|\mathbf{v}_n - \mathbf{v}_n^{hk}\|_V \times \|\theta_n - \theta_n^{hk}\|_{L^2(\Omega)} \leq \frac{D_V^2}{4\epsilon} \|\mathbf{v}_n - \mathbf{v}_n^{hk}\|_V^2 + \epsilon \|\theta_n - \theta_n^{hk}\|_{L^2(\Omega)}^2.$$

To continue, by using (48), we obtain

$$\psi(t_n, \theta_n; -\eta^h) + \psi(t_n, \theta_n^{hk}; \eta^h) \leq c_0 c^\varphi \|\theta_n - \theta_n^{hk}\|_E \times \|\eta^h\|_E,$$

and thus

$$\psi(t_n, \theta_n; -\eta^h) + \psi(t_n, \theta_n^{hk}; \eta^h) \leq c_0 c^\varphi \|\theta_n - \theta_n^{hk}\|_E^2 + c_0 c^\varphi \|\theta_n - \theta_n^{hk}\|_E \times \|\theta_n - \eta_n^h\|_E.$$

Consider the quantity for $n = 1, \dots, N$,

$$\mathcal{E}_n := \left(\frac{\varepsilon_n - \varepsilon_{n-1}}{k}, \varepsilon_n \right)_{L^2(\Omega)} + \langle K_n \theta_n - K_n \theta_n^{hk}, \varepsilon_n \rangle_{E' \times E}.$$

We have

$$\mathcal{E}_n \geq \frac{1}{2k} \left(\|\varepsilon_n\|_{L^2(\Omega)}^2 - \|\varepsilon_{n-1}\|_{L^2(\Omega)}^2 \right) + m_{\mathcal{K}_c} \|\varepsilon_n\|_E^2.$$

Now, we sum \mathcal{E}_j from $j = 1$ to $j = n$.

From (47), we have

$$\sum_{j=1}^n \left\| \frac{\theta_j - \theta_{j-1}}{k} - \dot{\theta}_j \right\|_{L^2(\Omega)}^2 \leq c k.$$

Under the condition that

$$D_T + c_0 c^\varphi < m_{\mathcal{K}_c}, \tag{60}$$

we can choose ϵ and ϵ_1 such that $\epsilon + \epsilon_1 + D_T + c_0 c^\varphi < m_{\mathcal{K}_c}$.

After some manipulation, we deduce the following error estimate for the temperature.

For some constant $c > 0$ independent of D_V and for $n = 1, \dots, N$,

$$\begin{aligned} & \|\theta_n - \theta_n^{hk}\|_{L^2(\Omega)}^2 + k \sum_{j=1}^n \|\theta_j - \theta_j^{hk}\|_E^2 \\ & \leq c \|\theta_0 - \theta_0^h\|_{L^2(\Omega)}^2 + c B_0^2 + c k^2 + c k B_1 + c B_2 M_\theta \\ & + c D_V^2 k \sum_{j=1}^n \|\mathbf{v}_j - \mathbf{v}_j^{hk}\|_V^2. \end{aligned} \tag{61}$$

Here,

$$\begin{aligned} M_\theta & := \max_{1 \leq n \leq N} \|\theta_n - \theta_n^{hk}\|_{L^2(\Omega)}, \\ B_1 & := \sum_{j=1}^N \|\theta_j - \eta_j^h\|_E^2, \\ B_2 & := \sum_{j=1}^N \|\theta_j - \eta_j^h - (\theta_{j+1} - \eta_{j+1}^h)\|_{L^2(\Omega)}. \end{aligned}$$

Third step. Estimate of $(\mathbf{v}_n - \mathbf{v}_n^{hk})_{1 \leq n \leq N}$.

The computation of the estimate for the velocity is similar as in [3], p. 241, which we refer for details. We mention only the main steps.

We obtain, for some constant $c > 0$ and for $n = 1, \dots, N$,

$$\begin{aligned} & \|\mathbf{v}_n - \mathbf{v}_n^{hk}\|_H^2 + k \sum_{j=1}^n \|\mathbf{v}_j - \mathbf{v}_j^{hk}\|_V^2 \\ & \leq c \|\mathbf{v}_0 - \mathbf{v}_0^h\|_H^2 + c \|\mathbf{u}_0 - \mathbf{u}_0^h\|_V^2 \\ & + c C_0 + c k^2 + c k (C_1 + \hat{C}_1) + c C_2 M_v \\ & + c k \sum_{j=1}^n \mathcal{R}_j^{hk} + c k \sum_{j=1}^n J_{v_j}^{hk} + c k \sum_{j=1}^n J_{\tau_j}^{hk} \\ & + \varepsilon k \sum_{j=0}^{n-1} \|\theta_j - \theta_j^{hk}\|_{L^2(\Omega)}^2 + c k \sum_{j=1}^{n-1} \left(k \sum_{i=1}^j \|\mathbf{v}_i - \mathbf{v}_i^{hk}\|_V^2 \right). \end{aligned}$$

Here, we denote by

$$\begin{aligned} M_v & := \max_{1 \leq n \leq N} \|\mathbf{v}_n - \mathbf{v}_n^{hk}\|_H; \\ C_0 & := \max_{1 \leq n \leq N} \|\mathbf{v}_n - \mathbf{w}_n^h\|_H; \\ C_1 & := \sum_{j=1}^N \|\mathbf{v}_j - \mathbf{w}_j^h\|_V^2; \\ \hat{C}_1 & := \sum_{j=1}^N \|\mathbf{v}_j - \mathbf{w}_j^h\|_V; \\ C_2 & := \sum_{j=1}^{N-1} \|(\mathbf{v}_j - \mathbf{w}_j^h) - (\mathbf{v}_{j+1} - \mathbf{w}_{j+1}^h)\|_H, \end{aligned}$$

and for $n = 1, \dots, N$,

$$\begin{aligned} \mathcal{R}_n^{hk} = & \left(\int_0^{t_n} \mathcal{B}(t_n - s) \boldsymbol{\varepsilon}(\mathbf{u}(s)) ds - k \sum_{m=0}^{n-1} \mathcal{B}(t_n - t_m) \boldsymbol{\varepsilon}(\mathbf{u}_m^{hk}), -\boldsymbol{\varepsilon}(\mathbf{e}_n) \right)_{\mathcal{H}} \\ & + \left(k \sum_{m=0}^{n-1} \mathcal{B}(t_n - t_m) \boldsymbol{\varepsilon}(\mathbf{u}_m^{hk}), \boldsymbol{\varepsilon}(\mathbf{w}_n^h) - \boldsymbol{\varepsilon}(\mathbf{v}_n) \right)_{\mathcal{H}}; \end{aligned}$$

and

$$J_{vn}^{hk} = j_v(t_n, \mathbf{u}_n, \mathbf{v}_n^{hk} - \mathbf{v}_n) + j_v(t_n, \mathbf{u}_{n-1}^{hk}, \mathbf{w}_n^{hk} - \mathbf{v}_n^{hk});$$

and

$$J_{\tau n}^{hk} = j_{\tau}(t_n, \mathbf{u}_n, \mathbf{v}_n^{hk}) - j_{\tau}(t_n, \mathbf{u}_n, \mathbf{v}_n) + j_{\tau}(t_n, \mathbf{u}_{n-1}^{hk}, \mathbf{w}_n^h) - j_{\tau}(t_n, \mathbf{u}_{n-1}^{hk}, \mathbf{v}_n^{hk}).$$

We have, for $n = 1, \dots, N$,

$$\begin{aligned} k \sum_{j=1}^n \mathcal{R}_j^{hk} \\ \leq c k^2 + c \|\mathbf{u}_0 - \mathbf{u}_0^h\|_V^2 + c k \sum_{j=1}^{n-1} \left(k \sum_{i=1}^j \|\mathbf{v}_i - \mathbf{v}_i^{hk}\|_V^2 \right) + c k (C_1 + \widehat{C}_1); \end{aligned}$$

and

$$\begin{aligned} k \sum_{j=1}^n J_{vj}^{hk} \\ \leq c k^2 + c \|\mathbf{u}_0 - \mathbf{u}_0^h\|_V^2 + c k \sum_{j=1}^{n-1} \left(k \sum_{i=1}^j \|\mathbf{v}_i - \mathbf{v}_i^{hk}\|_V^2 \right) \\ + c \varepsilon k \sum_{j=1}^n \|\mathbf{v}_j - \mathbf{v}_j^{hk}\|_V^2 + c k C_1 + c k \widehat{C}_1; \end{aligned}$$

and

$$\begin{aligned} k \sum_{j=1}^n J_{\tau j}^{hk} \\ \leq c k^2 + c \|\mathbf{u}_0 - \mathbf{u}_0^h\|_V^2 + c k \sum_{j=1}^{n-1} \left(k \sum_{i=1}^j \|\mathbf{v}_i - \mathbf{v}_i^{hk}\|_V^2 \right) \\ + c \varepsilon k \sum_{j=1}^n \|\mathbf{v}_j - \mathbf{v}_j^{hk}\|_V^2 + c k C_1 + c k \widehat{C}_1. \end{aligned}$$

Thus, we obtain the following error estimate for the velocity.

For some constant $c > 0$ and for $n = 1, \dots, N$,

$$\begin{aligned} \|\mathbf{v}_n - \mathbf{v}_n^{hk}\|_H^2 + k \sum_{j=1}^n \|\mathbf{v}_j - \mathbf{v}_j^{hk}\|_V^2 \\ \leq c \|\mathbf{v}_0 - \mathbf{v}_0^h\|_H^2 + c \|\mathbf{u}_0 - \mathbf{u}_0^h\|_V^2 \\ + c C_0 + c k^2 + c k (C_1 + \widehat{C}_1) + c C_2 M_v \\ + c \varepsilon k \sum_{j=0}^{n-1} \|\theta_j - \theta_j^{hk}\|_{L^2(\Omega)}^2 + c \varepsilon k \sum_{j=1}^n \|\mathbf{v}_j - \mathbf{v}_j^{hk}\|_V^2 \\ + c k \sum_{j=1}^{n-1} \left(k \sum_{i=1}^j \|\mathbf{v}_i - \mathbf{v}_i^{hk}\|_V^2 \right). \end{aligned} \tag{62}$$

To summarize, adding the three inequalities (59), (61), and (62) and choosing D_V and ε small enough, we obtain, for some constant $c > 0$ and for $n = 1, \dots, N$,

$$\begin{aligned}
 & \|\alpha_n - \alpha_n^{hk}\|_{L^2(\Omega)}^2 + k \sum_{j=1}^n \|\nabla(\alpha_j - \alpha_j^{hk})\|_{L^2(\Omega)}^2 + \|\theta_n - \theta_n^{hk}\|_{L^2(\Omega)}^2 \\
 & + k \sum_{j=1}^n \|\theta_j - \theta_j^{hk}\|_E^2 + \|\mathbf{v}_n - \mathbf{v}_n^{hk}\|_H^2 + k \sum_{j=1}^n \|\mathbf{v}_j - \mathbf{v}_j^{hk}\|_V^2 \\
 & \leq +c \|\mathbf{u}_0 - \mathbf{u}_0^h\|_V^2 + c \|\mathbf{v}_0 - \mathbf{v}_0^h\|_H^2 + c \|\boldsymbol{\sigma}_0 - \boldsymbol{\sigma}_0^h\|_{\mathcal{H}}^2 \\
 & + c \|\alpha_0 - \alpha_0^h\|_{L^2(\Omega)}^2 + c \|\theta_0 - \theta_0^h\|_{L^2(\Omega)}^2 \\
 & + c k \sum_{j=1}^n \|\alpha_j - \alpha_j^{hk}\|_{L^2(\Omega)}^2 \\
 & + c k^2 + c k \sum_{j=1}^{n-1} \left(k \sum_{i=1}^j \|\mathbf{v}_i - \mathbf{v}_i^{hk}\|_V^2 \right) \\
 & + c A_0 + c A_0^2 + c k \nabla A_1 + c k A_1 + c k A_2 + c B_0^2 + c k B_1 + c B_2 M_\theta \\
 & + c C_0 + c k C_1 + c k \hat{C}_1 + c C_2 M_v.
 \end{aligned} \tag{63}$$

To end, let us recall the discrete version of Gronwall’s inequality, see e.g. [2].

Consider a sequence $\{r_n\}_{0 \leq n \leq N} \subset \mathbb{R}^+$ and $a \in \mathbb{R}^+$.

Assume

$$r_n \leq a + c k \sum_{j=0}^{n-1} r_j, \quad 1 \leq n \leq N.$$

Then, we have

$$r_n \leq (a + c k r_0) (1 + c k)^{n-1} \leq (a + c k r_0) e^{cT}, \quad 1 \leq n \leq N.$$

Now, from Gronwall’s inequality, using estimation (63) and under condition (60), we conclude that for D_V small enough, then there exists some constant $c > 0$:

$$\begin{aligned}
 & \max_{1 \leq n \leq N} \left(\|\alpha_n - \alpha_n^{hk}\|_{L^2(\Omega)}^2 + k \sum_{j=1}^n \|\nabla(\alpha_j - \alpha_j^{hk})\|_{L^2(\Omega)}^2 + \|\theta_n - \theta_n^{hk}\|_{L^2(\Omega)}^2 \right. \\
 & \left. + k \sum_{j=1}^n \|\theta_j - \theta_j^{hk}\|_E^2 + \|\mathbf{v}_n - \mathbf{v}_n^{hk}\|_H^2 + k \sum_{j=1}^n \|\mathbf{v}_j - \mathbf{v}_j^{hk}\|_V^2 \right) \\
 & \leq +c \|\mathbf{u}_0 - \mathbf{u}_0^h\|_V^2 + c \|\mathbf{v}_0 - \mathbf{v}_0^h\|_H^2 + c \|\boldsymbol{\sigma}_0 - \boldsymbol{\sigma}_0^h\|_{\mathcal{H}}^2 \\
 & + c \|\alpha_0 - \alpha_0^h\|_{L^2(\Omega)}^2 + c \|\theta_0 - \theta_0^h\|_{L^2(\Omega)}^2 \\
 & + c k^2 + c A_0 + c A_0^2 + c k \nabla A_1 + c k A_1 + c k A_2 \\
 & + c B_0^2 + c k B_1 + c B_2^2 + c C_0 + c k C_1 + c k \hat{C}_1 + c C_2^2.
 \end{aligned} \tag{64}$$

As a typical example, let us consider $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}^*$, a polygonal domain. Let \mathcal{F}^h be a regular finite element partition of Ω . Let $V^h \subset V$, $E^h \subset E$, and $\mathcal{H}_{da}^h \subset \mathcal{H}_{da}$ be the finite element spaces consisting of piecewise polynomials of degree $\leq m$, with

$m \geq 1$, according to the partition \mathcal{T}^h . Denote by $\Pi_V^h : H^{m+1}(\Omega)^d \rightarrow V^h$, $\Pi_E^h : H^{m+1}(\Omega) \rightarrow E^h$, and $\Pi_K^h : H^m(\Omega) \rightarrow \mathcal{X}_{da}^h$ the finite element interpolation operators.

Recall (see e.g. [4]) that

$$\begin{cases} \|\mathbf{w} - \Pi_V^h \mathbf{w}\|_{H^r(\Omega)^d} \leq c h^{m+1-r} |\mathbf{w}|_{H^{m+1}(\Omega)^d}, & \forall \mathbf{w} \in H^{m+1}(\Omega)^d; \\ \|\eta - \Pi_E^h \eta\|_{H^r(\Omega)} \leq c h^{m+1-r} |\eta|_{H^{m+1}(\Omega)}, & \forall \eta \in H^{m+1}(\Omega); \\ \|\xi - \Pi_K^h \xi\|_{L^2(\Omega)} \leq c h^m |\xi|_{H^m(\Omega)}, & \forall \xi \in H^m(\Omega), \end{cases}$$

where $r = 0$ (for which $H^0 = L^2$) or $r = 1$.

We assume the following additional data and solution regularities:

$$\begin{cases} \mathbf{u}_0 \in H^{m+1}(\Omega)^d; & \alpha_0 \in H^m(\Omega); \\ \mathbf{v} \in C([0, T]; H^{2m+1}(\Omega)^d), & \dot{\mathbf{v}} \in L^1(0, T; H^m(\Omega)^d); \\ \theta \in C([0, T]; H^{m+1}(\Omega)), & \dot{\theta} \in W^{1,2}(0, T; H^m(\Omega)); \\ \dot{\alpha} \in W^{1,1}(0, T; H^m(\Omega)). \end{cases} \quad (65)$$

Then, we choose in (64) the elements

$$\mathbf{u}_0^h = \Pi_V^h \mathbf{u}_0, \quad \mathbf{v}_0^h = \Pi_V^h \mathbf{v}_0, \quad \theta_0^h = \Pi_E^h \theta_0, \quad \alpha_0^h = \Pi_K^h \alpha_0,$$

and

$$\mathbf{w}_j^h = \Pi_V^h \mathbf{v}_j, \quad \eta_j^h = \Pi_E^h \theta_j, \quad j = 1 \cdots N.$$

From assumption (65), we have

$$\begin{aligned} \|\mathbf{u}_0 - \mathbf{u}_0^h\|_V &\leq c h^m, & \|\mathbf{v}_0 - \mathbf{v}_0^h\|_H &\leq c h^m; \\ \|\theta_0 - \theta_0^{hk}\|_{L^2(\Omega)} &\leq c h^m, & \|\alpha_0 - \alpha_0^h\|_{L^2(\Omega)} &\leq c h^m; \\ A_0 &\leq c h^{m+1}, & B_0 &\leq c h^{m+1}, & C_0 &\leq c h^{2m+1}; \\ k A_1 &\leq c h^{2m}, & k B_1 &\leq c h^{2m}, & k C_1 &\leq c h^{2m}, & k \hat{C}_1 &\leq c h^{2m}; \\ A_2 &\leq c h^{2m}, & B_2 &\leq c h^m, & C_2 &\leq c h^m. \end{aligned}$$

Using these estimates in (64), we conclude to the following error estimate result.

Theorem 2 *We keep the assumptions of Theorem 1. Under the additional assumptions (47), (48), and (65), and condition (60), then for D_V small enough, we obtain the error estimate for the corresponding discrete solution $\{(\mathbf{v}_n^{hk}, \theta_n^{hk}, \alpha_n^{hk}), 1 \leq n \leq N\}$:*

$$\begin{aligned}
& \max_{1 \leq n \leq N} \|\mathbf{v}_n - \mathbf{v}_n^{hk}\|_H + \left(k \sum_{n=1}^N \|\mathbf{v}_n - \mathbf{v}_n^{hk}\|_V^2 \right)^{1/2} \\
& + \max_{1 \leq n \leq N} \|\theta_n - \theta_n^{hk}\|_{L^2(\Omega)} + \left(k \sum_{n=1}^N \|\theta_n - \theta_n^{hk}\|_E^2 \right)^{1/2} \\
& + \max_{1 \leq n \leq N} \|\alpha_n - \alpha_n^{hk}\|_{L^2(\Omega)} \\
& \leq c \left(h^{\frac{m+1}{2}} + k \right).
\end{aligned}$$

In particular, for $m = 1$, we have

$$\begin{aligned}
& \max_{1 \leq n \leq N} \|\mathbf{v}_n - \mathbf{v}_n^{hk}\|_H + \left(k \sum_{n=1}^N \|\mathbf{v}_n - \mathbf{v}_n^{hk}\|_V^2 \right)^{1/2} \\
& + \max_{1 \leq n \leq N} \|\theta_n - \theta_n^{hk}\|_{L^2(\Omega)} + \left(k \sum_{n=1}^N \|\theta_n - \theta_n^{hk}\|_E^2 \right)^{1/2} \\
& + \max_{1 \leq n \leq N} \|\alpha_n - \alpha_n^{hk}\|_{L^2(\Omega)} \\
& \leq c (h + k).
\end{aligned}$$

5 Numerical Computations

In this section, we provide numerical simulations in two-dimensional tests for the variational problem (QV) by using Matlab computation codes. We refer to the previous numerical scheme and use spaces of continuous piecewise affine functions $V^h \subset V$, $E^h \subset E$, and $\mathcal{K}_{da}^h \subset \mathcal{K}_{da}$ as families of approximating subspaces.

Here, we consider the following formulas:

$$\begin{aligned}
\mathcal{G}(t)(\boldsymbol{\tau}, \alpha) &= \mathcal{G}^0(t) \boldsymbol{\tau} - \alpha (d_{ij}(t)) \quad \text{in } \Omega; \\
C_e(t, \theta) &:= -\theta (c_{ij}(t)) \quad \text{in } \Omega; \\
p_v(t, \cdot, r) &= c_v(t) r_+ \quad \text{on } \Gamma_C; \\
p_\tau(t, \cdot, r) &= \mu_\tau(t) c_v(t) r_+ \quad \text{on } \Gamma_C; \\
\mathcal{K}_c(t, \nabla \theta) &= (k_{ij}(t)) \nabla \theta \quad \text{in } \Omega; \\
D_e(t, \mathbf{v}, \theta) &= -c_{ij}(t) \frac{\partial v_i}{\partial x_j} - \theta d_e(t) \quad \text{in } \Omega; \\
\phi_d(\sigma, \boldsymbol{\varepsilon}(\mathbf{u}), \alpha) &= -d_1 \|\sigma\|_{VM} - d_2 L_d(\alpha) \quad \text{in } \Omega; \\
\varphi(t, r) &= \frac{1}{2} k_e(t) (r - \theta_R(t))^2 \quad \text{on } \Gamma_C.
\end{aligned}$$

In view of the numerical simulations, we consider a rectangular open set, linear elastic, and linear visco-elastic operators, for a.e. $t \in (0, T)$:

$$\Omega = (0, L_1) \times (0, L_2);$$

$$\Gamma_F = (\{0\} \times [0, L_2]) \cup ([0, L_1] \times \{L_2\}) \cup (\{L_1\} \times [0, L_2]); \quad \Gamma_C = [0, L_1] \times \{0\};$$

$$(\mathcal{G}(t) \boldsymbol{\tau})_{ij} = \frac{E_Y(t) r_P(t)}{1-r_P^2(t)} (\tau_{11} + \tau_{22}) \delta_{ij} + \frac{E_Y(t)}{1+r_P(t)} \tau_{ij}, \quad 1 \leq i, j \leq 2, \boldsymbol{\tau} \in S_2;$$

$$(\mathcal{A}(t) \boldsymbol{\tau})_{ij} = \mu(t) (\tau_{11} + \tau_{22}) \delta_{ij} + \eta(t) \tau_{ij}, \quad 1 \leq i, j \leq 2, \boldsymbol{\tau} \in S_2;$$

$$(\mathcal{B}(t) \boldsymbol{\tau})_{ij} = B_1(t) (\tau_{11} + \tau_{22}) \delta_{ij} + B_2(t) \tau_{ij}, \quad 1 \leq i, j \leq 2, \boldsymbol{\tau} \in S_2.$$

Here, E_Y is the Young's modulus, r_P is the Poisson's ratio of the material, δ_{ij} denotes the Kronecker symbol, and μ and η are viscosity constants.

For computations, we considered the following data (IS unity), for $t \in (0, T)$:

$$L_1 = L_2 = 1, \quad T = 1;$$

$$\mu(t) = 3e^t, \quad \eta(t) = \frac{10}{1+t^2}, \quad E_Y(t) = \frac{2}{1+t}, \quad r_P(t) = \frac{0.1}{1+t^2}, \quad f_0(\mathbf{x}, t) = (0, -t);$$

$$f_F(\mathbf{x}, t) = (0, 0), \quad \mathbf{x} \in \{0\} \times (0, L_2);$$

$$f_F(\mathbf{x}, t) = (0.4t, \frac{0.3}{1+t}), \quad \mathbf{x} \in ((0, L_1) \times \{L_2\}) \cup (\{L_1\} \times (0, L_2));$$

$$d_{11}(t) = d_{22}(t) = d_{12}(t) = d_{21}(t) = 1;$$

$$c_{11}(t) = c_{12}(t) = c_{21}(t) = t, \quad c_{22}(t) = t^2;$$

$$k_{11}(t) = \frac{2}{1+t}, \quad k_{22}(t) = \frac{1+t}{2}, \quad k_{12}(t) = k_{21}(t) = 1;$$

$$k_e(t) = \frac{1+t}{2}, \quad d_e(t) = t^2, \quad q(t) = t;$$

$$g(t, \mathbf{x}) = x(L_1 - x)t, \quad \mu_\tau(t, \mathbf{x}) = 0.1x t^2,$$

$$c_v(t, \mathbf{x}) = 10t x^2, \quad \mathbf{x} = (x, 0) \in (0, L_1) \times \{0\};$$

$$\gamma = 0.1, \quad d_1 = 1/50, \quad d_2 = 1/20, \quad L_d(s) = e^s, \quad 0 \leq s \leq 1;$$

$$\mathbf{u}_0 = (0, 0), \quad \mathbf{v}_0 = (0, 0), \quad \alpha_0 = 1, \quad \theta_0 = 0.$$

Figure 1 represents the initial configuration.

In Figures 2, 3, and 4, we compute, respectively, the Von Mises norm, which gives a global measure of the stress, the temperature, and the damage at final time in the body at final time, for $\theta_R = 0$, respectively, for short and long memory viscoelasticity. In Figure 5, we show the evolution of the damage at the particular point $S = (L_1, L_2)$ (direction of the surface traction). We observe that the distribution of these parameters is changing for long memory, the deformation is more important, as well as for the damage, temperature, and stress in the neighborhood of the point S .

Finally in Figure 6, we show the distribution of the temperature and damage of the body for larger ground temperature. Here, we observe larger deformation, larger damage, and larger temperature in the neighborhood of the contact surface.

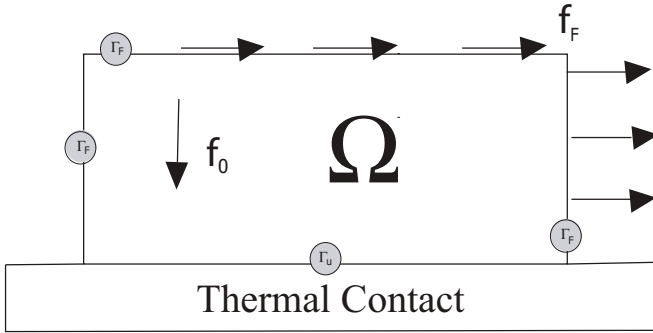


Fig. 1 Initial configuration

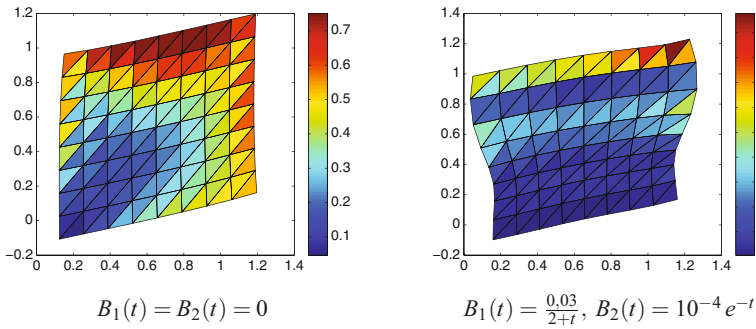


Fig. 2 Von Mises norm at final time, $\theta_R = 0$

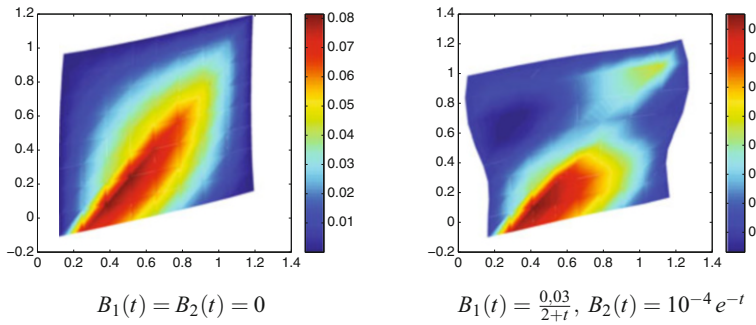


Fig. 3 Temperature field at final time, $\theta_R = 0$

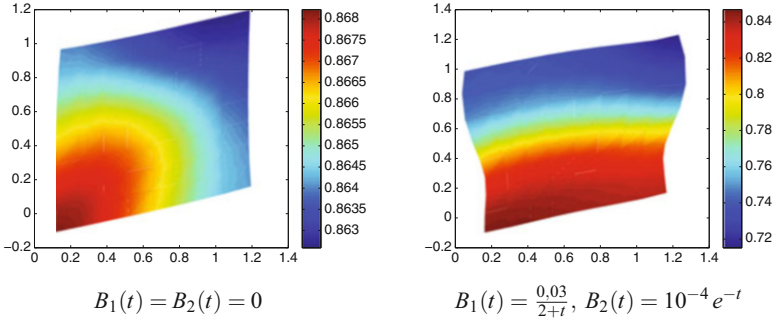


Fig. 4 Damage field at final time, $\theta_R = 0$

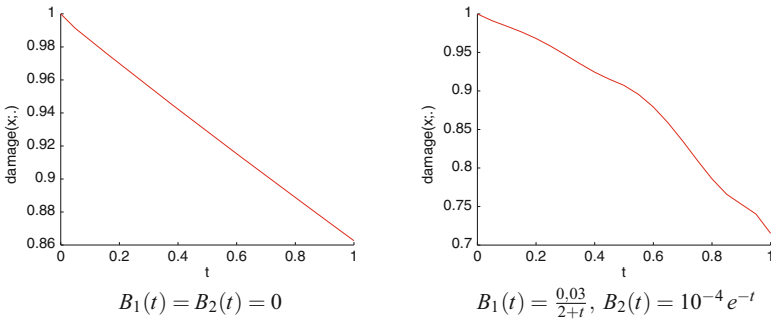


Fig. 5 Evolution of damage field at $x = (L_1, L_2), \theta_R = 0$

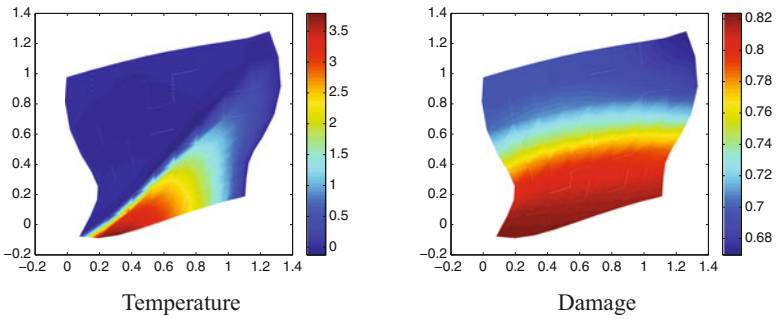


Fig. 6 Temperature and damage at final time, $B_1(t) = \frac{0.03}{2+t}, B_2(t) = 10^{-4} e^{-t}, \theta_R = 10$

References

1. V. Barbu, *Optimal Control of Variational Inequalities* (Pitman, London, 1984)
2. O. Chau, Ph.D. Thesis, Analyse variationnelle et numérique en mécanique du contact, University of Perpignan (2000)
3. O. Chau, Habilitation Thesis, Quelques problèmes d'évolution en mécanique de contact et en biochimie, University of La Reunion (2010)
4. P.G. Ciarlet, *The Finite Element Method for Elliptic Problems* (North Holland, Amsterdam, 1978)
5. G. Duvaut, J.L. Lions, *Les Inéquations en Mécanique et en Physique* (Dunod, Malakoff, 1972)
6. M. Frémond, B. Nedjar, Damage in concrete: the unilateral phenomenon. *Nucl. Eng. Design* **156**, 323–335 (1995)
7. M. Frémond, B. Nedjar, Damage, gradient of damage and principle of virtual work. *Int. J. Solids Struct.* **33**, 1083–1103 (1996)
8. N. Kikuchi, J.T. Oden, *Contact Problems in Elasticity* (SIAM, Philadelphia, 1988)
9. J.L. Lions, Quelques méthodes de résolution des problèmes aux limites non linéaires, Dunod et Gauthier-Villars (1969)
10. P.D. Panagiotopoulos, *Inequality Problems in Mechanics and Applications* (Birkhäuser, Basel, 1985)
11. P.D. Panagiotopoulos, *Hemivariational Inequalities, Applications in Mechanics and Engineering* (Springer, Berlin, 1993)
12. E. Zeidler, *Nonlinear Functional Analysis and its Applications, II/A, Linear Monotone Operators* (Springer, Berlin, 1997)

Mixed Concave–Convex Sub-Superlinear Schrödinger Equation: Survey and Development of Some New Cases



Riadh Chteoui, Anouar Ben Mabrouk, and Carlo Cattani

Abstract The present chapter is concerned with a whole review of the well known Schrödinger equation in a mixed case of nonlinearities. We precisely consider a general nonlinear model characterized by a superposition of linear, sub-linear, super-linear sometimes concave–convex power laws on the form $f(u) = |u|^{p-1}u \pm |u|^{p-1}u$. In a first part, we develop theoretical results on existence, uniqueness, classification as well as the behavior of the solutions of the ground state radial problem according to the power laws and the initial value. Next, in a second part, some examples are developed with graphical illustrations to confirm the theoretical results exposed previously. The graphs show coherent states between the theoretical findings and the numerical illustrations.

The chapter in its whole aim is a review of existing results about the studied problems reminiscent of some few cases that are not previously developed. We aim thus it will constitute a good reference especially for beginners in the field of nonlinear analysis of PDEs.

Mathematics Subject Classification 35J25, 35J60

R. Chteoui

Laboratory of Algebra, Number Theory and Nonlinear Analysis, Department of Mathematics, Faculty of Sciences, University of Monastir, Monastir, Tunisia

Department of Mathematics, Faculty of Sciences, University of Tabuk, Tabuk, Saudi Arabia

A. Ben Mabrouk

Laboratory of Algebra, Number Theory and Nonlinear Analysis, Department of Mathematics, Faculty of Sciences, University of Monastir, Monastir, Tunisia

Department of Mathematics, Higher Institute of Applied Mathematics and Computer Science, University of Kairouan, Kairouan, Tunisia

Department of Mathematics, Faculty of Sciences, University of Tabuk, Tabuk, Saudi Arabia
e-mail: anouar.benmabrouk@fsm.mu.tn

C. Cattani (✉)

Engineering School (DEIM), Tuscia University, Viterbo, Italy
e-mail: cattani@unitus.it

© Springer Nature Switzerland AG 2021

T. M. Rassias, P. M. Pardalos (eds.), *Nonlinear Analysis and Global Optimization*, Springer Optimization and Its Applications 167, https://doi.org/10.1007/978-3-030-61732-5_5

109

1 Introduction

In this chapter nonlinear equations issued from physical problems such as Schrödinger are considered in a general form of mixed nonlinearities characterized by the presence of both sub-linear (sometimes concave) and super-linear (convex) parts. In a first part, we focus on the ground state solutions of the continuous problem and investigate all possible cases relatively to the nonlinearities powers existing in the model. Existence, uniqueness, and classification of the solutions will be studied especially for the radial cases.

In addition we provided some numerical simulations to illustrate and show graphically the theoretical results.

The principal aim of this work is firstly to develop a whole and complete review on the models to be presented and to develop the cases that remain unsolved and thus to provide a complete study that we wish to be useful for researchers from different fields.

NLS equation is widely studied from both numerical and theoretical points of view. This is due to its link to real physical phenomena such as Newton's law and conservation of energy in classical mechanics, behavior of dynamical systems, the description of a particle in a non-relativistic setting in quantum mechanics, combustion, etc.

In the linear classical case, the solutions for Schrödinger equation for example are somehow known explicitly as in the cubic nonlinear form where the solutions are expressed as soliton type particles. Classical methods based on kernels especially Fourier permit to investigate the linear cases and some nonlinear specific ones with one nonlinearity. However, in the presence of many nonlinearities in the model, efforts remain to be done to provide a rigorous solution.

The Schrödinger equation in its general form is a prototypical dispersive nonlinear partial differential equation related to Bose–Einstein condensates and nonlinear optics, propagation of electric fields in optical fibers, self-focusing and collapse of Langmuir waves in plasma physics, behavior of rogue waves in oceans.

Based upon the analogy between mechanics and optics, Schrödinger established the classical derivation of his equation. By developing a perturbation method, he proved the equivalence between his wave mechanics equation and Heisenberg's matrix one, and thus introduced the time dependent version.

However, in the nonlinear case, the structure of the nonlinear Schrödinger equation is more complicated. It is also related to electromagnetic, ferromagnetic fields as well as magnums, high-power ultra-short laser self-channeling in matter, condensed matter theory, dissipative quantum mechanics, film equations, etc.

The Schrödinger equation is in fact a vector equation when separating the real and imaginary parts of the wave solution. In a specific basis of the state space this leads to a system of coupled equations. The same situation (system of Schrödinger equations) may be seen also in simultaneous particles (solitons) propagations or interactions. See for instance [30, 32, 34, 36, 38].

Schrödinger’s equation is initially expressed in a linear form as

$$\Delta\psi + \lambda(E - V(x))\psi = 0,$$

where ψ is the wave function, $\lambda = \frac{8\pi^2 m}{h^2}$, m is the mass, h is the well known Planck’s constant, E is the energy, and V is a potential energy.

The nonlinear Schrödinger (NLS) equation is more complicated. A well known model is the cubic NLS governing finitely many moving particles and which has been widely studied. It is stated on the form

$$ih \frac{\partial\psi(x, t)}{\partial t} = -\frac{h^2}{2m} \Delta\psi(x, t) + V(x, t)\psi(x, t) + Ng|\psi(x, t)|^2\psi(x, t). \tag{1}$$

$\psi(r, t)$ is a complex valued function known as condensate wave. m is the mass of the particle, $V(r, t)$ the exterior potential, N is the number of particles in the condensate. g is a coupling coefficient, and finally, V is the potential.

A general form of (1) is also met in plasma’s physics and the study of optic fibers. Such a generalization is expressed as

$$ih \frac{\partial\psi(x, t)}{\partial t} = -\frac{h^2}{2m} \Delta\psi(x, t) + V(x, t)\psi(x, t) + NgF(|\psi(x, t)|)\psi(x, t), \tag{2}$$

which may be expressed otherwise as

$$ih \frac{\partial\psi(x, t)}{\partial t} = -\frac{h^2}{2m} \Delta\psi(x, t) + V(x, t)\psi(x, t) + aF(|\psi(x, t)|)\psi(x, t), \tag{3}$$

The parameter a is a constant depending on N , g , h , and m . See [6, 37].

Schrödinger also established the classical derivation of his equation based on the analogy between mechanics and optics, and closer to De-Broglie’s formalism. A perturbation method based on Rayleigh in acoustics has been developed and introduced thus the time dependent equation

$$ih \frac{\partial\psi}{\partial t} = -\frac{h^2}{2m} \Delta\psi + V(x)\psi - \gamma |\psi|^{p-1} \psi \text{ in } \mathbb{R}^N, (N \geq 2), \tag{4}$$

where $p < p_c = \frac{2N}{N-2}$ for $N \geq 3$ and $p < +\infty$ if $N = 2$. In physical problems, a cubic nonlinearity corresponding to $p = 3$ is common and re-meets the well known Gross–Pitaevskii equation. In [29, 33], the potential V is assumed to be bounded and possessing a non-degenerate critical point at the origin.

2 The Stationary Problem

By taking in (4) $\gamma > 0$ and $h > 0$ sufficiently small and using a Lyapunov-Schmidt type reduction, Oh in [33] proved the existence of standing wave solutions of problem (4) of the form

$$\psi(x, t) = e^{-iEt/h}u(x). \tag{5}$$

This reduces the NLS equation (4) to the semi-linear elliptic equation

$$-\frac{h^2}{2m}\Delta u + (V(x) - E)u = |u|^{p-1}u$$

which by setting $x \rightsquigarrow hs$ and $z(s) = (2m)^{\frac{1}{p-1}}u(x)$, ($p \neq 1$) becomes

$$-\Delta z + 2m(V_h(x) - E)z = |z|^{p-1}z, \tag{6}$$

where $V_h(x) = V(hx)$. If furthermore, the potential V is translation-invariant with respect to some parameter ξ , the Equation (4) becomes invariant under the Galilean transformation

$$\psi(x, t) \mapsto \psi(x - t\xi, t) \exp(i\xi \cdot x/h - \frac{1}{2}i|\xi|^2t/h)\psi(x - t\xi, t).$$

In this case, it is well known that standing waves reproduce solitary waves traveling in the direction ξ .

In [35] a ground state solution for problem (6) has been proved to hold under suitable assumptions on the parameter h and the potential V . The problem reduces to a semi-linear elliptic equation

$$-\Delta u + V(x)u = f(x, u), \quad x \in \mathbb{R}^N. \tag{7}$$

Recently, problem (7) has been re-considered in [9–15] with a mixed model where no linear term exists, but in the contrary this was replaced by the odd extension $\lambda|u|^{q-1}u$ and the nonlinear term $f(u)$ was replaced by an odd extension $|u|^{p-1}u$. Some new difficulties appeared in the analysis. Using technical and direct computations from ODEs and classical inequalities such as Pohozaev’s and Sobolev’s ones the existence of some lower bound for λ to guarantee the positivity of solutions has been proved. A classification of radial solutions in some slightly critical cases relatively to the power p is investigated. The nonlinear function model is expressed as

$$f(s) = \pm|s|^{p-1} + \lambda|s|^{q-1}.$$

Associated NLS and heat equations have been also considered in both one and higher dimensional cases and some numerical developments have been investigated in two-dimensional case without applying the classical methods such as tri-diagonal systems. In [10], there has been provided a fascinating method based on Lyapunov–Sylvester operators which has been compared with classical methods. It is shown to be efficient from both time and error estimates. (See [9–15]).

Problem (7) has been also re-considered with V sign changing and the non-linearity on the form $f(x, u) = a(x)g(u)$ with a sub-linear function $g(u)$. Such problems in \mathbb{R}^N arise naturally in various branches of physics and present challenging mathematical difficulties. Whenever a bounded domain Ω is considered with Dirichlet boundary condition, multiplicity of solutions has been shown. When Ω is unbounded and especially on the whole space, the existence and multiplicity of nontrivial solutions have been widely investigated, both for sub-linear and super-linear nonlinearities. See [1–5, 8–18, 21–25, 27, 28].

Balabane et al. [7] proved that for each integer k , there is a radial compactly supported solution with k zeros in its support, provided that

$$V = -1 \text{ and } g(u) = |u|^{-2\theta}u, \text{ where } \theta \in]0, \frac{1}{2}[.$$

It is also proved the existence of infinitely many solutions for Equation (7) with a sub-nonlinearity

$$f(x, u) = |u|^{p-1}u, \quad 0 < p < 1$$

and where the potentials V and a satisfying $V > 0, a > 0, V \in (C^0(\mathbb{R}^N), \mathbb{R}), a$ continuous, $a \in L^{\frac{2}{1-p}}(\mathbb{R}^N)$, and where

$$m\{x \in B(y, r); V(x) \leq M\} \rightarrow 0 \text{ as } |y| \rightarrow +\infty, \quad \forall M > 0,$$

for some $r > 0$. the measure m stands for the Lebesgue measure on \mathbb{R}^N . In the critical case $p = 1 + \frac{4}{n}$, Carles and Zhang proved the global existence for some sufficient conditions. The supercritical case $p \geq 1 + \frac{4}{n}$ has been also considered by Carles who proved the existence of blow up solutions whenever the initial solution has a negative energy.

In the present chapter we focus on the radial solutions of the stationary problem associated with an evolutive NLS equation. We consider precisely an elliptic problem of the form

$$\begin{cases} u'' + \frac{d-1}{r}u' + ug(u) = 0, & r \in (0, \infty) \\ u'(0) = 0, & u(0) = a, \end{cases} \tag{8}$$

where a is a real number parameter and g is the model function characterized by mixed nonlinearities

$$g(u) = |u|^{p-1} \pm \lambda |u|^{q-1}. \tag{9}$$

We will study existence, uniqueness, and classification of the solutions relatively to the power nonlinearity parameters p, q , the initial value $u(0) = a$, and the real parameter $\lambda > 0$.

In some specific cases the last parameter λ may be compared to the eigenvalues of the Laplacian operator Δ . Indeed, for $q = 1$, the problem becomes a radial version of

$$-\Delta u = |u|^{p-1}u \pm \lambda u. \tag{10}$$

For example, positive solutions lead to the famous Brezis–Nirenberg problem

$$\begin{cases} \Delta u + u^p + \lambda u = 0 & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \\ u > 0 & \text{in } \Omega, \end{cases} \tag{11}$$

where Ω is a bounded domain with smooth boundary in \mathbb{R}^N , $N \geq 3$ and p depending on p_c . In this case, the bounds of λ are based in some parts on the presence of the linear term λu which allows a comparison with the eigenvalues of $-\Delta$. They are related also to the Rayleigh quotient and the Pohozaev’s identity. Let λ_1 be the first eigenvalue of $-\Delta$ in $\Omega = B$ the unit ball of \mathbb{R}^N with zero Dirichlet boundary condition. It is shown that bounded positive solutions are all radial and exist only in the following cases,

- (i) $p < p_c$ and $\lambda < \lambda_1$.
- (ii) $p = p_c$ and $\lambda \in (\lambda^*, \lambda_1)$, where $\lambda^* = 0$ if $N \geq 4$ and $\lambda^* = \frac{\lambda_1}{4}$ if $N = 3$.
- (iii) $p > p_c$ and $\lambda \in (\lambda_1^+, \lambda_1)$, for some λ_1^+ positive.

See for example [19, 20]. In [12, 13] similar results on the existence of positive bounded solutions of (36) have been established. The main and first difference with (11) is the absence of the linear term which is replaced by the odd extension $\lambda |u|^{q-1}u$ and the positive term u^p is replaced by its odd extension $|u|^{p-1}u$. Using technical and direct computations from (36) and using Pohozaev’s and Sobolev’s inequalities existence of some bounds for λ and a full classification of radially symmetric solutions of problem (36) have been established.

In the remaining parts of the present chapter we will adopt the following notations:

$$f(u) = ug(u) = |u|^{p-1}u \pm \lambda |u|^{q-1}u \tag{12}$$

and

$$F(u) = \int_0^u f(s)ds = \frac{1}{p+1}|u|^{p+1} \pm \frac{1}{q+1}|u|^{q+1}.$$

We define also the functional energy

$$E(r) = \frac{1}{2}u'(r)^2 + F(u(r))$$

which satisfies easily

$$E'(r) = -\frac{d-1}{r}u'(r)^2 < 0$$

which means that it is a non-increasing function of the variable r .

Relatively to the convexity–concavity (sub-linearity/super-linearity) parameters p and q of the problem there are 10 cases to investigate as shown in Figure 1. In the rest of the chapter we will consider the defocusing case $\lambda = \mp 1$ in the model function f defined by (12). This is always possible by acting a scaling on the function u in (36). Indeed, for $\lambda > 0$ let $K, \alpha > 0$ be positive real numbers such that

$$\lambda = K^{p-q} \quad \text{and} \quad \alpha^2 = K^{p-1}.$$

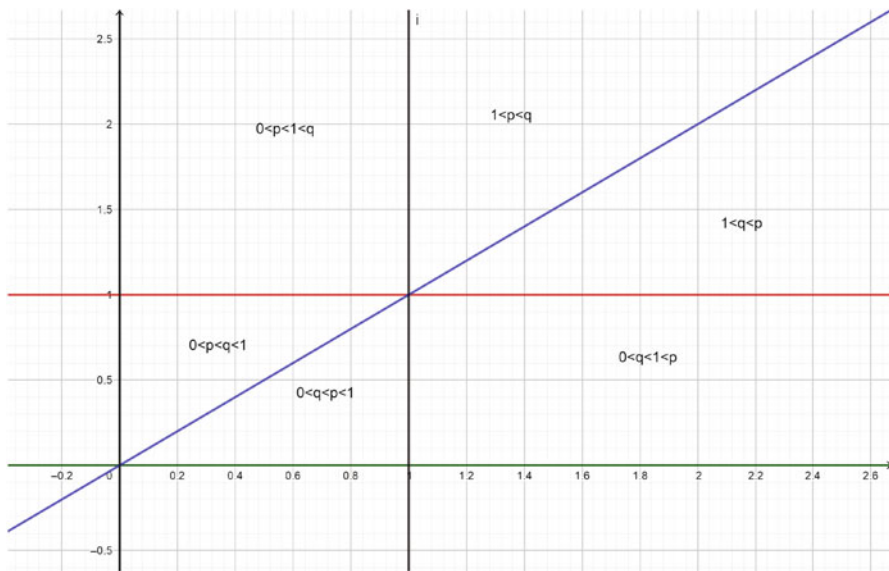


Fig. 1 The different regions relatively to p and q

The function v defined by $u(r) = Kv(\alpha r)$ satisfies

$$\begin{cases} v'' + \frac{N-1}{t}v + |v|^{p-1}v \pm |v|^{q-1}v = 0, & t \in (0, \infty) \\ v'(0) = 0, & v(0) = \frac{a}{K}. \end{cases} \tag{13}$$

Consequently, in the rest of the chapter our study will focus on the radial problem

$$\begin{cases} u'' + \frac{d-1}{t}u + |u|^{p-1}u - |u|^{q-1}u = 0, & t \in (0, \infty) \\ u(0) = a, & u'(0) = 0, \end{cases} \tag{14}$$

where $d \geq 1$ stands for the dimension of the Euclidean space \mathbb{R}^d .

2.1 Mixed Sub-linear Defocusing Case $0 < p < q < 1$

In this case, we consider the nonlinear model functions g, f , and F with $\pm\lambda = -1$ and $0 < p < q < 1$ as follows:

$$g(u) = |u|^{p-1} - |u|^{q-1},$$

$$f(u) = |u|^{p-1}u - |u|^{q-1}u$$

and

$$F(u) = \frac{1}{p+1}|u|^{p+1} - \frac{1}{q+1}|u|^{q+1}.$$

By analogy with the behavior of the present models, similar problems have been already studied in [11–14]. Existence, uniqueness, nodal solutions, and group invariant ones have been studied in detail.

The first result in the present work dealing with nodal radial solutions is stated as follows. Let

$$u_{p,q} = \left(\frac{p}{q}\right)^{\frac{1}{q-p}}, \quad \underline{u}_{p,q} = \left(\frac{1+q}{1+p}\right)^{\frac{1}{q-p}} \quad \text{and} \quad \bar{u}_{p,q} = \left(\frac{1-p}{1-q}\right)^{\frac{1}{q-p}}.$$

We have immediately

$$0 < u_{p,q} < 1 < \underline{u}_{p,q} < \bar{u}_{p,q}$$

and

$$f'(u_{p,q}) = F(\underline{u}_{p,q}) = g'(\overline{u}_{p,q}) = 0.$$

Remark that

- g is even, non-increasing on $(0, \overline{u}_{p,q})$, and non-decreasing on $(\overline{u}_{p,q}, +\infty)$ with $g(1) = 0$, $\lim_{u \rightarrow 0} g(u) = +\infty$ and $\lim_{u \rightarrow +\infty} g(u) = 0$.
- f is odd, non-decreasing on $(0, u_{p,q})$, and non-increasing on $(u_{p,q}, +\infty)$ with $f(0) = f(1) = 0$ and $\lim_{u \rightarrow +\infty} f(u) = -\infty$.
- F is even, non-increasing on $(0, 1)$, and non-decreasing on $(1, +\infty)$ with $F(0) = F(u_{p,q}) = 0$ and $\lim_{u \rightarrow +\infty} F(u) = 0$.

Theorem 1 *Assume that $u(0) = a \in]-1, 1[\setminus\{0\}$. The solution of problem (14) is extended to ∞ and oscillatory. Furthermore, there exists $(t_k)_k$ and $(z_k)_k$ satisfying $u(z_k) = u'(t_k) = 0$ and*

$$0 = t_0 < z_1 < t_1 < z_2 < t_2 < \dots < z_k < t_k < z_{k+1} < \dots \uparrow +\infty. \tag{15}$$

Moreover $u(t_k)$ is non-increasing to 0 as k goes to ∞ .

Proof Without loss of the generality, we may assume that $a \in (0, 1)$. At $r = 0$, we get

$$du''(0) = -f(a) < 0.$$

Consequently, $u''(r) < 0$ on some small interval $(0, \delta)$. This yields that $u'(r) < 0$ on $(0, \delta)$ and that u is non-increasing on $(0, \delta)$. Furthermore, $u(r) < a, \forall r \in (0, \delta)$. It holds that

$$-a < u(r) < a, \forall r \in (0, \infty). \tag{16}$$

Indeed, if it is not. Let $r_0 \in (0, \infty)$ be the first point such that $u(r_0) = \pm a$. At this point the functional energy $E(r)$ satisfies

$$E(r_0) = \frac{1}{2}u'(r_0)^2 + F(a) < E(0) = F(a).$$

As a consequence of (16), we get

$$g(u(r)) \geq L = g(a) > 0, \forall r \in (0, \infty). \tag{17}$$

Moreover, Equation (16) implies that u cannot remain non-increasing on the whole interval $(0, \infty)$. Let next v be the solution of the problem

$$\begin{cases} v'' + \frac{d-1}{r}v' + Lv = 0, & r \in (0, \infty) \\ v'(0) = 0, & v(0) = a \end{cases} \tag{18}$$

It is straightforward that v is oscillatory. So, let $\zeta < \xi$ be two consecutive zeros of v with $v(r) > 0$ for all $r \in]\zeta, \xi[$. We will prove that u vanishes at least once on $] \zeta, \xi [$. If the latter is not the case, then $u(r) < 0$ or $u(r) > 0$ on $] \zeta, \xi [$. Suppose $u(r) > 0$. By multiplying the Equations (36) and (18) by v and u , respectively, and integrating on $] \zeta, \xi [$, we obtain

$$\xi^{d-1}\omega(\xi) - \zeta^{d-1}\omega(\zeta) = \int_{\zeta}^{\xi} s^{d-1}(L - f(u(s)))u(s)v(s)ds,$$

where $\omega = u'v - uv'$. So that

$$\zeta^{d-1}u(\zeta)v'(\zeta) - \xi^{d-1}u(\xi)v'(\xi) < 0$$

which is not true. Then, since it is well known that v is oscillatory, u is also oscillatory.

We now prove the existence of the sequence $(t_k)_k$. Notice that the existence of t_k follows from a simple application of Rolle's theorem. Suppose now that $u(r) > 0$ for $r \in]z_k, z_{k+1}[$ and that

$$\text{cardinality}(\{ T \in]z_k, z_{k+1}[; u'(t) = 0 \}) \geq 2.$$

Let $\xi < \zeta$ be two consecutive zeros of u' in $]z_k, z_{k+1}[$. Then, Equation (36) shows that $f(s) = 0$ for $s \in]\xi, \zeta[$. Which is contradictory since u is not constant on $] \xi, \zeta [$.

Theorem 2 *Whenever $u(0) = a > 1$, problem (14) has a unique solution u which is strictly increasing to ∞ .*

Proof The existence and uniqueness are results of the well known Cauchy-Lipschitz Theorem which is satisfied here. It suffices to consider the system

$$\begin{cases} v = u' \\ v' = \frac{1-d}{r}u - f(u) \\ u(0) = a, \quad v(0) = 0. \end{cases}$$

Setting next

$$X = \begin{pmatrix} u \\ v \end{pmatrix}$$

we get

$$\begin{cases} X' = \tilde{F}(r, X), \\ X(0) = \begin{pmatrix} a \\ 0 \end{pmatrix}, \end{cases}$$

where

$$\tilde{F}(r, X) = \begin{pmatrix} v \\ \frac{1-d}{r}u - f(u) \end{pmatrix}.$$

It is straightforward that \tilde{F} is Lipschitz continuous relatively to the variable X . Hence, the application of Cauchy–Lipschitz Theorem follows.

We now study the behavior of the solution u . For $r = 0$, the Equation (36) yields that

$$du''(0) = -f(a) > 0.$$

Consequently, $u''(r) > 0$ on a small interval $(0, \delta)$. Similarly to the previous case, we conclude that u is non-decreasing on $(0, \delta)$. We claim that the solution u remains non-decreasing on $(0, \infty)$. Indeed, if it is not, let $r_0 > 0$ be the first critical point of u . That is, r_0 is the first point satisfying

$$r_0 > 0, \quad u'_a(r_0) = 0.$$

We have

$$(r^{d-1}u')' = -r^{d-1}f(u(r)).$$

Integrating from 0 to r_0 , we obtain

$$\int_0^{r_0} (r^{d-1}u')' dr = - \int_0^{r_0} r^{d-1} f(u(r)) dr.$$

Or equivalently,

$$\int_0^{r_0} r^{d-1} f(u(r)) dr = 0.$$

Which leads to a contradiction as $f(u(r)) < 0$ on $(0, r_0)$. Hence, the solution u is strictly increasing on $(0, +\infty)$. If it is bounded, the limit on ∞ will be one of the zeros of the function f . However, f vanishes at 0 and ± 1 . Observing that $u \geq a > 1$ on $(0, +\infty)$, we get a contradiction.

Lemma 1 *Whenever $u(0) = 1$ (respectively, $u(0) = 0$), problem (36) has the unique trivial solution $u \equiv 1$ (respectively, $u \equiv 0$).*

2.2 The Defocusing Sub-linear Case $p = 1, 0 < q < 1$

In this section we consider the defocusing sub-linear problem

$$\begin{cases} u'' + \frac{d-1}{r}u' + u - |u|^{q-1}u = 0, & r \in (0, +\infty), \\ u(0) = a, & u'(0) = 0. \end{cases} \tag{19}$$

In the present case, the generic functions are expressed as follows:

$$g(u) = 1 - |u|^{q-1},$$

$$f(u) = ug(u) = u - |u|^{q-1}u,$$

and

$$F(u) = \frac{1}{2}u^2 - \frac{1}{q+1}|u|^{q+1}.$$

The essential points included in the study are

$$u_q = \left(\frac{1}{q}\right)^{\frac{1}{q-1}} < 1 < \bar{u}_q = \left(\frac{q+1}{2}\right)^{\frac{1}{q-1}}.$$

The first result in this section is the following.

Theorem 3 *Whenever $u(0) = a \in]-1; 1[\setminus\{0\}$, the problem (19) has a unique solution which is oscillating around ± 1 with a finite number of zeros. Furthermore, there exist unique sequences $(t_k)_k$ and $(r_k)_k$ such that*

$$r_k < t_k < r_{k+1}, \quad u(r_k) = \pm 1, \quad u'(t_k) = 0, \quad k \geq 1. \tag{20}$$

Proof It is easy to see that u is non-decreasing on a small interval $(0, \delta)$. If it remains non-decreasing on $(0, \infty)$, two possibilities may occur. $u \nearrow \infty$ or $u \nearrow 1$ as $r \nearrow \infty$. In the first case, we get for $d = 1$ and r large enough,

$$u(r) < -\frac{r^2}{2} + K_1r + K_2,$$

for some constants K_1 and K_2 . For $d = 2$, we obtain similarly,

$$u(r) < -\frac{r^2}{4} + K_1 \log(r) + K_2,$$

for r large enough. Finally, for $d > 2$

$$u(r) < -\frac{r^2}{2d} + \frac{K_1}{r^{d-2}} + K_2,$$

for r large enough, where already K_1, K_2 are constants in \mathbb{R} . Consequently, $u \rightarrow -\infty$ as $r \rightarrow \infty$, which is a contradiction. Now, if the second case occurs, it implies that u behaves at ∞ like the solution v of the problem

$$\begin{cases} v'' + \frac{d-1}{r}v' + (1-q)(v-1) = 0, & r \in (0, \infty) \\ v'(0) = 0 & , v(0) = a \end{cases} \tag{21}$$

Observing that v is oscillatory, we obtain a contradiction.

It results from these cases that u cannot be non-decreasing on the whole interval $(0, \infty)$. So, it is oscillatory. We claim that u oscillates indefinitely around 1. Indeed, let t_1 be the first point in $(0, +\infty)$ such that $u'(t_1) = 0$. It holds that $u(t_1) > 1$. If not, by integrating Equation (36) from 0 to t_1 we obtain

$$0 = -\int_0^{t_1} r^{d-1} f(u(r)) > 0,$$

which is contradictory. Thus, u crosses the line $y = 1$ once in $(0, t_1)$ leading to a unique point $r_1 \in (0, t_1)$ such that $u(r_1) = 1$. Next, using similar techniques, we prove that u cannot remain greater than 1 in the rest of its domain. (Consider the same equation on $(t_1, +\infty)$ with initial data $u(t_1)$ and $u'(t_1)$). Consequently we prove that there exist unique sequences $(t_k)_k$ and $(r_k)_k$ such that

$$r_k < t_k < r_{k+1}, \quad u(r_k) = 1, \quad u'(\zeta_k) = 0, \quad k \geq 1. \tag{22}$$

Next, observing that E is decreasing as a function of r , we deduce that the sequence of maxima $(u(t_k))_k$ goes to 1 and therefore u .

Theorem 4 *Whenever $u(0) = a \in]1, \bar{u}_q[$, the problem (14) has a unique solution u which is oscillatory around ± 1 . Furthermore, there exist unique sequences $(t_k)_k$ and $(r_k)_k$ such that*

$$r_k < t_k < r_{k+1}, \quad u(r_k) = \pm 1, \quad u'(t_k) = 0, \quad k \geq 1. \tag{23}$$

Proof We proceed as previously by studying the behavior of the solution u at the origin. We have in the present case

$$du''(0) = -f(a) < 0.$$

Consequently, $u''(r) < 0$ on an interval $(0, \varepsilon)$, for some ε with $0 < \varepsilon \ll 1$. Hence u' is non-increasing on $(0, \varepsilon)$. Which means that

$$u'(r) < u'(0) = 0, \quad \forall r \in (0, \varepsilon).$$

Therefore, u is non-increasing on $(0, \varepsilon)$. We now study the possibility that u remains or not non-increasing on $(0, +\infty)$. Four situations should be investigated.

Case 1. u is decreasing with limit $-\infty$ as r goes to $+\infty$. It follows that $f(u(r)) \rightarrow -\infty$ whenever $r \rightarrow +\infty$. There exists thus $r_0 > 0$ (large enough) such that

$$f(u(r)) < -1, \quad \forall r > r_0.$$

Using Equation (19) this yields that

$$(r^{d-1}u')' > r^{d-1}, \quad \forall r > r_0.$$

Integrating from r_0 to $r > r_0$, we obtain

$$u'(r) > \frac{r}{d} + Kr^{1-d}, \quad \forall r > r_0,$$

where

$$K = r_0^{d-1}u'(r_0) - \frac{r_0^d}{d}.$$

For $d = 1$, this results in

$$u(r) > \frac{r^2}{2} + Kr - \frac{r_0^2}{2} - Kr_0 + u(r_0), \quad \forall r > r_0.$$

Which is contradictory with the fact that $\lim_{r \rightarrow +\infty} u(r) = -\infty$.

For $d = 2$, we obtain

$$u(r) > \frac{r^2}{4} + K \log(r) - \frac{r_0^2}{4} - K \log(r_0) + u(r_0), \quad \forall r > r_0.$$

Which is contradictory for the same reason as previously.

For $d > 2$, we get

$$u(r) > \frac{r^2}{2d} + \frac{K}{2-d}r^{2-d} - \frac{r_0^2}{2d} - \frac{K}{2-d}r_0^{2-d} + u(r_0), \quad \forall r > r_0.$$

Which is contradictory also for the same reason.

Case 2. The solution u is decreasing with limit 1 as r goes to $+\infty$. In this case considering the behavior of f near the point 1 we get

$$f(u) = (1 - q)(u - 1) + (u - 1)\varepsilon(u - 1),$$

where $\varepsilon(t) \rightarrow 0$ whenever $t \rightarrow 0$. So, the solution u behaves like the solution v of the equation

$$v'' + \frac{d - 1}{r}v + (1 - q)(v - 1) = 0$$

whenever $r \rightarrow +\infty$. Hence, as the solution v is oscillating around 1 indefinitely, we get a contradiction.

Case 3. The solution u is decreasing with limit -1 as r goes to $+\infty$. It may be checked by similar arguments as the previous case. So, this case cannot occur also.

Case 4. The solution u is decreasing with limit 0 as r goes to $+\infty$. Consider the energy functional associated with problem (19),

$$E(r) = \frac{1}{2}u'^2(r) + F(u(r)).$$

It is straightforward that E is decreasing. Hence,

$$E(r) < E(0) = F(a) < 0, \quad \forall r > 0.$$

Consequently,

$$0 = \lim_{r \rightarrow +\infty} E(r) \leq E(0) = F(a) < 0.$$

Which is contradictory. As a result, u is oscillatory. By following similar techniques as above, we prove the remaining part of the theorem.

The following lemma yields a localization of the critical points t_k of u .

Lemma 2 *Let $a \in]1, \bar{u}_q[$ and u the solution of (19). Let $r_0 > 0$ be the first critical point of u . Then $-a < u(r_0) < -1$.*

Indeed, assume by the contrary that $u(r_0) \geq 1$. We have

$$(r^{d-1}u'(r))' = -r^{d-1}f(u(r)), \quad \forall r.$$

Integrating on the interval $(0, r_0)$ we obtain

$$0 = \int_0^{r_0} (r^{d-1}u'(r))' dr = - \int_0^{r_0} r^{d-1}f(u(r))dr < 0.$$

Which is a contradiction. Consequently, for the critical point r_0 , the following situations should be examined:

1. $0 < u(r_0) < 1$.
2. $u(r_0) = 0$.
3. $-1 < u(r_0) < 0$.
4. $u(r_0) = -1$.
5. $-a < u(r_0) < -1$.

Case 1: $u(r_0) = b \in]0, 1[$ and $u'(r_0) = 0$. So, at $r = r_0$ that $u''(r_0) = -f(b) > 0$. Thus, there exists $\varepsilon > 0$ small enough for which $u''(r) > 0$ on $(r_0 - \varepsilon, r_0 + \varepsilon)$. Hence, u' is non-decreasing on $(r_0 - \varepsilon, r_0 + \varepsilon)$. Consequently, r_0 is a local minimum of u . We claim that u is not increasing on the whole interval $(r_0, +\infty)$. Indeed, by similar arguments as above, it is straightforward that $|u(r)| \leq a$ for all r . Hence, whenever u is increasing on $(r_0, +\infty)$, it has the only limit $l = 1$ as $r \rightarrow +\infty$. Thus, as previously, we prove that u becomes oscillating infinitely around 1. Which contradicts the fact of being increasing on $(r_0, +\infty)$. So, let r_1 be the next critical point of u . Two situations may hold. $b < u(r_1) \leq 1$ or $u(r_1) > 1$. In the first case, we get immediately

$$0 = \int_{r_0}^{r_1} (r^{d-1} u')' dr = - \int_{r_0}^{r_1} r^{d-1} f(u(r)) dr > 0.$$

Which is contradictory. Hence, the second case occurs. Continuing with similar techniques, we prove that the solution u is oscillating infinitely around 1.

Case 2: $u(r_0) = u'(r_0) = 0$. As $a \in]1, \bar{u}_q]$ we get $E(a) = F(a) \leq 0$ and $E(r_0) = 0 < E(a)$ which is contradictory.

Case 3: $u(r_0) = b \in]-1, 0[$ and $u'(r_0) = 0$. Recall firstly that $\forall r \in (0, +\infty)$, we have

$$-a \leq u(r) \leq a.$$

As $a \in (1, \bar{u}_q)$, we get from the behavior of F that $F(a) < F(b)$. Consequently, $E(r_0) = F(b) > E(0) = F(a)$ which is contradictory with the fact that E is decreasing.

Case 4. $u(r_0) = -1$ and $u'(r_0) = 0$. Let r_1 be such that

$$0 < r_1 < r_0 \quad \text{and} \quad u(r_1) = 0.$$

It holds that u is decreasing on (r_1, r_0) and that $u(r) \in (-1, 0)$ for all $r \in (r_1, r_0)$. Furthermore, $u'(r_1) < 0$ and $f(u(r)) > 0$ for all $r \in (r_1, r_0)$. Therefore, by multiplying Equation (19) by r^{d-1} and integrating on (r_1, r_0) we get

$$0 < -r_1^{d-1} u'(r_1) = - \int_{r_1}^{r_0} r^{d-1} f(u(r)) dr < 0.$$

Which is contradictory.

Case 5. $-a < u(r_0) = b < -1$ and $u'(r_0) = 0$. Whenever $a \in (1, \bar{u}_q)$, we get $F(b) > F(a)$. Which means in terms of energy that $E(r_0) > E(0)$. Which is a contradiction. Next, for $a \in (\bar{u}_q, +\infty)$, r_0 is the first local minimum for u . As previously, u cannot be increasing on the whole interval $(r_0, +\infty)$. Let r_1 the next critical point of u . Hence, r_1 is a local maximum obviously. By following similar techniques as above, we prove that u oscillates indefinitely around ± 1 .

Theorem 5 *Whenever $u(0) = a \in]\bar{u}_q, +\infty[$, the problem (19) has a unique solution u which is oscillating infinitely around ± 1 with its limit being ± 1 .*

Proof Proceeding as previously, it holds that u is non-increasing on a small interval $(0, \varepsilon)$. If it remains non-increasing on the whole interval $(0, \infty)$, it should decrease to the limit 0 as $r \rightarrow \infty$. Denote

$$h(r) = -\frac{d-1}{r}u' + u^q.$$

We get in one hand

$$u'' + u = h(r).$$

On the other hand, it is easy to see that $h(r) \downarrow 0$ as $r \rightarrow +\infty$. Next, from standard techniques we get

$$u(r) = \left(C_0 - \int_0^r h(u(t)) \cos t dt \right) \cos r + \left(C_1 + \int_0^r h(u(t)) \cos t dt \right) \sin r.$$

It suffices now to prove that the integrals

$$\int_0^{+\infty} h(u(t)) \cos t dt \quad \text{and} \quad \int_0^{+\infty} h(u(t)) \sin t dt$$

are convergent which is true due to Abel’s criterion for integrals. Hence, for $r \rightarrow +\infty$, we may have the estimation

$$u \sim A \cos(r) + B \sin(r)$$

for r large enough, which is contradictory with the monotony of u .

We now study the positions of the critical points of u as previously. Recall firstly that $-a < u(r) < a$ on $(0, \infty)$. Let next $r_0 > 0$ be the first strictly positive critical point of u . As previously, we may have many situations:

1. $u(r_0) \geq 1$.
2. $0 < u(r_0) < 1$.
3. $u(r_0) = 0$.

4. $-1 < u(r_0) < 0$.
5. $u(r_0) = -1$.
6. $-a < u(r_0) < -1$.

Case 1: $u(r_0) = b \in [1, a[$ and $u'(r_0) = 0$. By integrating Equation (36) on the interval $(0, r_0)$ we obtain

$$0 = \int_0^{r_0} (r^{d-1}u'(r))' dr = - \int_0^{r_0} r^{d-1} f(u(r)) dr < 0.$$

Which is a contradiction.

Case 2: $u(r_0) = b \in]0, 1[$ and $u'(r_0) = 0$. It holds at $r = r_0$ that $u''(r_0) = -f(b) > 0$. Thus, there exists $\varepsilon > 0$ small enough for which $u''(r) > 0$ on $(r_0 - \varepsilon, r_0 + \varepsilon)$. Hence, u' is non-decreasing on $(r_0 - \varepsilon, r_0 + \varepsilon)$. Consequently, r_0 is a local minimum of u . We claim that u is not increasing on the whole interval $(r_0, +\infty)$. Indeed, by similar arguments as above, whenever u is increasing on $(r_0, +\infty)$, it has the only finite positive limit $l = 1$ as $r \rightarrow +\infty$. Thus, as previously, we prove that u becomes oscillating infinitely around 1. Which contradicts the fact of being increasing on $(r_0, +\infty)$. So, let r_1 be the next critical point of u , ($r_1 > r_0$). Two situations may hold. $b < u(r_1) \leq 1$ or $u(r_1) > 1$. In the first case, we get immediately

$$0 = \int_{r_0}^{r_1} (r^{d-1}u')' dr = - \int_{r_0}^{r_1} r^{d-1} f(u(r)) dr > 0.$$

Which is contradictory. Hence, the second case occurs. Continuing with similar techniques, we show that the solution u is oscillating infinitely around its limit being equal to 1.

Case 3: $u(r_0) = u'(r_0) = 0$. As $a \in]\bar{u}_q, +\infty[$, we obtain $g(u(r)) \rightarrow -\infty$ whenever $r \rightarrow r_0$. Consequently, for all $A > 0$, there exists $\eta > 0$ small enough such that

$$g(u(r)) < -A, \quad \forall r \in (r_0 - \eta, r_0 + \eta).$$

Consequently,

$$u(r)g(u(r)) < -Au(r), \quad \forall r \in (r_0 - \eta, r_0 + \eta).$$

Which yields that

$$u'' + \frac{d-1}{r}u' + ug(u) < u'' + \frac{d-1}{r}u' - Au, \quad \forall r \in (r_0 - \eta, r_0 + \eta). \quad (24)$$

Now, observe that

$$u(r_0) = u'(r_0) = u''(r_0) = 0.$$

At the point $r = r_0$, Equation (24) leads to a contradiction.

Case 4: $u(r_0) = b \in]-1, 0[$ and $u'(r_0) = 0$. As $a \in (\bar{u}_q, +\infty)$, we get at $r = r_0$, $u''(r_0) = -f(b) < 0$. Consequently, $u''(r)$ remains negative on a small interval $I_\varepsilon = (r_0 - \varepsilon, r_0 + \varepsilon)$ for some $\varepsilon > 0$ small enough. Hence, u' is decreasing on I_ε . Which in turn yields that r_0 is a local maximum of u . Which is contradictory.

Case 5: $u(r_0) = -1$ and $u'(r_0) = 0$. Let r_1 be such that

$$0 < r_1 < r_0 \quad \text{and} \quad u(r_1) = 0.$$

It holds that u is decreasing on (r_1, r_0) and that $u(r) \in (-1, 0)$ for all $r \in (r_1, r_0)$. Furthermore, $u'(r_1) < 0$ and $f(u(r)) > 0$ for all $r \in (r_1, r_0)$. Therefore, by multiplying Equation (19) by r^{d-1} and integrating on (r_1, r_0) we get

$$0 < -r_1^{d-1}u'(r_1) = - \int_{r_1}^{r_0} r^{d-1} f(u(r))dr < 0.$$

Which is contradictory.

Case 6: $-a < u(r_0) = b < -1$ and $u'(r_0) = 0$. As $a \in (\bar{u}_q, +\infty)$, r_0 is the first local minimum for u . As previously, u cannot be increasing on the whole interval $(r_0, +\infty)$. Let r_1 be the next critical point of u . Hence, r_1 is a local maximum obviously. The position of $u(r_1)$ may be localized as previously, so that we get oscillations around ± 1 .

Theorem 6

1. For all $a \neq 0$, problem (19) has a unique solution u .
2. Whenever $u(0) = a \in (0, \bar{u}_q)$, the following situation holds:

$$u(\zeta) = 0, \text{ for some } \zeta \implies u'(\zeta) \neq 0.$$

Furthermore, the solution u cannot be compactly supported.

3. For all $a \in (0, \bar{u}_q)$, problem (19) has a unique positive solution u .

Proof

1. We explicit the case $a \in (0, 1)$. The remaining cases may be treated by analogue arguments. Denote

$$M_a = \left\{ u \in C((0, \delta)); a \leq u(r) \leq 2a, \quad \forall r \in (0, \delta) \right\},$$

where

$$0 < \delta < \min \left\{ \sqrt{\frac{2}{|f'(\frac{a}{2})|d}}, \sqrt{\frac{2}{d}}, \sqrt{\frac{ad}{2|f(\bar{u}_q)|}} \right\}. \tag{25}$$

It is easy that u satisfies

$$u(r) = a - r^2 \int_0^1 \int_0^1 x s^{d-1} f(u(rxs)) ds dx.$$

Denote next $\Phi : M_a \rightarrow M_a \cap C^2$ defined by

$$\Phi(u(r)) = a - r^2 \int_0^1 \int_0^1 x s^{d-1} f(u(rxs)) ds dx.$$

Φ is well defined because of the fact that

$$\begin{aligned} |\phi(u(r)) - a| &\leq \left| r^2 \int_0^1 \int_0^1 x s^{d-1} f(u(rxs)) ds dx \right| \\ &\leq \frac{\delta^2}{d} |f(p)| \\ &< \frac{\delta}{2}. \end{aligned}$$

Hence, $\Phi(u) \in M_a$. On the one hand,

$$\Phi(u(r)) = a - \int_0^r \int_0^1 s^{d-1} (f(u(ts))) ds dt$$

is a primitive of

$$\varphi(t) = -\frac{1}{t^{d-1}} \int_0^t x^{d-1} f(u(x)) dx$$

which is C^2 on $(0, \delta)$ because of the continuity of the function ψ defined by

$$\psi(x) = x^{d-1} f(u(x)), \quad x \in (0, \delta).$$

We will prove that Φ satisfies the fixed point theorem. Indeed, let $u, v \in M_a$. We may write that

$$\begin{aligned} \|\phi(u) - \phi(v)\|_\infty &\leq K \|u - v\|_\infty r^2 \int_0^1 \int_0^1 x s^{d-1} ds dx \\ &\leq K \frac{d\delta^2}{2} \|u - v\|_\infty \end{aligned}$$

where $K = \max(|f'(\frac{a}{2})|, 1)$. Hence, Φ is contractive due to (25). Thus, Φ has a unique fixed point $u \in M_a$. Consequently, for all $a \neq 0$, problem (19) has a unique solution u .

We will prove now that the solution u of problem (19) is global on $(0, +\infty)$. Suppose by contrast that there exists $t_0 \in (0, +\infty)$ such that

$$\lim_{t \rightarrow t_0} |u(t)| = +\infty.$$

As we already know that the energy function $E(r)$ is non-increasing, then

$$F(u(r)) \leq E(r) \leq E(0) = F(a) < 0, \quad \forall r > 0,$$

which yields that

$$\lim_{t \rightarrow t_0} F(|u(t)|) = +\infty < F(a),$$

which is a contradiction.

2. Assume that the assertion holds. We get

$$0 = F(u(\zeta)) = E(\zeta) < E(0) = F(a) < 0.$$

Which is a contradiction. Next, whenever the solution u is compactly supported, it satisfies the assertion above with ζ being the upper bound of its support. So, it is not possible.

3. We will prove in this part that the solutions already found in Theorem 6 are strongly related to the initial values in the sense that whenever $u(0) = a \in (0, \bar{u}_q)$, the oscillations are around 1 (the positive zero of f) and that these solutions remain in fact positive on $(0, \infty)$. Indeed, by evaluating the energy $E(r)$, we get

$$F(u(r)) \leq E(r) \leq E(0) = F(a) < 0, \quad \forall r > 0.$$

Since F is even and coercive, there exists a unique positive $\alpha \neq a$ which satisfies $F(a) = F(\alpha)$. We immediately get

$$\alpha \leq u(r) \leq a \quad \text{or} \quad a \leq u(r) \leq \alpha.$$

Hence, the result follows.

2.3 The Mixed Sub-linear/Linear Case $0 < p < 1$ and $q = 1$

In this case the problem (36) becomes

$$\begin{cases} u'' + \frac{d-1}{r}u' + |u|^{p-1}u - u = 0, & r \in (0, +\infty), \\ u(0) = a, \quad u'(0) = 0. \end{cases} \quad (26)$$

The key functions will be expressed as

$$g(u) = |u|^{p-1} - 1, \quad f(u) = |u|^{p-1}u - u \text{ and } F(u) = \frac{1}{p+1}|u|^{p+1} - \frac{1}{2}u^2.$$

Denote similarly to previous cases

$$u_p = \left(\frac{1}{p}\right)^{\frac{1}{p-1}} \text{ and } \bar{u}_p = \left(\frac{p+1}{2}\right)^{\frac{1}{p-1}}.$$

We observe obviously that

$$u_p < 1 < \bar{u}_p \text{ and } f'(u_p) = F(\bar{u}_p) = 0.$$

The first main result of this part is stated as follows.

Theorem 7 *Whenever $u(0) = a \in]-1, 1[\setminus\{0\}$, the solution u is oscillating infinitely around 0. Furthermore, whenever $(z_k)_k$ is the sequence of the nodes of u on $(0, +\infty)$, we have $u'(z_k) \neq 0$.*

Proof Without loss of the generality we may assume that $a \in (0, 1)$. Let $L > 0$ be such that

$$L = \inf_{s \in (-a, a)} g(s).$$

We will apply Sturm’s comparison theorem to prove that u is oscillatory around 0 infinitely on $(0, +\infty)$. Indeed, consider the solution v of the problem

$$\begin{cases} v'' + \frac{d-1}{r}v' + Lv = 0, & r \in (0, +\infty), \\ v(0) = a, & v'(0) = 0. \end{cases} \tag{27}$$

It is straightforward that v is oscillating around 0 infinitely on $(0, +\infty)$. Let $\xi < \zeta$ be two consecutive zeros of v on $(0, +\infty)$ and assume that u is non-sign changing on (ξ, ζ) (for instance $u > 0$ on (ξ, ζ)). By multiplying Equations (26) and (27) by $r^{d-1}v$ and $r^{d-1}u$, respectively, integrating on (ξ, ζ) and subtracting we get

$$0 < \zeta^{d-1}v'(\zeta)u(\zeta) - \xi^{d-1}v'(\xi)u(\xi) = \int_{\xi}^{\zeta} r^{d-1}[L - g(u)]uvdr < 0$$

which is a contradiction. As a result, u vanishes at least once on (ξ, ζ) .

We now prove the second part of the Lemma. Assume that for some k the assertion

$$u(z_k) = u'(z_k) = 0$$

holds. So, for some $\varepsilon > 0$ small enough we have

$$u'' + \frac{d-1}{r}u' + ug(u) > u'' + \frac{d-1}{r}u' + 20u, \quad r \in (z_k - \varepsilon, z_k + \varepsilon).$$

At the point $r = z_k$ we obtain a contradiction.

We now state the second result.

Theorem 8 *Whenever $u(0) = a \in (1, +\infty)$, the solution u is increasing towards $+\infty$ on $(0, +\infty)$. Furthermore,*

$$u(r) > -\frac{f(a)}{d}r^2, \quad \forall r \in (0, +\infty).$$

Proof At the origin $r = 0$ we have $du''(0) = -f(a) > 0$. Therefore, $u''(r) > 0$ on some interval $(0, \varepsilon)$ for some $\varepsilon > 0$ small enough. On this interval we obtain immediately $u'(r) > 0$ as it is increasing and $u'(0) = 0$. Consequently, the solution u is also increasing on $(0, \varepsilon)$. If it did not remain increasing on the whole interval $(0, +\infty)$, let $r_0 > 0$ be the first point such that $u'(r_0) = 0$. Equation (26) multiplied by r^{d-1} and integrated on $(0, r_0)$ yields that

$$0 = \int_0^{r_0} (r^{d-1}u')' dr = - \int_0^{r_0} (r^{d-1}f(u(r)))dr > 0.$$

Hence, a contradiction. Consequently, u is increasing on $(0, +\infty)$.

Let now $u_\infty = \lim_{r \rightarrow +\infty} u(r)$. Assuming that $u_\infty < +\infty$ yields that $f(u_\infty) = 0$ which is impossible as $u_\infty \geq a > 1$. It remains now to prove the last part of the theorem. Observe that $u(r) \geq a$ for all $r \in (0, +\infty)$. Consequently, as f is decreasing on $(1, +\infty)$, we obtain $f(u(r)) < f(a)$ for all $r \in (0, +\infty)$. Hence,

$$0 = u'' + \frac{d-1}{r}u' + f(u(r)) < u'' + \frac{d-1}{r}u' + f(a), \quad \forall r \in (0, +\infty).$$

Therefore,

$$(r^{d-1}u')' + r^{d-1}f(a) > 0, \quad \forall r \in (0, +\infty).$$

Integrating on the interval $(0, r)$ we obtain

$$r^{d-1}u' > -\frac{f(a)}{d}r^d, \quad \forall r \in (0, +\infty).$$

Or equivalently,

$$u'(r) > -\frac{f(a)}{d}r, \quad \forall r \in (0, +\infty).$$

As a result,

$$u(r) > -\frac{f(a)}{2d}r^2, \quad \forall r \in (0, +\infty).$$

2.4 A Mixed Sub-linear Defocusing Case $0 < q < p < 1$

Consider the problem

$$\begin{cases} u'' + \frac{d-1}{r}u' + |u|^{p-1}u - |u|^{q-1}u = 0, & r \in (0, \infty) \\ u(0) = a & , u'(0) = 0. \end{cases} \tag{28}$$

Denote

$$g(u) = |u|^{p-1} - |u|^{q-1},$$

$$f(u) = ug(u) = (|u|^p - |u|^q)sign(u)$$

and

$$F(u) = \frac{1}{p+1}|u|^{p+1} - \frac{1}{q+1}|u|^{q+1}.$$

Denote also

$$u_{p,q} = \left(\frac{q}{p}\right)^{\frac{1}{p-q}}, \quad \underline{u}_{p,q} = \left(\frac{1+p}{1+q}\right)^{\frac{1}{p-q}} \quad \text{and} \quad \bar{u}_{p,q} = \left(\frac{1-q}{1-p}\right)^{\frac{1}{p-q}}.$$

We have immediately

$$u_{p,q} < 1 < \underline{u}_{p,q} < \bar{u}_{p,q}$$

and

$$f'(u_{p,q}) = F(\underline{u}_{p,q}) = g'(\bar{u}_{p,q}) = 0.$$

The following technical lemma is needed:

Lemma 3 *For all $a \in (0, 1)$ the solution u of problem (28) is not increasing on $(0, +\infty)$. Moreover, u did not reach a again on $(0, \infty)$.*

Proof Assume that the converse is true. So, the solution u has a limit as r goes upwards $+\infty$. As $a > 0$, this limit is either 1 or $+\infty$. In the first case we get

$$f(u(r)) = (p - q)(u(r) - 1) + (u(r) - 1)o(u(r) - 1), \quad \text{as } r \rightarrow +\infty.$$

Consequently, u behaves like the solution v of the problem

$$v'' + \frac{d-1}{r}v' + (v-1) = 0$$

as $r \rightarrow +\infty$. As v is oscillatory, u is also. Which is a contradiction of being increasing. For the second case, this leads to the following assertion:

$$\forall A > 0, \exists r_0 > 0 \text{ such that } f(u(r)) > A, \forall r \in (r_0, +\infty).$$

Consequently,

$$(r^{d-1}u')' < -r^{d-1}A, \forall r \in (r_0, +\infty).$$

As a result,

$$u'(r) < -\frac{A}{d}r + \left(\frac{Ar_0^d}{d} + r_0^{d-1}u'(r_0)\right)r^{1-d}, \forall r \in (r_0, +\infty).$$

Which yields that $u'(r) \rightarrow -\infty$. Which in turn contradicts the fact of u being increasing.

The last part of the lemma is an easy application of the fact that the energy $E(r)$ is non-increasing.

As a result of this lemma we get the following first main result.

Theorem 9 $\forall a \in (0, 1)$, the solution u is oscillating around 1 infinitely on the interval $(0, +\infty)$ with a finite number of zeros.

Proof As at 0, we have $du''(0) = -f(a) > 0$, the solution u starts by increasing from the origin. As previously, u cannot be increasing on the whole interval $(0, +\infty)$. Let $r_0 > 0$ be its first critical point on $]0, +\infty)$. We claim that $u(r_0) > 1$. If not, we get from Equation (28)

$$0 = \int_0^{r_0} r^{d-1} f(u(r)) dr < 0,$$

which is contradictory. Consider next the second critical point $r_1 > r_0$ of u . Here also we claim similarly that $a < u(r_1) < 1$. If not, we get in a similar way as previously

$$0 = \int_{r_0}^{r_1} r^{d-1} f(u(r)) dr > 0,$$

which is contradictory. So, assume that we get $r_k, k = 0, 1, \dots, n$ for some $n \in \mathbb{N}$ and assume for instance that u is decreasing on (r_{n-1}, r_n) . We claim that the solution u cannot be increasing on the whole interval $(r_n, +\infty)$. Indeed, if this occurs, let L

be the limit of $u(r)$ as $r \rightarrow +\infty$. It is straightforward that $L \neq 1$ as u is not oscillatory. Next, whenever $L = +\infty$ we get as previously

$$u'(r) < -\frac{A}{d}r + \left(\frac{Ar_d^d}{d} + r_d^{d-1}u'(r_d)\right)r^{1-d}, \quad \forall r \in (r_d, +\infty)$$

for some A and r_d positive large enough. So, a contradiction. So, the two possible limits are impossible. Thus u is not strictly increasing there. Hence there exists $r_{n+1} > r_n$ such that $u'(r_{n+1}) = 0$. Now, similar techniques yield that $u(r_{n+1}) > 1$. And so on.

Now assume that we constructed $r_k, k = 0, 1, \dots, n$ for some $n \in \mathbb{N}$ and assume for instance that u is increasing on (r_{n-1}, r_n) . We claim that the solution u cannot be decreasing on the whole interval $(r_n, +\infty)$. Indeed, if this occurs, let L be the limit of $u(r)$ as $r \rightarrow +\infty$. We have in one hand $1 < a \leq L$. In the other hand, the possible values of L are $0, -1$, and $-\infty$ which are in contradiction with the bounds of it. Thus u is not strictly decreasing there. Hence there exists $r_{n+1} > r_n$ such that $u'(r_{n+1}) = 0$. Now, similar techniques yield that $a < u(r_{n+1}) < 1$. And so on. As a result u is oscillatory around 1 infinitely on $(0, +\infty)$.

Theorem 10 *For all $a \in (1, \frac{a}{p,q})$, the solution u is oscillatory around ± 1 .*

Proof We have at the point $r = 0, du''(0) = -f(a) < 0$. Hence, $u''(r) < 0$ on some small interval $(0, \varepsilon)$. Thus, u' is decreasing there on $(0, \varepsilon)$. Which means that $u' < 0$ on $(0, \varepsilon)$. As a result, u is decreasing on $(0, \varepsilon)$. Assume that u remains decreasing on the whole interval $(0, +\infty)$. So, it has a limit L at $+\infty$. Now, as $|u|$ is bounded by a , the limit may be 0 or ± 1 . The latter cannot occur as it yields that u is oscillating around ± 1 by using the equivalence $f(u) = (p - q)(u \pm 1)$ near ± 1 . Next, whenever $L = 0$, we recall that

$$E(r) = \frac{1}{2}u'^2(r) + F(u(r)) < F(a) < 0.$$

Consequently,

$$0 = E(\infty) \leq F(a) < 0.$$

Which is contradictory. Let next $r_0 > 0$ be the first critical point of u on $(0, +\infty)$. One of the following situations is true.

- (i) $u(r_0) \in]0, 1[$ and thus u is oscillating infinitely around 1.
- (ii) $u(r_0) \in]-a, -1[$ and thus u is oscillating infinitely around -1 .

Assume that the opposite cases occur instead. Whenever $u(r_0) \geq 1$ we get

$$0 = \int_0^{r_0} (r^{d-1}u'dr = - \int_0^{r_0} (r^{d-1}f(u(r))dr < 0.$$

Which is impossible. Next, for $u(r_0) = 0$, we get at $r = r_0$,

$$E(r_0) = 0 < E(r) = F(a) < 0.$$

Which is a contradiction. Now, whenever $-1 < u(r_0) = b < 0$, we get $u''(r_0) = -f(b) < 0$. So it remains negative on some small interval $(r_0 - \varepsilon, r_0 + \varepsilon)$. Hence, u' is decreasing on $(r_0 - \varepsilon, r_0 + \varepsilon)$. As $u'(r_0) = 0$, it holds that r_0 is a local maximum of u . Which is contradictory. Assume in the next case that $u(r_0) = -1$. Let r_1 be such that

$$0 < r_1 < r_0, \quad u(r_1) = 1.$$

It is immediate that $u'(r_1) < 0$ and that $f(u(r)) < 0$ on (r_1, r_0) . Hence, Equation (28) yields that

$$0 > r_1^{d-1} u'(r_1) = \int_{r_1}^{r_0} (r^{d-1} u')' dr = \int_{r_1}^{r_0} r^{d-1} f(u(r)) dr < 0.$$

Which is a contradiction.

It results from these cases that only the situations (i) and (ii) above may occur.

Theorem 11 *For all $a \in (\underline{u}_{p,q}, +\infty)$, the solution u is oscillatory around ± 1 with finite number of zeros. Furthermore, whenever ζ is a zero of u on $(0, +\infty)$, we have $u'(\zeta) \neq 0$.*

Proof We have as previously u is decreasing on $(0, \varepsilon)$ for some $\varepsilon > 0$ small enough. Whenever u remains decreasing on the whole interval $(0, +\infty)$, it has a limit L at $+\infty$. Now, as $|u|$ is bounded by a , the limit may be 0 or ± 1 . The latter cannot occur as it yields that u is oscillating around ± 1 by using the equivalence $f(u) = (p - q)(u \pm 1)$ near ± 1 . Next, for $L = 0$, consider the function

$$g(r) = -\frac{d-1}{r} u' + u - f(u).$$

Proceeding as previously we get a contradiction. As a result, the solution u is oscillating.

In the following part, we will determine the possible value around which the solution u oscillates infinitely.

Lemma 4 *The following situation cannot occur. There exist sequences (r_k) , (t_k) , (z_k) , and (ζ_k) satisfying*

- i. $t_{2k-1} < z_{2k-1} < \zeta_{2k-1} < z_{2k} < t_{2k} < r_{2k} < \zeta_{2k} < r_{2k+1}, \quad \forall k.$
- ii. $u(r_k) = -u(z_k) = 1, \quad u(t_k) = u'(\zeta_k) = 0, \quad \forall k.$
- iii. u is increasing strictly on $(\zeta_{2k-1}, \zeta_{2k})$ and decreasing strictly on $(\zeta_{2k}, \zeta_{2k+1}), \quad \forall k.$

Proof Assume by contrast that the situation occurs. Hence, using the functional energy $E(r)$, we observe that

$$E(\zeta_k) = F(|u(\zeta_k)|), \quad k \in \mathbb{N}$$

is a decreasing sequence. Therefore, $(|u(\zeta_k)|)_k$ is also decreasing. As it satisfies further $|u(\zeta_k)| \geq 1$, it is therefore convergent to a limit $L \geq 1$. We claim that $L = 1$. Indeed, as (ζ_k) goes to infinity and $u'(\zeta_k) = 0$ for all k , we should have $u''(\zeta_k) \rightarrow 0$ as $k \rightarrow +\infty$. This yields from the ODE satisfied by u that $f(u(\zeta_k)) \rightarrow 0$ as $k \rightarrow +\infty$. So, because of the fact that $|u(\zeta_k)| \geq 1$ for all k , the limit should be equal to 1. Observe next that for r large enough and $k \in \mathbb{N}$ unique such that

$$\zeta_{2k} \leq r < \zeta_{2k+1}$$

or

$$\zeta_{2k+1} \leq r < \zeta_{2k+2},$$

we have

$$E(\zeta_{2k}) \leq E(r) < E(\zeta_{2k+1})$$

or

$$E(\zeta_{2k+1}) \leq E(r) < E(\zeta_{2k+2}),$$

which means that

$$\lim_{r \rightarrow +\infty} E(r) = \frac{q-p}{(1+p)(1+q)}.$$

Similarly, we get

$$\lim_{k \rightarrow +\infty} E(t_k) = \frac{q-p}{(1+p)(1+q)},$$

which means that

$$\lim_{k \rightarrow +\infty} u'^2(t_k) = \frac{q-p}{(1+p)(1+q)} < 0$$

which is a contradiction.

Lemma 5 *The following situation cannot occur. There exist sequences $(r_k)_k$, $(t_k)_k$, $(\zeta_k)_k$ satisfying*

$$i. \quad t_{2k-1} < \zeta_{2k-1} < t_{2k} < r_{2k} < \zeta_{2k} < r_{2k+1} < t_{2k+1}, \quad \forall k.$$

- ii. $u(r_k) = 1$, $u(\zeta_{2k+1}) = -1$, $u(\zeta_{2k}) \geq 1$, $u(t_k) = u'(\zeta_k) = 0$, $\forall k$.
- iii. u is non-decreasing on $(\zeta_{2k-1}, \zeta_{2k})$ and non-increasing on $(\zeta_{2k}, \zeta_{2k+1})$, $\forall k$.

Proof Assume as in Lemma 4 that the situation occurs. Again, using the functional energy $E(r)$, we show that $(E(u(\zeta_{2k})))_k$ is a constant equal to $F(-1)$, which contradicts its monotony.

Lemma 6 *The following situation cannot occur. There exist sequences $(r_k)_k$, $(t_k)_k$, $(\zeta_k)_k$ satisfying*

- i. $t_{2k-1} < \zeta_{2k-1} < t_{2k} < r_{2k} < \zeta_{2k} < r_{2k+1} < t_{2k+1}$, $\forall k$.
- ii. $u(r_k) = 1$, $u(\zeta_{2k+1}) \in (-1, 0)$, $u(\zeta_{2k}) \geq 1$, $u(t_k) = u'(\zeta_k) = 0$, $\forall k$.
- iii. u is non-decreasing on $(\zeta_{2k-1}, \zeta_{2k})$ and non-increasing on $(\zeta_{2k}, \zeta_{2k+1})$, $\forall k$.

Proof Whenever the situation occurs, there holds for each k that u is non-decreasing on $(\zeta_{2k+1} - \delta, \zeta_{2k+1})$ and non-increasing on $(\zeta_{2k+1}, \zeta_{2k+1} + \delta)$. Which is a contradiction.

Lemma 7 *The following situation cannot occur. There exist sequences (r_k) , (t_k) satisfying*

- i. $t_{2k-1} < r_{2k} < t_{2k} < r_{2k+1} < t_{2k+1}$, $\forall k$.
- ii. $u(r_k) = 1$, $u(t_{2k+1}) = 0$, $u(t_{2k}) \geq 1$, $u'(t_k) = 0$, $\forall k$.
- iii. u is non-decreasing on (t_{2k-1}, t_{2k}) and non-increasing strictly on (t_{2k}, t_{2k+1}) , $\forall k$.

Proof This situation is obviously impossible as if not, we get $E(t_{2k+1}) = 0$, $\forall k$. However $(E(t_{2k}))_k$ converges to $F(1) = \frac{q-p}{(1+p)(1+q)} < 0$. Which is a contradiction.

Remark 1 Similar results may be obtained by replacing 1 by -1 in the Lemmas 4–7. As a result of these situations, we conclude that the solution u of Theorem 11 oscillates around 0, 1, or -1 . Furthermore, we have

1. Whenever u oscillates around ± 1 , it has the limit also ± 1 , respectively, as $r \rightarrow \infty$, and thus has a finite number of zeros.
2. Whenever u oscillates infinitely around 0, it has the limit 0 as $r \rightarrow \infty$.

We now claim that the last situation (2) in Remark 1 above cannot occur. Indeed, assume that it occurs and denote

$$\dots < r_{2k-1} < \zeta_{2k-1} < r_{2k} < \zeta_{2k} < r_{2k+1} < \dots$$

such that

$$u(r_k) = u'(\zeta_k) = 0, \forall k.$$

The following situations hold immediately.

- (i) The sequence $(u(\zeta_{2k}))_k$ is increasing to 1.
- (ii) The sequence $(u(\zeta_{2k+1}))_k$ is decreasing to -1 .

Which yields that the sequence $(E(\zeta_k))_k$ is convergent to $F(1)$. Next, because of the monotony of $E(r)$, we deduce that $(E(r_k))_k$ is also convergent to $F(1) < 0$. However,

$$E(r_k) = \frac{1}{2}u'(r_k)^2 \geq 0.$$

Which is a contradiction. In fact we may prove further that whenever u tends to 0 at ∞ , it behaves like the solution v of the problem

$$v'' + \frac{d-1}{r}v' - |v|^{q-1}v = 0.$$

Now, recall that v tends to ∞ when r goes to ∞ , which is contradictory. This achieves the proof.

2.5 A Mixed Sub-linear/Super-Linear Defocusing Case $0 < p < 1 < q$

In this section, we consider the problem

$$\begin{cases} u'' + \frac{d-1}{r}u' + ug(u) = 0, & r \in (0, \infty), \\ u(0) = a & , u'(0) = 0, \end{cases} \tag{29}$$

where g is always the model nonlinear function

$$g(u) = |u|^{p-1} - |u|^{q-1},$$

with a nonlinear convex part $|u|^{p-1}$, $0 < p < 1$ and a nonlinear (may be concave) one $|u|^{q-1}$, $q > 1$. We also have

$$f(u) = ug(u) = |u|^{p-1}u - |u|^{q-1}u \text{ and } F(u) = \frac{1}{p+1}|u|^{p+1} - \frac{1}{q+1}|u|^{q+1}.$$

The following points

$$u_{p,q} = \left(\frac{p}{q}\right)^{\frac{1}{q-p}} \text{ and } \bar{u}_{p,q} = \left(\frac{1+q}{1+p}\right)^{\frac{1}{q-p}}$$

satisfy

$$u_{p,q} < 1 < \bar{u}_{p,q}$$

and

$$f'(u_{p,q}) = F(\bar{u}_{p,q}) = 0.$$

Theorem 12 *Whenever $u(0) = a \in] - 1, 1[\setminus\{0\}$, the problem (29) has a unique solution u which is oscillating infinitely around 0. Furthermore, whenever $(z_k)_k$ is the sequence of the nodes of u on $(0, +\infty)$, we have $u'(z_k) \neq 0$.*

Proof Let $a \in (0, 1)$. Let $L > 0$ be such that

$$L = \inf_{s \in (-a,a)} g(s).$$

We will apply Sturm’s comparison theorem to prove that u is oscillatory around 0 infinitely on $(0, +\infty)$. Indeed, consider the solution v of the problem

$$\begin{cases} v'' + \frac{d-1}{r}v' + Lv = 0 & , \quad r \in (0, +\infty), \\ v(0) = a & , \quad v'(0) = 0. \end{cases} \tag{30}$$

It is straightforward that v is oscillating around 0 infinitely on $(0, +\infty)$. Let $\xi < \zeta$ be two consecutive zeros of v on $(0, +\infty)$ and assume that u is non-sign changing on (ξ, ζ) (For instance $u > 0$ on (ξ, ζ)). By multiplying Equations (29) and (30) by $r^{d-1}v$ and $r^{d-1}u$, respectively, integrating on (ξ, ζ) and subtracting we get

$$0 < \zeta^{d-1}v'(\zeta)u(\zeta) - \xi^{d-1}v'(\xi)u(\xi) = \int_{\xi}^{\zeta} r^{d-1}[L - g(u)]uvdr < 0$$

which is a contradiction. As a result, u vanishes at least once on (ξ, ζ) .

We now prove the second part of the Lemma. Assume that for some k the assertion

$$u(z_k) = u'(z_k) = 0$$

holds. So, for some $\varepsilon > 0$ small enough we have

$$u'' + \frac{d-1}{r}u' + ug(u) > u'' + \frac{d-1}{r}u' + u, \quad r \in (z_k - \varepsilon, z_k + \varepsilon).$$

At the point $r = z_k$ we obtain a contradiction.

Theorem 13 $\forall a \in (1, +\infty)$, the problem (29) has a unique solution u which is increasing to ∞ .

Proof We have $du''(0) = -f(a) > 0$. So $u'' > 0$ on some small interval $(0, \varepsilon)$. Hence, u' is increasing on $(0, \varepsilon)$. As $u'(0) = 0$, so $u' > 0$ on $(0, \varepsilon)$. Hence, u is

increasing on $(0, \varepsilon)$. Assume that it is not increasing on the whole interval $(0, +\infty)$. Let $r_0 > 0$ be the first critical point on $(0, +\infty)$. We get

$$0 = \int_0^{r_0} (r^{d-1} u')' dr = - \int_0^{r_0} r_1^{d-1} f(u(r)) dr > 0.$$

Which is contradictory. Consequently, u is increasing on $(0, +\infty)$ with its limit being equal to ∞ .

2.6 Mixed Super-Linear/Sub-linear Defocusing Case $0 < q < 1 < p$

Denote as in the previous cases

$$u_{p,q} = \left(\frac{p}{q}\right)^{\frac{1}{p-q}} \quad \text{and} \quad \bar{u}_{p,q} = \left(\frac{1+p}{1+q}\right)^{\frac{1}{p-q}}.$$

We have

$$u_{p,q} < 1 < \bar{u}.$$

Theorem 14 *Whenever $u(0) = a \in (-1, 1)$, the problem (14) has a unique solution u which is oscillating infinitely around ± 1 with limit ± 1 . Moreover, whenever $u(r) = \pm 1$ we have $u'(r) \neq 0$.*

Proof For $a \in (0, 1)$, the solution u starts by increasing on $(0, \delta)$ for $\delta > 0$ small enough. However, due to the energy functional $E(r)$, the potential $F(u)$ and Sturm–Liouville comparison theorem, the solution u cannot remain increasing on the whole interval $(0, \infty)$. Let $r_0 > 0$ be the first critical point of u on $(0, \infty)$. Whenever $u(r_0) \leq 1$, we get

$$0 = \int_0^{r_0} (r^{d-1} u')' dr = - \int_0^{r_0} r^{d-1} f(u) dr > 0.$$

Which is a contradiction. So, $u(r_0) > 1$. Next, let $r_1 > 0$ be the second critical point of u on $(0, \infty)$. The same argument used previously shows that $u(r_1) < 1$. Continuing the procedure, we show that u is oscillating around 1 infinitely. Applying the fact that the functional energy $E(r)$ is decreasing, we show that the limit is 1. Next, using again the functional energy $E(r)$, it holds that whenever $u(r) = \pm 1$ we obtain $u'(r) \neq 0$.

Theorem 15 *For $u(0) = a \in (1, \bar{u}_{p,q})$, the problem (14) has a unique solution u which is oscillating around 1 infinitely on $(0, +\infty)$ with its limit being equal to 1.*

Proof At the origin, we have $du''(0) = -f(a) < 0$. So, as for some previous cases, the solution u starts by decreasing on some small interval $(0, \varepsilon)$. Assume that it remains decreasing on the whole interval $(0, +\infty)$. So, it has a limit L as $r \rightarrow +\infty$. As u is bounded by $\pm a$, this limit is one of the real zeros of the nonlinear function f . So, $L = \pm 1$ or $L = 0$. The first case where $L = \pm 1$ yields that u has the same behavior at ∞ as the solution v of the problem

$$u'' + \frac{d-1}{r}u' + (p-q)(u \pm 1) = 0,$$

which is oscillating. This contradicts the fact of being decreasing. Next, whenever $L = 0$, we get using the energy E ,

$$0 = F(0) = \lim_{r \rightarrow +\infty} F(u(r)) \leq E(1) < E(0) = F(0) = 0.$$

Which is a contradiction. So the solution u is not monotone on the whole interval $(0, \infty)$.

Let next $r_0 > 0$ be the first critical point of u on $(0, +\infty)$. As because of the energy, $-a \leq u(r) \leq a$ for all $r \in (0, +\infty)$, the following situations may hold:

- (a.) $1 \leq u(r_0) < a$.
- (b.) $0 < u(r_0) < 1$.
- (c.) $u(r_0) = 0$.
- (d.) $-1 < u(r_0) < 0$.
- (e.) $u(r_0) = -1$.
- (f.) $-a < u(r_0) < -1$.

We now investigate each case to show its compatibility with the problem (14). The case (a.) seems to be impossible as it yields that

$$0 = \int_0^{r_0} (r^{d-1}u')' dr = - \int_0^{r_0} r_1^{d-1} f(u(r)) dr < 0,$$

which is a contradiction. The case (c.) yields that

$$E(r_0) = 0 < E(0) = F(a) < 0.$$

Which is impossible. The case (d.) yields that on a small interval $(r_0 - \varepsilon, r_0 + \varepsilon)$, we have u is increasing on $(r_0 - \varepsilon, r_0)$ and decreasing on $(r_0, r_0 + \varepsilon)$. Which means that r_0 is a local maximum of u . Which is contradictory.

We now investigate the case (e.) where $u(r_0) = -1$. Let r_1 be such that

$$0 < r_1 < r_0 \text{ and } u(r_1) = 0.$$

It is obvious that $u'(r_1) < 0$. By multiplying next the Equation (36) by r^{d-1} and integrating on (r_1, r_0) we obtain

$$0 > r_1^{d-1}u'(r_1) = \int_{r_1}^{r_0} r^{d-1}u'f(u(r))dr > 0.$$

Which is impossible. Now for the case (f.) which states that $-a < u(r_0) = b < -1$ we get

$$E(r_0) = F(b) < E(0) = F(a).$$

Which is contradictory with the fact that $F(b) > F(a)$. Consequently, the only possible case that may occur is (b.).

We now investigate the behavior of the solution in this case. By applying similar techniques as in the previous sections, it holds that the solution u is oscillating around 1 infinitely on $(0, +\infty)$.

Theorem 16 *Whenever $u(0) = a \in (\bar{u}_{p,q}, +\infty)$, there are three classes of solutions.*

1. *non-increasing with limit 0 at infinity (and thus positive) (this class contains also the compactly supported solutions).*
2. *oscillating around ± 1 with finite number of zeros.*

Proof

1. As previously, the solution u starts by decreasing on some small interval $(0, \varepsilon)$. Assume it remains decreasing on the whole domain $(0, \infty)$. Its limit is immediately 0. Denote $g(r) = \frac{u''(r)}{u(r)}$. It is easy to see that $g(r) \rightarrow \infty$ whenever $r \rightarrow \infty$. Let also $G(r) = \int_R^r g(t)dt$, where $R > 0$ large enough. Denote next, $X(r) = (u(r) \ u'(r))$. We obtain the two-dimensional one-order differential equation $X'(r) = A(r)X(r)$, where $A(r) = \begin{pmatrix} 0 & 1 \\ g(r) & 0 \end{pmatrix}$. Let next $\bar{A}(r) = \int_R^r A(t)dt$. We know from classical techniques of differential equations that $X(t) = \exp(\bar{A}(r))X_0$, where X_0 is a constant vector. Next, standard techniques yield that

$$u(r) = \frac{1}{2}e^{\lambda(r)} \left(C_1 + C_2 \frac{r-R}{\lambda(r)} \right) + \frac{1}{2}e^{-\lambda(r)} \left(C_1 - C_2 \frac{r}{\lambda(r)} \right),$$

where $\lambda(r) = \sqrt{G(r)(r-R)}$. Next, as $u(r) \downarrow 0$ when $r \rightarrow \infty$, we should have

$$C_1 + C_2 \lim_{r \rightarrow \infty} \frac{r - R}{\lambda(r)} = 0.$$

Whenever $C_2 = 0$, we get immediately $C_1 = 0$ and thus u is compactly supported. For $C_2 \neq 0$, we get $G(r) \sim \alpha(r - R)$ at infinity, where $\alpha = \frac{C_1^2}{C_1} = \omega^2$, ($\omega > 0$). (In fact we have here also $C_1 \neq 0$). As a result, we get $u(r) \sim Ce^{-\omega r}$ at infinity. Which is the solutions declared (positive and decreasing to 0).

2. We will show now that the solutions which did not remain decreasing on the whole domain $(0, \infty)$ are necessarily oscillating around ± 1 infinitely. This may be shown by following similar techniques as in Section 3.

2.7 Mixed Linear/Super-Linear Defocusing Case $p = 1$ and $q > 1$

In this section we are focusing on the case $p = 1$ and $q > 1$. The functions g, f , and F are written as follows:

$$g(u) = 1 - |u|^{q-1}, \quad f(u) = ug(u) = u - |u|^{q-1}u \quad \text{and} \quad F(u) = \frac{1}{2}u^2 - \frac{1}{q+1}|u|^{q+1}.$$

In this case, we have two essential points that affect the behavior of these functions and thus affect in turn the behavior of the solution(s) of problem (14). These are

$$u_q = \left(\frac{1}{q}\right)^{1/(q-1)} \quad \text{and} \quad \bar{u}_q = \left(\frac{q+1}{2}\right)^{1/(q-1)}.$$

Remark that

$$0 < u_q < 1 < \bar{u}_q \quad \text{and} \quad f'(u_q) = F(\bar{u}_q) = 0.$$

Theorem 17 *Whenever $u(0) = a \in (-1, 1)$, the problem (14) has a unique solution u which is oscillating infinitely around 0.*

Proof Denote $L = \inf_{s \in (-a, a)} g(s)$. It is easy that $L > 0$ and thus the solution v of the problem

$$v'' + \frac{N-1}{r}v' + Lv = 0, \quad v(0) = a, \quad v'(0) = 0$$

is oscillating infinitely around 0. As a consequence, the solution u is also oscillating infinitely around 0 (Sturm–Liouville theory).

Theorem 18 *Whenever $u(0) = a \in (1, +\infty)$, the problem (14) has a unique solution u which is strictly increasing to $+\infty$. More precisely,*

$$u(r) \geq a - \frac{f(a)}{2d}r^2.$$

Proof The solution u exists and is unique because of the Lipschitz theorem. It is also easy to see that u is non-decreasing on $(0, \delta)$ for some $\delta > 0$ small enough. Furthermore, because of the energy $E(r)$, the solution $u \geq a$ on its whole domain. Hence, it holds from (36) that $f(u) \leq f(a)$ on $(0, \infty)$. Which yields that

$$u(r) \geq a - \frac{f(a)}{2d}r^2.$$

Now, assume that u does not remain strictly increasing and let $r_0 > 0$ its first critical point. That is, the first point in $]0, +\infty[$ for which we have $u'(r_0) = 0$, u is strictly increasing on $(0, r_0)$ and strictly decreasing on $(r_0, r_0 + \varepsilon)$ for some $\varepsilon > 0$. Of course the solution u cannot remain constant on $(r_0, +\infty)$ also because of Lipschitz theorem. Now, multiplying the first equation in (36) by r^{d-1} and integrating from 0 to r_0 we get

$$\int_0^{r_0} r^{d-1} f(u(r)) dr = 0,$$

which is contradictory as $f(u(r)) < 0$ on $(0, r_0)$.

2.8 Mixed Super-Linear/Linear Defocusing Case $q = 1$ and $p > 1$

In this section we are focusing on the case $q = 1$ and $p > 1$. The functions g , f , and F are written as follows:

$$g(u) = 1 - |u|^{p-1}, \quad f(u) = ug(u) = |u|^{p-1}u - u, \quad F(u) = \frac{1}{p+1}|u|^{p+1} - \frac{1}{2}u^2.$$

In this case, we have two essential points

$$u_p = \left(\frac{1}{p}\right)^{1/(p-1)} \quad \text{and} \quad \bar{u}_p = \left(\frac{p+1}{2}\right)^{1/(p-1)}$$

with

$$0 < u_p < 1 < \bar{u}_p.$$

Theorem 19 *Whenever $u(0) = a \in (-1, 1)$, the problem (14) has a unique solution u which is oscillating infinitely around ± 1 with limit ± 1 , respectively.*

Proof It is easy to see that u starts as non-decreasing on $(0, \delta)$ for some $\delta > 0$ small enough. Let next $\bar{a} > \bar{u}_p$ be such that $F(\bar{a}) = F(a)$. Because of the energy $E(r)$ we conclude that $a \leq u(r) \leq \bar{a}$ on $(0, \infty)$. Consequently, if u remains non-decreasing on the whole interval $(0, \infty)$, it should be increasing to its unique limit 1. However, if this occurs, the solution u will behave at ∞ as the solution v of the system

$$v'' + \frac{d-1}{r}v' + (p-1)(v-1) = 0, \quad r \in (0, \infty), \quad v(0) = a, \quad v'(0) = 0$$

which is oscillating infinitely on $(0, \infty)$. Next, let $(\zeta_k)_k$ be the sequence of extremum points of u on its domain $(0, \infty)$. Using again the energy $E(r)$ we show that $(u(\zeta_k))_k$ has the limit 1 as k goes to infinity.

Theorem 20 *Whenever $u(0) = a \in (1, \bar{u}_p)$, the problem (14) has a unique solution u which is oscillating around ± 1 .*

Proof It is easy to show that $u(r) \in]\underline{a}, a[$ for all $r > 0$, where \underline{a} is the unique point in $(0, 1)$ such that $F(\underline{a}) = F(a)$. Moreover, the solution u is not strictly decreasing on $(0, +\infty)$. Indeed, assume in the contrary that it is. So, it has a limit l as r goes to $+\infty$. This limit l should be equal to 1. However, in this case, we deduce as previously that the solution u will have the same behavior as the solution v of the problem

$$v'' + \frac{d-1}{r}v' + (p-1)(v-1) = 0, \quad v(0) = a, \quad v'(0) = 0.$$

Observe now that v is oscillating we get a contradiction with the fact that u is non-increasing on $(0, \infty)$. Next, as the solution u is not strictly decreasing on $(0, +\infty)$ it is oscillating. We will study its behavior relatively to its first critical point which will be denoted by r_0 and b its value there. Hence, for the moment, one of the following cases may occur:

- (i.) $b \in [1, a[$.
- (ii.) $b \in]0, 1[$.

For the case (i.), we get $u(r) \in (1, a)$ for $r \in (0, r_0)$. By multiplying the Equation (36) by r^{d-1} and integrating from 0 to r_0 we obtain

$$\int_0^{r_0} r^{d-1} f(u(r)) dr = 0,$$

which is contradictory as this integral is positive. Now, it remains as a consequence the case (ii.), which yields that u is a solution of the problem

$$\begin{cases} u'' + \frac{d-1}{r}u' + f(u) = 0, & r \in (r_0, +\infty), \\ u(r_0) = b & , u'(r_0) = 0, \end{cases}$$

which oscillates around 1.

Theorem 21 *Whenever $u(0) = a \in (\bar{u}_p, +\infty)$, there are three classes of solutions.*

1. *non-increasing with limit 0 at infinity (and thus positive) (this class contains also the compactly supported solutions).*
2. *oscillating around ± 1 with finite number of zeros.*

Proof

1. As previously, the solution u starts by decreasing on some small interval $(0, \varepsilon)$. Assume it remains decreasing on the whole domain $(0, \infty)$. Its limit is immediately 0. Denote $g(r) = \frac{u''(r)}{u(r)}$. It is easy to see that $g(r) \rightarrow \infty$ whenever $r \rightarrow \infty$. Let also $G(r) = \int_R^r g(t)dt$, where $R > 0$ large enough. Denote next, $X(r) = (u(r) \ u'(r))$. We obtain the two-dimensional one-order differential equation $X'(r) = A(r)X(r)$, where $A(r) = \begin{pmatrix} 0 & 1 \\ g(r) & 0 \end{pmatrix}$. Let next $\bar{A}(r) = \int_R^r A(t)dt$. We know from classical techniques of differential equations that $X(t) = \exp(\bar{A}(r))X_0$, where X_0 is a constant vector. Next, standard techniques yield that

$$u(r) = \frac{1}{2}e^{\lambda(r)} \left(C_1 + C_2 \frac{r-R}{\lambda(r)} \right) + \frac{1}{2}e^{-\lambda(r)} \left(C_1 - C_2 \frac{r}{\lambda(r)} \right),$$

where $\lambda(r) = \sqrt{G(r)(r-R)}$. Next, as $u(r) \downarrow 0$ when $r \rightarrow \infty$, we should have

$$C_1 + C_2 \lim_{r \rightarrow \infty} \frac{r-R}{\lambda(r)} = 0.$$

Whenever $C_2 = 0$, we get immediately $C_1 = 0$ and thus u is compactly supported. For $C_2 \neq 0$, we get $G(r) \sim \alpha(r-R)$ at infinity, where $\alpha = \frac{C_1^2}{C_2^2} = \omega^2$, ($\omega > 0$). (In fact we have here also $C_1 \neq 0$). As a result, we get $u(r) \sim Ce^{-\omega r}$ at infinity. Which is the solutions declared (positive and decreasing to 0).

2. We will show now that the solutions which do not remain decreasing on the whole domain $(0, \infty)$ are necessarily oscillating around ± 1 infinitely. This may be shown by following similar techniques as in Section 3.

2.9 Mixed Super-Linear/Super-Linear Defocusing Case $1 < p < q$

In this section we will study the mixed case $1 < p < q$. In this case, we have three essential points that affect the behavior of the solutions of problem (36). These are

$$u_{p,q} = \left(\frac{p}{q}\right)^{1/(q-p)}, \quad \underline{u}_{p,q} = \left(\frac{p-1}{q-1}\right)^{1/(q-p)} \quad \text{and} \quad \bar{u}_{p,q} = \left(\frac{q+1}{p+1}\right)^{1/(q-p)}.$$

It holds that

$$0 < \underline{u}_{p,q} < u_{p,q} < 1 < \bar{u}_{p,q}$$

and

$$g'(\underline{u}_{p,q}) = f'(u_{p,q}) = F(\bar{u}_{p,q}) = 0.$$

The first result in this section is stated as follows.

Theorem 22

- a.** Whenever $u(0) = a \in (0, 1)$, the solution u of problem (36) is non-increasing or oscillating infinitely around 0. Moreover, u has the limit 0 as $r \rightarrow \infty$.
- b.** Whenever $u(0) = a \in (1, \infty)$, the solution u of problem (36) is non-decreasing with limit 0 as $r \rightarrow \infty$. Moreover,

$$u(r) \geq a - \frac{f(a)}{2d}r^2, \quad \forall r \geq 0.$$

- c.** Whenever $u(0) = a \in (-1, 0)$, the solution u of problem (36) is non-decreasing or oscillating infinitely around 0. Moreover, u has the limit 0 as $r \rightarrow \infty$.
- d.** Whenever $u(0) = a \in (-\infty, -1)$, the solution u of problem (36) is non-increasing with limit 0 as $r \rightarrow \infty$. Moreover,

$$u(r) \leq a - \frac{f(a)}{2d}r^2, \quad \forall r \geq 0.$$

Proof Assertions **c.** and **d.** may be obtained using the fact that f is an odd function. So, we will develop only the proofs of assertions **a.** and **b.**

Proof of Assertion a. Using the energy functional $E(r)$, it holds that $u_a(r) \in] - a, a[$ for all $r > 0$. The solution u starts by decreasing near 0. Hence, two cases may occur.

- i. u remains decreasing on its whole domain $(0, \infty)$.
- ii. u is not monotone (decreasing) on its whole domain $(0, \infty)$.

Whenever u satisfies assertion i., it is obvious that it is decreasing to 0 as $r \rightarrow \infty$. Now, assume that assertion ii. holds and let $r_1 > 0$ be the first critical point of u on $(0, \infty)$. We claim that $u(r_1) < 0$. Indeed, if not, we get by integrating Equation (36),

$$0 = \int_0^{r_1} r^{d-1} f(u(r))dr > 0,$$

which is a contradiction. So as the claim. Let next $r_2 > r_1$ be the next critical point. We claim here that $u(r_2) > 0$. If not, we get by the same argument

$$0 = \int_{r_1}^{r_2} r^{d-1} f(u(r))dr < 0,$$

which is a contradiction. Assume now that we have constructed the critical points $r_0 = 0 < r_1 < \dots < r_{2n} < r_{2n+1}$ such that $u(r_{2n}) > 0 > u(r_{2n+1})$. By applying the previous technique we get a solution u oscillating around 0. We now prove that u has the limit 0 in this case also. Indeed, let $z_k, k \in \mathbb{N}$ be the zeros of u in $(0, \infty)$. we get immediately,

$$E(r_{2n+2}) \leq E(z_{2n+1}) < E(r_{2n+1}) < E(z_{2n}) < E(r_{2n}),$$

or equivalently,

$$F(u(r_{2n+2})) \leq \frac{u'(z_{2n+1})^2}{2} \leq F(u(r_{2n+1})) \leq \frac{u'(z_{2n})^2}{2} \leq F(u(r_{2n})),$$

which yields immediately that the only limit of u is 0 as $r \rightarrow \infty$.

Proof of Assertion b. Assume in the contrary that it is not. u starts by increasing near 0. Let $r_0 > 0$ be the first critical point of u . By multiplying Equation (36) by r^{d-1} and integrating from 0 to r_0 we obtain

$$0 = \int_0^{r_0} r^{d-1} f(u(r))dr < 0,$$

which is contradictory. Hence, u is strictly increasing unboundedly. Moreover, it is straightforward that $u \geq a$ on $[0, \infty)$. Consequently, as f is decreasing on $(1, \infty)$, we get $f(u(r)) \leq f(a)$, for all $r \in (0, \infty)$. Consequently

$$(r^{d-1}u'(r))' \geq -r^{d-1}f(a), \quad \forall r \in (0, \infty).$$

By integrating twice on $(0, r)$, we get

$$u(r) \geq a - \frac{f(a)}{2d}r^2, \quad \forall r \geq 0.$$

2.10 Mixed Super-Linear/Super-Linear Defocusing Case $1 < q < p$

In this section we consider the problem

$$\begin{cases} u'' + \frac{d-1}{r}u' + |u|^{p-1}u - |u|^{q-1}u = 0, & r \in (0, \infty) \\ u(0) = a & , u'(0) = 0, \end{cases} \tag{31}$$

where $1 < q < p$. With the original notations g, f , and F , we get

$$\underline{u}_{p,q} < u_{p,q} < 1 < \bar{u}_{p,q},$$

where these points satisfy

$$g'(\underline{u}_{p,q}) = f'(u_{p,q}) = F(\bar{u}_{p,q}) = 0.$$

In other words,

$$\underline{u}_{p,q} = \left(\frac{q-1}{p-1}\right)^{1/(p-q)}, \quad u_{p,q} = \left(\frac{q}{p}\right)^{1/(p-q)}, \quad \bar{u}_{p,q} = \left(\frac{p+1}{q+1}\right)^{1/(p-q)}.$$

Theorem 23

- a. Whenever $u(0) = a \in (0, \bar{u}_{p,q})$, $a \neq 1$, the solution u of problem (36) is oscillating infinitely around 1. Moreover, u has the limit 1 as $r \rightarrow \infty$.
- b. Whenever $u(0) = a \in (\bar{u}_{p,q}, \infty)$, the solution u of problem (36) is oscillating around ± 1 .
- c. Whenever $u(0) = a \in (-\bar{u}_{p,q}, 0)$, $a \neq -1$, the solution u of problem (36) is oscillating infinitely around -1. Moreover, u has the limit -1 as $r \rightarrow \infty$.

Proof As in the previous section(s), assertion **c.** may be obtained using the fact that f is an odd function. So, we will develop only the proofs of assertions **a.** and **b.**

Proof of Assertion a. We will split the proof into two cases. So, let firstly, $a \in (0, 1)$ and $\bar{a} \in (1, \bar{u}_{p,q})$ be such that $F(\bar{a}) = F(a)$. Using the energy functional $E(r)$, it holds that $u_a(r) \in]a, \bar{a}[$ for all $r > 0$. It is easy to see that u starts by increasing near 0. If it continues to be non-decreasing on $(0, \infty)$, it should have exactly the limit 1 as $r \rightarrow \infty$. Hence, it behaves at ∞ as the solution v of the problem

$$v'' + \frac{d-1}{r}v + (p-q)(v-1) = 0, \quad r \in (0, \infty), \quad v(0) = a, \quad v'(0) = 0,$$

which is oscillating. This contradicts the monotony of u . Consequently, u is not monotone on its whole domain $(0, \infty)$. Now, applying similar techniques as in

previous cases we get a solution u which is oscillating around 1 infinitely with its limit being equal to 1.

Assume now that $a \in (1, \bar{u}_{p,q})$ and let $\underline{a} \in (0, 1)$ be such that $F(\underline{a}) = F(a)$. Using as previously the energy functional $E(r)$, it holds that $u_a(r) \in]\underline{a}, a[$ for all $r > 0$. It is also easy to see that u starts by decreasing near 0. If it continues to be decreasing on $(0, \infty)$, it should have as in the previous case the limit 1 as $r \rightarrow \infty$. Hence, as previously, u is oscillating, which is contradictory. Consequently, u is not monotone on its whole domain $(0, \infty)$. By applying similar techniques as in previous sections we get a solution u which is oscillating around 1 infinitely with its limit being equal to 1.

Proof of Assertion b. The solution u starts by decreasing near 0. As previously, because of the energy $E(r)$, it holds that $u(r) \in (-a, a)$ for all $r \geq 0$. Whenever the solution u remains to be non-increasing on $(0, \infty)$ it has the limit 0 as $r \rightarrow \infty$. Denote

$$h(r) = -\frac{d-1}{r}u' - f(u) + u.$$

We get in one hand

$$u'' + u = h(r).$$

On the other hand, it is easy to see that $h(r) \downarrow 0$ as $r \rightarrow +\infty$. So, proceeding as in Section 3, we get a contradiction. As a result, the solution u is not monotone on its whole domain $(0, \infty)$. Proceeding as in previous sections we conclude that u is infinitely oscillating around 1 or -1 with a finite number of zeros.

3 Some Graphical Illustrations

In this section we provide some numerical examples that illustrate the theoretical results presented previously. We will see that as it is stated in the theory the oscillating behavior, nodes as well as the asymptotic behavior of the solution depend obviously on the interval of the initial value and the power laws in the nonlinear part as well as their order.

3.1 Example 1: The Mixed Sub-linear Case $0 < p < q < 1$

We consider in this example the case where $p = \frac{1}{8}$, $q = \frac{1}{4}$, and $u(0) = a = 0.1$. Remark that $0 < q < p < 1$ and $u(0) = a \in (0, 1)$. We obtain consequently the following sub-linear case:

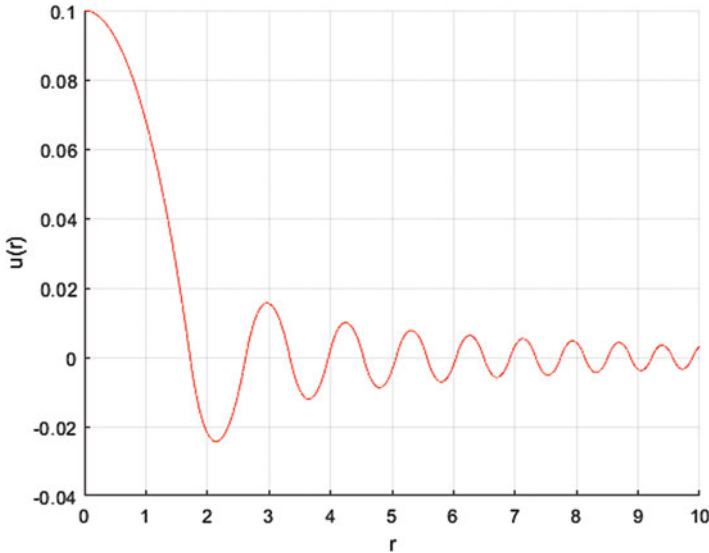


Fig. 2 The solution u for $p = \frac{1}{8}$, $q = \frac{1}{4}$, and $u(0) = 0.1$

$$\begin{cases} u'' + \frac{d-1}{r}u' + |u|^{-7/8}u - |u|^{-3/4}u = 0, & r \in (0, \infty) \\ u'(0) = 0, & u(0) = a. \end{cases} \tag{32}$$

Figure 2 shows the coherence with the theoretical result of Theorem 1 for $a = 0.1$. The solution u of problem (32) is oscillating infinitely around 0 with its limit being 0. The same result is confirmed by Figure 3 where we fixed the parameters of the problem in the same intervals as previously. We precisely fixed $p = \frac{1}{8}$, $q = \frac{1}{4}$, and $u(0) = a = 0.35$. Remark that here also $0 < p < q < 1$ and $u(0) \in (0, 1)$. Next we consider already for the same sub-linear powers as above the initial conditions out of the interval $(-1, 1)$. For $u(0) = a = 1.1$ we get a coherent result with Theorem 2 as shown in Figure 4. The solution is clearly increasing to infinity.

3.2 Example 2: The Mixed Linear/Sub-linear Case $p = 1$ and $0 < q < 1$

We consider in this example the case where $p = 1$ and $q = \frac{1}{4}$. We will show that according to the values of these power laws and the initial value $u(0) = a$ the numerical illustrations are completely coherent with the theoretical results of Theorems 3, 4, and 5. We obtain consequently the following problem:

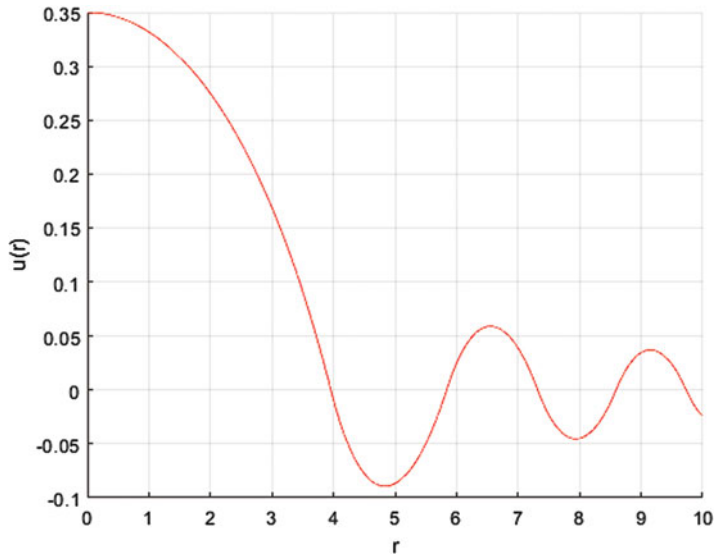


Fig. 3 The solution u for $p = \frac{1}{8}$, $q = \frac{1}{4}$, and $u(0) = 0.35$

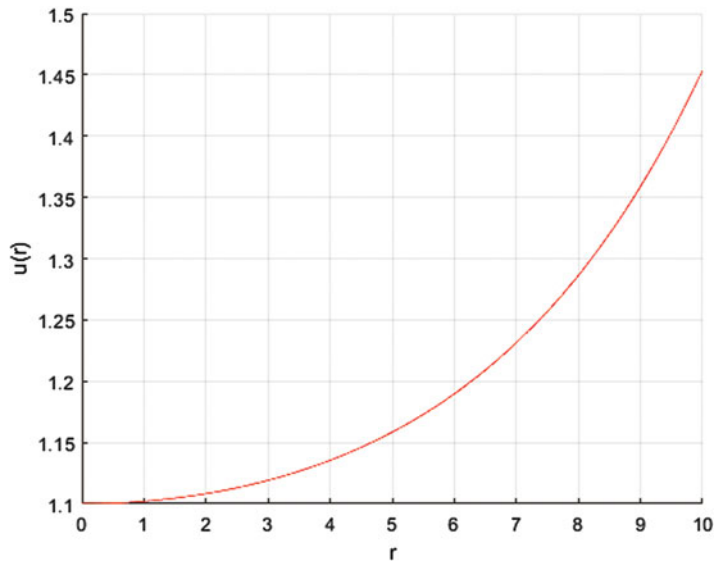


Fig. 4 The solution u for $p = \frac{1}{8}$, $q = \frac{1}{4}$, and $u(0) = 1.1$

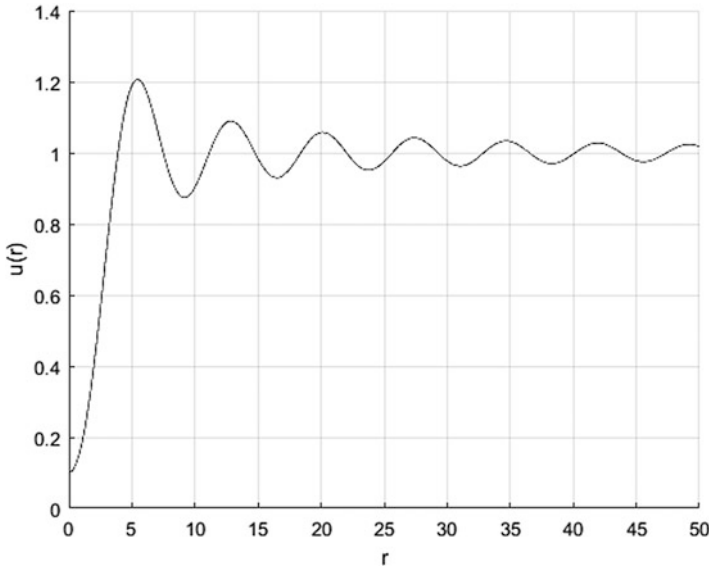


Fig. 5 The solution u for $p = 1, q = \frac{1}{4}$, and $u(0) = 0.1$

$$\begin{cases} u'' + \frac{d-1}{r}u' + u - |u|^{-3/4}u = 0, & r \in (0, \infty) \\ u'(0) = 0 & , u(0) = a. \end{cases} \tag{33}$$

In the first case we fix $u(0) = a = 0.1 \in (0, 1)$. Figure 5 shows the coherence with the theoretical result of Theorem 3. The solution u of problem (33) is oscillating infinitely around 1 with no zeros and thus it constitutes a positive solution. Next we consider already for the same linear/sub-linear problem with the same powers as above and the initial conditions out of the interval $(-1, 1)$ but in the interval $(1, \bar{u}_q)$. Recall that in the present case $\bar{u}_q = (\sqrt[3]{\frac{8}{5}})^4 = 1.8714$. In the first case we fix $u(0) = a = 1.8$. The result is illustrated by Figure 6 which shows a coherence with Theorem 4. The third case concerns the same problem (33) with an initial value $u(0) = a = 2 \in (\bar{u}_q, \infty)$. The solution is illustrated by Figure 7 which shows a coherence with Theorem 5. Next we fix $u(0) = a = 5.25 \in (\bar{u}_q, \infty)$. The solution is illustrated by Figure 8 which shows a coherence with Theorem 5. Finally, we fix $u(0) = a = 6 \in (\bar{u}_q, \infty)$. The solution is illustrated by Figure 9 which shows a coherence with Theorem 5.

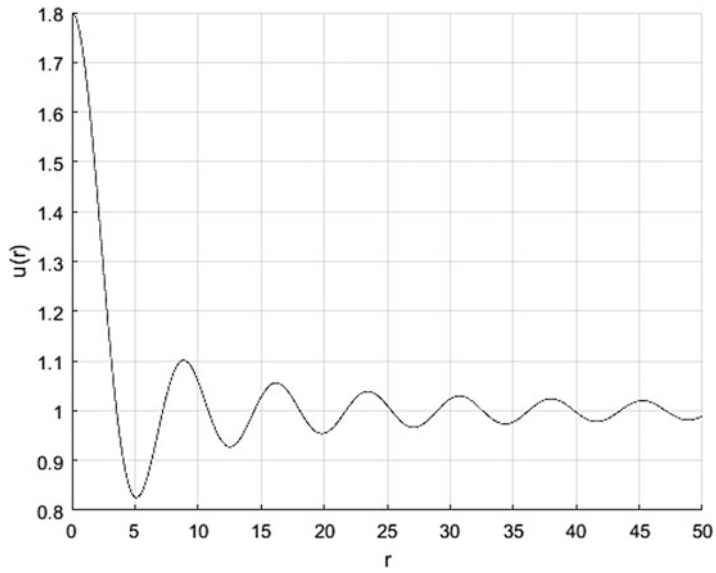


Fig. 6 The solution u for $p = 1$, $q = \frac{1}{4}$, and $u(0) = 1.8$

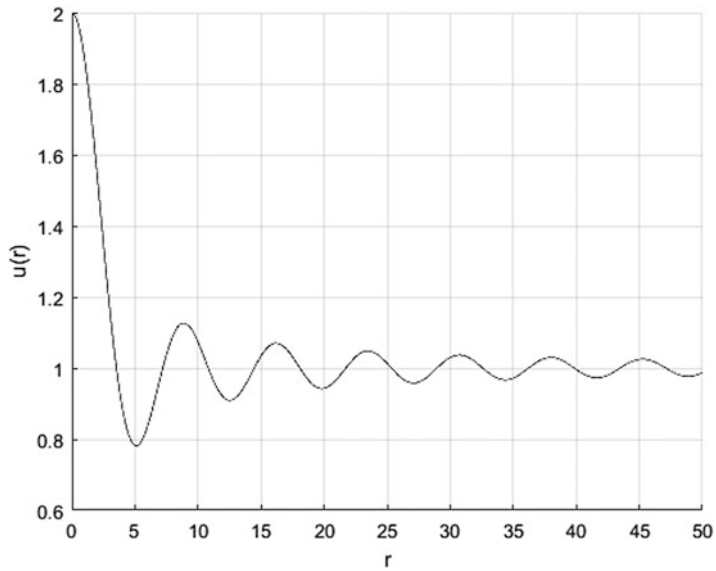


Fig. 7 The solution u for $p = 1$, $q = \frac{1}{4}$, and $u(0) = 2$

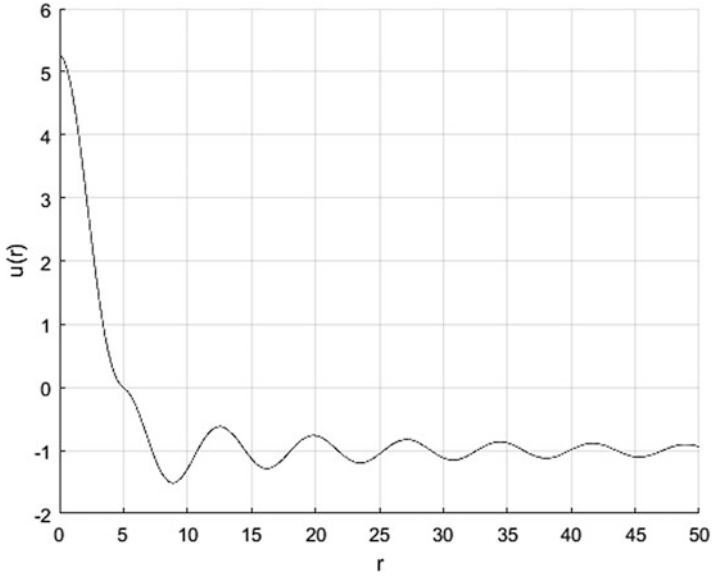


Fig. 8 The solution u for $p = 1, q = \frac{1}{4}$, and $u(0) = 5.25$

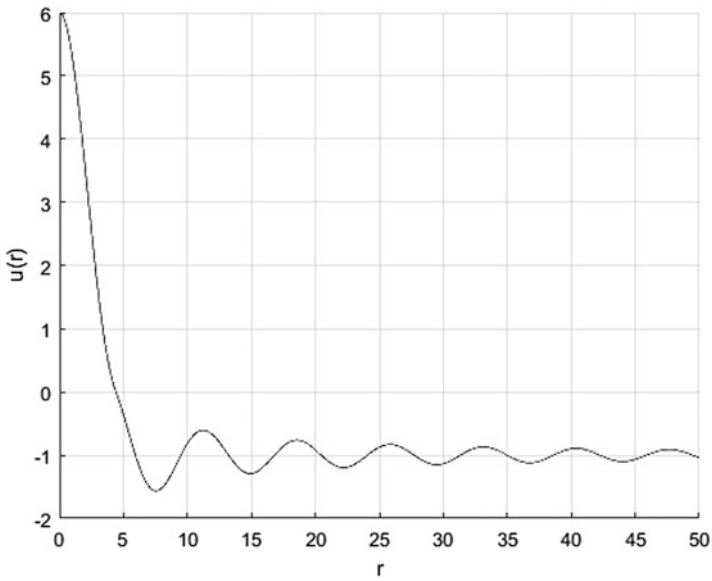


Fig. 9 The solution u for $p = 1, q = \frac{1}{4}$, and $u(0) = 6$

3.3 Example 3: The Mixed Sub-linear/Sub-linear Case $0 < q < p < 1$

We consider in this example the case where $p = \frac{1}{2}$ and $q = \frac{1}{4}$. We will show that in this case also according to the values of these power laws and the initial value $u(0) = a$ the numerical illustrations are completely coherent with the theoretical results of Theorems 9, 10, and 11. We obtain consequently the following case:

$$\begin{cases} u'' + \frac{d-1}{r}u' + |u|^{-1/2}u - |u|^{-3/4}u = 0, & r \in (0, \infty) \\ u'(0) = 0, & u(0) = a \end{cases} \quad (34)$$

In the first case we fix $u(0) = a = 0.75 \in (0, 1)$. Figure 10 shows the coherence with the theoretical result of Theorem 9. The solution u of problem (34) is oscillating infinitely around 1 with no zeros. It constitutes consequently a positive solution. Next we consider already for the same problem with the same powers as above and the initial conditions out of the interval $(-1, 1)$ but in the interval $(1, \underline{u}_{p,q})$. Recall that in the present case $\underline{u}_{p,q} = (\frac{6}{5})^4 = 2.0736$. In the first case we fix $u(0) = a = 1.5$. The result is illustrated by Figure 11 which shows a coherence with Theorem 10. The third case concerns the same problem (34) with an initial value $u(0) = a = 6.5 \in (\underline{u}_{p,q}, \infty)$. The solution is illustrated by Figure 12 which shows a coherence with Theorem 11. Next we fix $u(0) = a = 6.75 \in (\underline{u}_{p,q}, \infty)$. The solution is illustrated by Figure 13 which shows a coherence with Theorem 11.

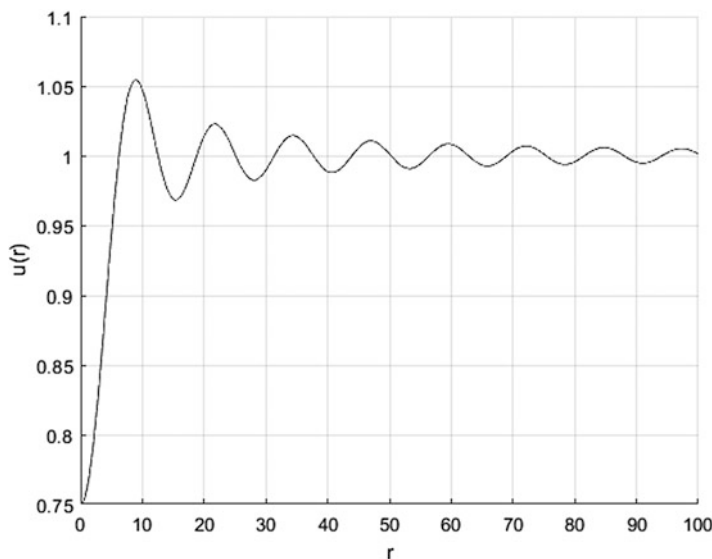


Fig. 10 The solution u for $p = \frac{1}{2}$, $q = \frac{1}{4}$, and $u(0) = 0.75$

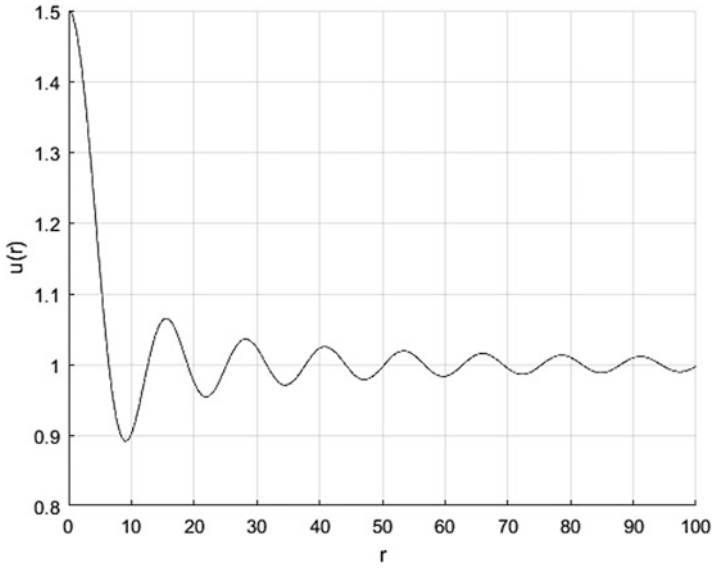


Fig. 11 The solution u for $p = \frac{1}{2}, q = \frac{1}{4}$, and $u(0) = 1.5$

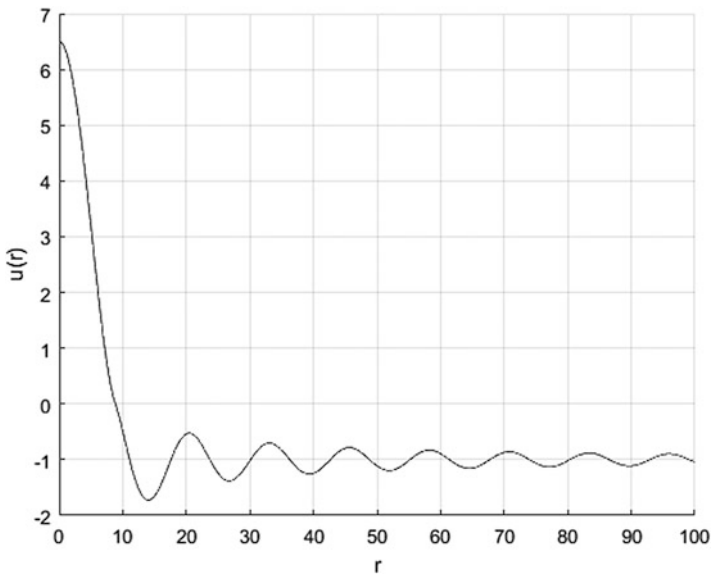


Fig. 12 The solution u for $p = \frac{1}{2}, q = \frac{1}{4}$, and $u(0) = 6.5$

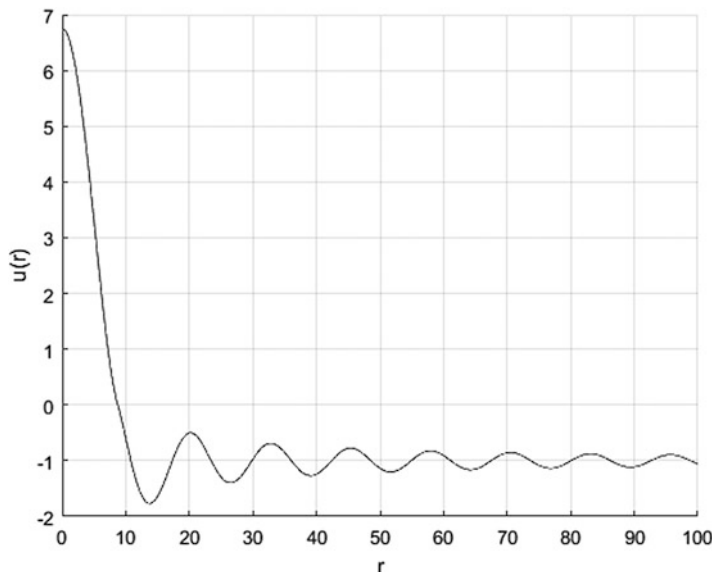


Fig. 13 The solution u for $p = \frac{1}{2}$, $q = \frac{1}{4}$, and $u(0) = 6.75$

3.4 Example 4: The Mixed Sub-linear/Linear Case $0 < p < 1$ and $q = 1$

We consider in this example the case where $p = \frac{1}{4}$ and $q = 1$. We will show that according to the values of these power laws and the initial value $u(0) = a$ the numerical illustrations are completely coherent with the theoretical results of Theorems 7 and 8. The following nonlinear problem is studied:

$$\begin{cases} u'' + \frac{d-1}{r}u' + |u|^{-3/4}u - u = 0, & r \in (0, \infty) \\ u'(0) = 0 & , u(0) = a. \end{cases} \quad (35)$$

In the first case we fix $u(0) = a = 0.35 \in (0, 1)$. Figure 14 shows the coherence with the theoretical result of Theorem 7. The solution u of problem (35) is oscillating infinitely around 1 with no zeros and thus it constitutes a positive solution.

Next we consider already for the same linear/sub-linear problem with the same powers as above and the initial conditions out of the interval $(1, \infty)$. In the first case we fix $u(0) = a = 1.01$. The result is illustrated by Figure 15 which shows a coherence with Theorem 8. The third case concerns the same problem (35) with an initial value $u(0) = a = 1.25 \in (1, \infty)$. The solution is illustrated by Figure 16 which shows a coherence with Theorem 8.

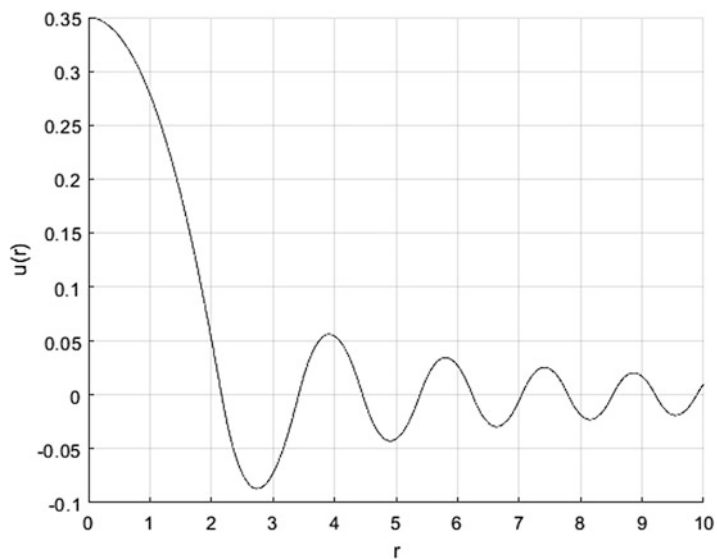


Fig. 14 The solution u for $p = \frac{1}{4}$, $q = 1$, and $u(0) = 0.35$

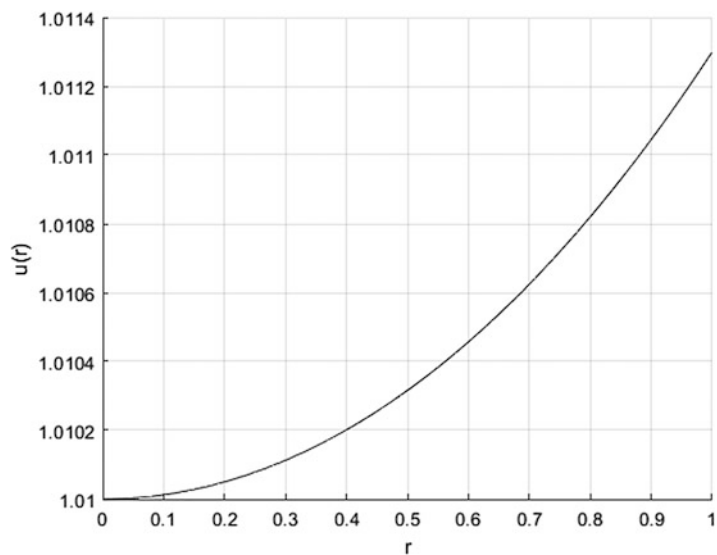


Fig. 15 The solution u for $p = \frac{1}{4}$, $q = 1$, and $u(0) = 1.01$

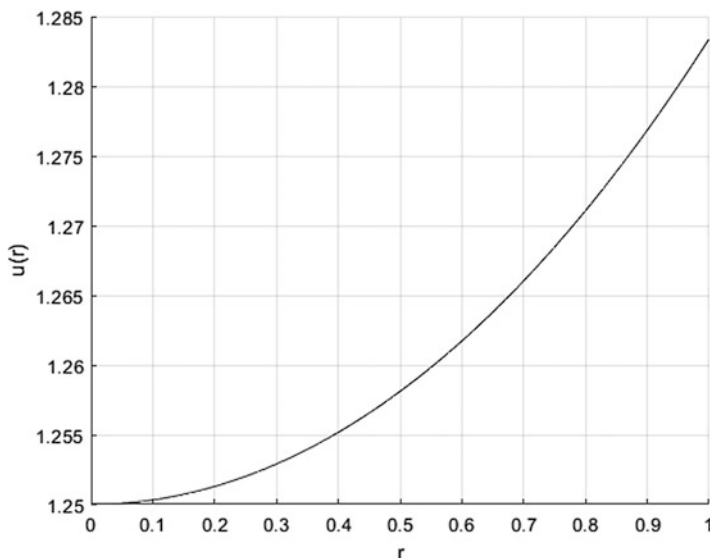


Fig. 16 The solution u for $p = \frac{1}{4}, q = 1$, and $u(0) = 1.2$

4 Conclusion

The present chapter investigates the problem of existence, uniqueness, and classification of the solutions of an elliptic problem derived from the famous Schrödinger equation in a mixed nonlinear case. We precisely considered an elliptic problem of the form

$$\begin{cases} u'' + \frac{d-1}{r}u' + |u|^{p-1}u - |u|^{q-1}u = 0, & r \in (0, \infty) \\ u'(0) = 0 & , u(0) = a, \end{cases} \tag{36}$$

where a is a real number parameter and real p, q are power laws that may be sub-linear, super-linear sometimes leading to convex and concave cases. A full study relatively to these power laws and the initial value $u(0) = a$ has been developed.

As we have mentioned in the introduction, the chapter in its whole aim is a review of existing results about the studied problems more than a development of new ones reminiscent of some few cases that are not previously developed. We aim thus it will constitute a good reference especially for beginners in the field of nonlinear analysis of PDEs.

References

1. A. Adimurthi, S.L. Yadava, Elementary proof of the nonexistence of nodal solutions for the semilinear elliptic equations with critical Sobolev exponent. *Nonlinear Anal.* **14**(9), 785–787 (1990)
2. A. Ambrosetti, Critical points and nonlinear variational problems. *Priprints di Mathematica*, 118, Scuola Normale Superiore, Pisa (1991)
3. A. Ambrosetti, M. Struwe, A note on the problem $-\Delta u = \lambda u + u|u|^{2^*-2}$. *Manusc. Math.* **54**, 373–379 (1986)
4. A. Ambrosetti, H. Brezis, G. Cerami, Combined effects of concave and convex nonlinearities in some elliptic problems. *J. Funct. Anal.* **122**, 519–543 (1994)
5. F.V. Atkinson, H. Brezis, L.A. Peletier, Solutions d'équations elliptiques avec exposant de Sobolev critique qui changent de signe. (French) [Nodal solutions of elliptic equations with critical Sobolev exponents] *C. R. Acad. Sci. Paris I Math.* **306**(16), 711–714 (1988)
6. J. Avron, I. Herbst, B. Simon, Schrödinger operators with electromagnetic fields. III. Atoms in homogeneous magnetic field. *Commun. Math. Phys.* **79**, 529–572 (1981)
7. B. Balabane, J. Dolbeault, H. Ounaies, Nodal solutions for a sublinear elliptic equation. *Nonlinear Anal.* **52**, 219–237 (2003)
8. R. Bamon, I. Flores, M. del Pino, Positive solutions of elliptic equations in \mathbb{R}^N with a super-subcritical nonlinearity. *C. R. Acad. Sci. Paris* **330**, 187–191 (2000)
9. A. Ben Mabrouk, M. Ayadi, A linearized finite-difference method for the solution of some mixed concave and convex nonlinear problems. *Appl. Math. Comput.* **197**, 1–10 (2008)
10. A. Ben Mabrouk, M. Ayadi, Lyapunov type operators for numerical solutions of PDEs. *Appl. Math. Comput.* **204**, 395–407 (2008)
11. A. Ben Mabrouk, M.L. Ben Mohamed, On some critical and slightly super-critical sub-superlinear equations. *Far East J. Appl. Math. (Special Volume of PDEs)* **23**(1), 73–90 (2006)
12. A. Ben Mabrouk, M.L. Ben Mohamed, Nodal solutions for some nonlinear elliptic equations. *Appl. Math. Comput.* **186**, 589–597 (2007)
13. A. Ben Mabrouk, M.L. Ben Mohamed, Phase plane analysis and classification of solutions of a mixed sublinear-superlinear elliptic problem. *Nonlinear Anal.* **70**, 1–15 (2009)
14. A. Ben Mabrouk, M.L. Ben Mohamed, Nonradial solutions of a mixed concave-convex elliptic problem. *J. Part. Diff. Equ.* **24**(4), 313–323 (2011)
15. A. Ben Mabrouk, M.L. Ben Mohamed, K. Omrani, Finite difference approximate solutions for a mixed sub-superlinear equation. *J. Appl. Math. Comput.* **187**, 1007–1016 (2007)
16. A.G. Bratsos, A linearized finite-difference method for the solution of the nonlinear cubic Schrödinger equation. *Commun. Appl. Anal.* **4**(1), 133–139 (2000)
17. A.G. Bratsos, A linearized finite-difference scheme for the numerical solution of the nonlinear cubic Schrödinger equation. *Korean J. Comput. Appl. Math.* **8**(3), 459–467 (2001)
18. A.G. Bratsos, C. Tsituras, D.G. Natsis, Linearized numerical schemes for the Boussinesq equation. *Appl. Num. Anal. Comp. Math.* **2**(1), 34–53 (2005)
19. H. Brezis, L. Nirenberg, Characterization of the range of some non-linear operators and application to boundary value problems. *Ann. Scuola Norm. Super. Pisa* **4**(5), 225–326 (1978)
20. H. Brezis, L. Nirenberg, Positive solutions of nonlinear elliptic equations involving critical Sobolev exponents. *Commun. Pure Appl. Math.* **36**(4), 437–477 (1983)
21. H. Brezis, L. Veron, Removable singularities for some non linear elliptic equations. *Arch. Rat. Mech. Anal.* **75**, 1–6 (1980)
22. C.J. Budd, A.R. Humphries, Weak finite-dimensional approximations of semi-linear elliptic PDEs with near-critical exponents. *Asymptot. Anal.* **17**(3), 185–220 (1998)
23. G. Cerami, D. Fortunato, M. Struwe, Bifurcation and multiplicity results for nonlinear elliptic problems involving critical Sobolev exponents. *Ann. Inst. H. Poincaré Anal. Non Linéaire.* **1**(5), 341–350 (1984)
24. G. Cerami, S. Solimini, M. Struwe, Some existence results for superlinear elliptic boundary value problems involving critical exponents. *J. Funct. Anal.* **69**(3), 289–306 (1986)

25. R. Chteoui, A. Ben Mabrouk, H. Ounaies, Existence and properties of radial solutions of a sub-linear elliptic equation. *J. Part. Differ. Equ.* **28**(1), 30–38 (2015)
26. R. Chteoui and A. Ben Mabrouk, A generalized Lyapunov-Sylvester computational method for numerical solutions of NLS equation with singular potential. *Anal. Theory Appl.* **33**, 333–354 (2017).
27. C. Cortazar, M. Elgueta, P. Felmer, On a semilinear elliptic problem in \mathbb{R}^N with a non-Lipschitzian nonlinearity. *Adv. Differ. Equ.* **1**(2), 199–218 (1996)
28. M. Dehghan, Finite difference procedures for solving a problem arising in modeling and design of certain optoelectronic devices. *Math. Comput. Simul.* **71**, 16–30 (2006)
29. A. Floer, A. Weinstein, Nonspreading wave packets for the cubic Schrödinger equation with a bounded potential. *J. Funct. Anal.* **69**, 397–408 (1986)
30. A. Hasegawa, Y. Kodama, *Solitons in Optical Communications* (Academic Press, San Diego, 1995)
31. C. Lanski, Quadratic central differential identities of prime rings. *Nova J. Alg. Geom.* **1**(2), 185–206 (1992)
32. B.A. Malomed, Variational methods in nonlinear fiber optics and related fields. *Progress Opt.* **43**, 69–191 (2002)
33. Y.G. Oh, Existence of semi-classical bound states of nonlinear Schrödinger equations with potentials of the class (V_a) . *Commun. Partial Differ. Equ.* **13**, 1499–1519 (1988)
34. M. Onorato, A.R. Osborne, M. Serio, S. Bertone, Freak waves in random oceanic sea states. *Phys. Rev. Lett.* **86**, 5831–5834 (2001)
35. P.H. Rabinowitz, Multiple critical points of perturbed symmetric functionals. *Trans. Am. Math. Soc.* **272**(2), 753–769 (1982)
36. J. Serrin, M. Tang, Uniqueness of ground states for quasilinear elliptic equations. *Ind. Univ. Math. J.* **49**(3), 897–923 (2000)
37. C. Sulem, P.-L. Sulem, The nonlinear Schrödinger equation, in *Self-focusing and Wave Collapse, Applied Mathematical Sciences* (Springer, New York, 1999), p. 139
38. V. E. Zakharov, *Collapse and Self-focusing of Langmuir Waves*, eds. by A.A. Galeev, R.N. Sudan. *Handbook of Plasma Physics, Basic Plasma Physics*, vol 2 (Elsevier, North-Holland, 1984), pp. 81–121

An Optimization Model for a Network of Organ Transplants with Uncertain Availability



Gabriella Colajanni and Patrizia Daniele

Abstract Thanks to advances in modern medicine and the presence of an increasingly efficient organizational network, nowadays transplantation can save thousands of lives every year. In our paper we present a supply chain model with transplant centers and donor hospitals, where we assume that the medical teams move to the hospitals, take the organs, and go back to the transplant centers, using the most suitable transport mode. Since the availability of organs in each donor hospital is unknown a priori, we introduce a random variable which gives us an expected value of such an availability. The aim of the model is to obtain a social optimum in which we intend to minimize the total costs, given by transport costs of both teams and organs, as well as those of transplant patients, the costs of removal, of transplantation and of post-transplantation, the costs of disposal of diseased or non-functioning organs and of the damaged ones, and the penalties. We deduce the associated variational inequality formulation and an existing result for the solution. Finally, we present some numerical examples.

1 Introduction

Transplantation is a rapidly evolving sector and represents a true frontier of modern surgery. In recent years the progress of scientific research has made possible interventions considered unachievable only a few years ago.

Transplantation is a surgical procedure that involves the replacement of a damaged or missing organ or tissue with another taken from the same individual (homotransplant or autograft), from another individual (allograft) or from an individual of a different species (xenotransplantation). It is often used as a synonym for grafting (although in this case the transfer of organs or tissues is carried out without a surgical anastomosis). In this case we talk about “removal” of an organ

G. Colajanni · P. Daniele (✉)

Department of Mathematics and Computer Science, University of Catania, Catania, Italy
e-mail: daniele@dmi.unict.it

© Springer Nature Switzerland AG 2021

T. M. Rassias, P. M. Pardalos (eds.), *Nonlinear Analysis and Global Optimization*,
Springer Optimization and Its Applications 167,
https://doi.org/10.1007/978-3-030-61732-5_6

163

or tissue from a donor organism, whereas the term “explant” must be reserved for the surgical removal of an organ previously transplanted and removed for various reasons.

Transplantation is performed in authorized facilities based on certain minimum structural, technological, and organizational requirements. Thanks to advances in modern medicine and the presence of an increasingly efficient organizational network, transplantation today is a routine intervention that can save thousands of lives every year.

Transplantation is used when severe organ failures or severe blood diseases cannot be treated with other medical treatments; indeed, transplantation is often a life-saving therapy, as in the case in which the severe insufficiency concerns the heart, the liver, the lungs, the intestine. For the kidney and pancreas, transplantation is the natural replacement therapy, much more effective and tolerable than dialysis or insulin administration. In other cases, we talk about an “improvement” intervention, such as for tissue transplantation.

At the base of the transplant there is the donation, a voluntary, conscious, free, and anonymous act.

Therefore, two phases are identified: the organ removal from a subject called *donor* (which can be a living person or died through brain or circulatory death), and the subsequent transplantation or grafting of the same onto a subject called *recipient*, usually with the removal of the sick native counterpart. It is possible to transplant organs (kidney, liver, heart, lung, intestine), tissues (corneas, bone, cartilage, heart valve, blood vessels, skin), or complex ensembles (hand).

There are different types of transplants. In this paper we take into account only the orthotopic transplant (the original malfunctioning organ is removed, and the donor organ is placed in the same anatomical position as the original organ), which is more widespread, and we ignore the heterotopic transplant (a new organ is placed side by side with the old one which is no longer functioning, but remains in its place; this type of transplant is also called auxiliary).

In almost all countries there is a shortage of organs for transplantation. Countries often have formal systems to manage the order determination process for recipients of available organs and organ donors.

The main criterion for donor/recipient matching is obviously that of compatibility, based on tissue typing and valued by the scientific leadership in the field; however, the choice is also influenced by other parameters, such as the age and general state of health of the recipient (although, in recent years, the improvement of transplantation techniques has allowed this operation to be performed even in patients of advanced age).

In most countries, there are various lengths of waiting times that can affect who receives the organ, due to the different availabilities of organs, medical factors, and the position on the waiting list or, in some countries, the existence of the targeted donation.

It should be noted that the number of organs needed for transplants is almost always insufficient to cover waiting lists quickly, so the mortality rate among the listed patients can be high.

In Italy, the *National Transplant Center* (CNT) is the technical-scientific organization responsible for coordinating the National Transplant Network, which is used by the Ministry of Health, the Regions and the Autonomous Provinces.

The CNT carries out functions of guidance, coordination, regulation, training, and supervision of the transplantation network, as well as operational functions for the allocation of organs and for national transplant programs (the urgency program, the pediatric program, the hyperimmune program, the split-liver program, the cross-over program for kidney, organ exchange with foreign countries, returns and surpluses).

The *National Transplant Network* is one of the clinical networks of the Italian National Health System, that is, an organizational model aimed at taking care of patients with formalized and coordinated methods among all the professionals and structures operating in the area.

As already mentioned, at the national level we find the CNT, while at the regional level there are the Regional or Interregional Centers for Transplantation (CRT), which are public structures that, among the different tasks, coordinate, on a regional level, the activities of procurement, donation, and transplant and proceed with the allocation of organs. In addition, at the local level there are *donor hospitals* (public health facilities where organs, tissues, and hematopoietic stem cells are taken for transplantation purposes) and *transplant centers* (facilities where there is a team authorized to perform interventions for transplantation of organs, tissues, and hematopoietic stem cells).

The *Transplant Information System* (SIT) is an IT infrastructure for managing data connected to the activity of the National Transplant Network. The SIT was established by Law on April 1, 1999 n. 91 under the New Health Information System; through the SIT it is possible to guarantee the transparency and traceability of donation, collection, and transplantation processes.

In 2018, 1,924,017 declarations of intention were registered in the SIT; 44,908 are the new donors registered in the Italian Bone Marrow Donor Registry (IBMDR); 1689 are the organ donors (deceased and living) with 318 living donors (there was an increase of 94%, compared to 20 years ago); 13,482 tissue donors; 268 hematopoietic stem cell donors (registered in the IBMDR registry and cord blood donations); 3725 organ transplants (deceased and living donors), with 318 living organ transplants, with an increase of 54% since 1998; 16,468 tissue transplants; 848 hematopoietic stem cell transplants (from unfamiliar donors).

In Italy, as reported in Table 1, the number of transplants is quite high and the average waiting time has decreased for different types of organs.

Numerous papers confirm the importance of the transplantation topic, which is highlighted by a lot of research studies in both medical and mathematical literature. In this latter area some themes are treated, such as: optimization of times, fundamental in the transplant process (see, for example, [1, 2]); the best allocation of organs to transplant centers [4, 9, 10]; the management of waiting lists [3, 7]; cost minimization of each stage of the transplant process [5].

One of the most important aspects throughout all the donation-transplant process is the logistics which is used to manage and to coordinate all phases which are

Table 1 Data on the number of transplants and the average waiting time in 2002 and 2018

Organ	Number of operations in 2018	Waiting time in 2002 (in months)	Waiting time in 2018 (in months)
Lung	143	14	12
Heart	233	81	13
Liver	1159	7	5
Kidney	1831	32	24

Table 2 Organs and cold ischemia times

Organ	Cold ischemia time
Heart	4–6 h
Lung	4–6 h
Liver	12–18 h
Kidney	48–72 h
Pancreas	12–24 h

necessary to reach the goal as quickly as possible. Logistics concerns with the medical teams, the organs, the biological materials, and the patients transport.

The organs, once taken, require special procedures for their preservation for a transplant. The maximum extracorporeal storage time varies from organ to organ and is based on the storage liquid and the temperature. They can be transported and stored based on the relative cold ischemia times which are the maximum storage times after removal and before transplantation. In Table 2 we indicate the main cold ischemia times. Therefore, logistics is a very important aspect in transplantation because an improper management can lead to delays and problems during the whole process.

In this paper, we aim at presenting a mathematical model, based on networks, which allows us to minimize the total costs associated with organ transplants, ensuring that ischemia times are respected and entirely using the quantities of organs available at the donor hospitals, which are random variables.

The paper is structured as follows. In the next section we introduce the network underlying the organs transplant system, consisting of transplant centers and donor hospitals which can own different types of available organs. We then present the cost functions associated with: transportation, organ removals, waste disposals, and post-transplants. We also deal with the uncertainty of organ availability and we present the costs associated with penalties. Therefore, we show the mathematical model. In Section 3 we determine the optimality conditions for the national health service and derive the variational inequality formulation introducing the Lagrange multipliers associated with the constraints and we also report existence results for the solution. In Section 4 we apply the model to some numerical examples.

2 The Mathematical Model

In this section we first illustrate the network related to the transportation process, and then, we describe the mathematical model.

The network structure underlying the organs transplant system, depicted in Figure 1, consists of three tiers of nodes: the first level is represented by the transplant centers from which the medical teams reach the donor hospitals, which constitute the second tier, where they perform the organ removal and then the medical teams and the organs go back to the transplant centers, the third level, where the organ transplant is performed.

Therefore, the nodes at the highest and at the lowest levels of the network represent the m transplant centers, with a typical one denoted by i ; while the intermediate level consists of n donor hospitals, with a typical one denoted by j .

We underline that each donor hospital can have S different kinds of available organs, such as kidney, liver, pancreas, intestine, heart, as well as lungs, and so on (and we denote by s the typical one).

In the organs transport network we also consider v different transportation services, such as ambulance, airplane, helicopter, etc., and we denote the typical one by k . We can remark that in the network the different transportation services are denoted by parallel edges, since they are not decision makers.

Let g_{ijk} be the quantity of medical teams moving from the transplant center i to the donor hospital j using the k -th transportation service and we group such flows into the vector $G^1 \in \mathbb{R}_+^{mnv}$. Let c_{ijk}^{TE} be the transportation costs associated with the medical teams from the transplant center i to the donor hospital j using the k -th transportation service and we assume c_{ijk}^{TE} as a function of g_{ijk} :

$$c_{ijk}^{TE} = c_{ijk}^{TE}(g_{ijk}), \quad \forall i = 1, \dots, m, \forall j = 1, \dots, n, \forall k = 1, \dots, v.$$

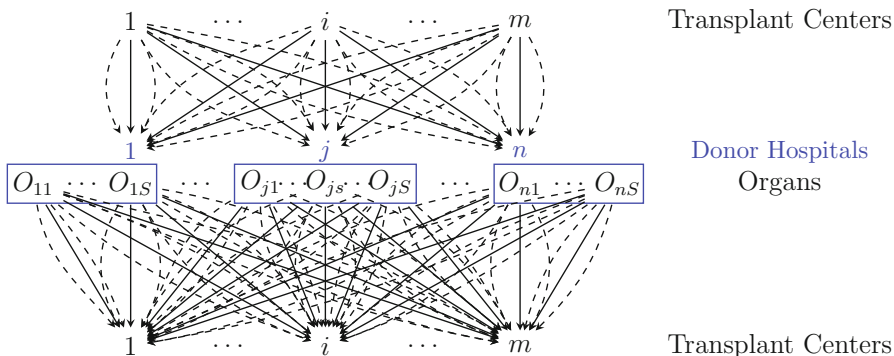


Fig. 1 Organs transport network

Let g_{ij} be the amount of organs that the medical teams in the transplant center i intend to take from the hospital j , and we group such quantities into the vector $G^2 \in \mathbb{R}_+^{mn}$.

As a consequence, the number of medical teams cannot exceed the amount of organs to be taken, namely:

$$\sum_{k=1}^v g_{ijk} \leq g_{ij} \quad \forall i = 1, \dots, m, \quad \forall j = 1, \dots, n. \quad (1)$$

Further, we assume that the quantity of organs that all the teams in i intend to take from j is less than or equal to a large enough upper bound:

$$g_{ij} \leq \bar{M} \quad \forall i = 1, \dots, m, \quad \forall j = 1, \dots, n. \quad (2)$$

Let us assume that we have different kinds of organs (kidney, liver, pancreas, intestine, heart, as well as lungs, ...) and we denote by s the typical one, where $s = 1, \dots, S$. Let O_{js} be the class of organs of type s available in the donor hospital j . Then, let \tilde{g}_{jsik} be the quantity of organs of the class O_{js} sent from the donor hospital j to the transplant center i using the k -th transportation service and we group such flows into the vector $G^3 \in \mathbb{R}_+^{nSmv}$. Let c_{jik}^{TO} be the transportation costs associated with the organs from the donor hospital j to the transplant center i using the k -th transportation service and we assume c_{jik}^{TO} as a function of the sum of all organs sent from j to i via k :

$$c_{jik}^{TO} = c_{jik}^{TO} \left(\sum_{s=1}^S \tilde{g}_{jsik} \right), \quad j = 1, \dots, n, \quad i = 1, \dots, m, \quad k = 1, \dots, v.$$

We also assume that the number of medical teams moving from i cannot exceed the number of organs which are transported to i , that is:

$$\sum_{j=1}^n \sum_{k=1}^v g_{ijk} \leq \sum_{j=1}^n \sum_{s=1}^S \sum_{k=1}^v \tilde{g}_{jsik} \quad \forall i = 1, \dots, m. \quad (3)$$

Let \tilde{g}_i be the quantity of organs transplanted at the center i and we group such quantities into the vector $G^4 \in \mathbb{R}_+^m$. Let \tilde{c}_i^S be the health costs due to the transplant at the center i and we assume such costs as a function of \tilde{g}_i :

$$\tilde{c}_i^S = \tilde{c}_i^S(\tilde{g}_i), \quad \forall i = 1, \dots, m.$$

Let c_i^{POST} be the costs incurred in the center i during the post-transplant process and we assume they are a function of \tilde{g}_i :

$$c_i^{POST} = c_i^{POST}(\tilde{g}_i), \quad \forall i = 1, \dots, m.$$

Let c_j^W be the unit special waste disposal cost at the donor hospital j (for instance, explanted organs which are unfit for transplant). Let $\beta_j \in [0, 1]$ be the portion of explanted organs discarded in the donor hospital j . Further, let \tilde{c}_i^W be the unit special waste disposal cost at the transplant center i (for instance, diseased organs to be replaced).

Let $\gamma_{jik} \in [0, 1]$ be the portion of organs reaching the transplant center i , but which cannot be transplanted, due to damage in the transportation, for instance. As a consequence, in every transplant center i the quantities of organs which must be wasted is given by the sum of the quantity of harvested unhealthy organs (that is equal to \tilde{g}_i) and the portion of those which are damaged during transport: $\tilde{g}_i + \sum_{j=1}^n \sum_{s=1}^S \sum_{k=1}^v \gamma_{jik} \tilde{g}_{jsik}$. Hence, the relationship among \tilde{g}_i , \tilde{g}_{jsik} and γ_{jik} is given by

$$\tilde{g}_i = \sum_{j=1}^n \sum_{k=1}^v [(1 - \gamma_{jik}) (\sum_{s=1}^S \tilde{g}_{jsik})], \tag{4}$$

namely the number of transplanted organs is the same as the number of transported organs minus the wasted ones.

Let \tilde{g}_{js} denote the quantity of organs of the class O_{js} available at the donor hospital j , which is a random variable with probability density function given by $f_{js}(t)$. Let P_{js} be the probability distribution function of \tilde{g}_{js} , that is:

$$P_{js}(G_{js}) = P_{js}(\tilde{g}_{js} \leq G_{js}) = \int_0^{G_{js}} f_{js}(t) d(t).$$

Therefore, the expected values of \tilde{g}_{js} , $\forall j = 1, \dots, n$, $\forall s = 1, \dots, S$, are given by

$$g_{js} = \mathbb{E}[\tilde{g}_{js}] = \int_0^\infty t f_{js}(t) d(t).$$

We denote by $g_j = \sum_{s=1}^S g_{js}$ the expected value of the quantity of organs available at the donor hospital j .

Therefore, the number of medical teams moving to j needs to be less than or equal to g_j , which is the expected value of available organs in j :

$$\sum_{i=1}^m \sum_{k=1}^v g_{ijk} \leq g_j \quad \forall j = 1, \dots, n, \tag{5}$$

and the number of organs which are transported from the donor hospital j must be equal to the number of healthy and usable explanted organs:

$$\sum_{i=1}^m \sum_{k=1}^v \tilde{g}_{jsik} = (1 - \beta_j) g_{js} \quad \forall j = 1, \dots, n, \quad \forall s = 1, \dots, S. \tag{6}$$

In the past, it often happened that a patient entered in a waiting list for an organ transplant, when he received the call from the transplant center because a compatible organ was available, he had some difficulties in reaching the transplant center (because almost always the transplant center is not in the same city of residence). From the moment in which the patient is convened, it is necessary that he reaches the transplant center as quickly as possible. Hence, an agreement State-Regions Agreement 55/CSR of March 25, 2015 has been announced, which provides that the Regions or the Autonomous Provinces have to take charge of all transports carried out as part of the collection and transplant activity, including the transport of patient transplant candidates on the occasion of the received convocation.

We recall that every organ of the type s has a cold ischemia time \bar{R}_s . Specifically, “cold ischemia time during organ transplantation begins when the organ is cooled with a cold perfusion solution after organ procurement surgery, and ends after the tissue reaches physiological temperature during implantation procedures” (<http://www.reference.md/files/D050/mD050377.html>).

Since quickness is a key factor in the transplant process, we introduce:

- $t_{kij_s}^p$, which is the time that the patient assigned to the organ in the class O_{j_s} needs to reach i , via k ;
- t_{jik} , which is the time that the medical team takes to go from j to i , via k ;
- $t_{j_s}^{PR}$, which is the time that the team takes to pick up the organ of the class O_{j_s} ;
- \bar{R}_s , which is the maximum ischemia time for the organ of type s .

Therefore, when $\tilde{g}_{jsik} \geq 0$, it results to be:

$$\max\{t_{jik}, t_{kij_s}^p - t_{j_s}^{PR}\} \leq \bar{R}_s, \tag{7}$$

$$j = 1, \dots, n, \quad s = 1, \dots, S, \quad k = 1, \dots, v, \quad i = 1, \dots, m, \quad p = 1, \dots, P.$$

Let $y_{kij_s}^p \geq 0$ be the number of patients who have to be transported via k , from j to i , for the transplantation of the organ in the class O_{j_s} , such that:

$$\sum_{z=1}^v \tilde{g}_{jsiz} = \sum_{p=1}^P \sum_{k=1}^v y_{kij_s}^p, \quad i = 1, \dots, m, \quad j = 1, \dots, n, \quad s = 1, \dots, S, \tag{8}$$

namely we require that the number of patients who need to be transported to i for the transplant of the organ in the class O_{j_s} is the same as the number such organs transported in i .

Hence, constraint (7) can be also rewritten as:

$$t_{jik}\tilde{g}_{jsik} \leq \bar{R}_s\tilde{g}_{jsik},$$

$$\forall j = 1, \dots, n, \forall s = 1, \dots, S, \forall i = 1, \dots, m, \forall k = 1, \dots, v; \tag{9}$$

$$t_{kij}^p y_{kij}^p - t_{js}^{PR} y_{kij}^p \leq \bar{R}_s y_{kij}^p,$$

$$\forall k = 1, \dots, v, \forall i = 1, \dots, m, \forall j = 1, \dots, n, \forall s = 1, \dots, S, \forall p = 1, \dots, P. \tag{10}$$

Let c_{kij}^p be the transportation cost for the patient assigned to the organ in O_{js} of i , via k , and let us assume that such a cost is a function of the number of patients transported from i via k :

$$c_{kij}^p = c_{kij}^p(y_{kij}^p)$$

$$\forall k = 1, \dots, v, \forall i = 1, \dots, m, \forall j = 1, \dots, n, \forall s = 1, \dots, S, \forall p = 1, \dots, P.$$

Let us denote by $\Delta_j^- \equiv \max\{0, \tilde{g}_j - \sum_{i=1}^m g_{ij}\}$ e $\Delta_j^+ \equiv \max\{0, \sum_{i=1}^m g_{ij} - \tilde{g}_j\}$, $\forall j = 1, \dots, n$, the lack of medical teams ready for transplant and the excess of medical teams in j , then we have

$$\mathbb{E}[\Delta_j^-] = \int_0^{\infty} \sum_{i=1}^m g_{ij} \left(t - \sum_{i=1}^m g_{ij} \right) f_j(t) d(t), \quad \forall j = 1, \dots, n,$$

$$\mathbb{E}[\Delta_j^+] = \int_0^{\sum_{i=1}^m g_{ij}} \sum_{i=1}^m g_{ij} \left(\sum_{i=1}^m g_{ij} - t \right) f_j(t) d(t), \quad \forall j = 1, \dots, n.$$

We also assume that δ^- e δ^+ are the penalties to be paid in the case of an available and unused body or an excess of teams, respectively.

Let c_j^S be the health costs due to the organ removal at the hospital j and we assume such costs as a function of $\sum_{i=1}^m g_{ij} - \mathbb{E}[\Delta_j^+]$:

$$c_j^S = c_j^S\left(\sum_{i=1}^m g_{ij} - \mathbb{E}[\Delta_j^+]\right), \quad \forall j = 1, \dots, n.$$

Thus, the aim of the model consists in finding the optimal quantities of medical teams that must be moved from the transplant centers to the donor hospitals, choosing the most appropriate means, the optimal quantities of organs to be taken, to be transported (choosing the vehicle), and to be implanted, in order to minimize

transport costs of both teams and organs, as well as those of transplant patients, the costs of removal, of transplantation, and of post-transplantation, the costs of disposal of diseased or non-functioning organs and of those damaged, the penalties.

The problem formulation is as follows:

$$\begin{aligned}
\min \left\{ \sum_{i=1}^m \sum_{j=1}^n \sum_{k=1}^v 2c_{ijk}^{TE}(g_{ijk}) + \sum_{j=1}^n c_j^S \left(\sum_{i=1}^m g_{ij} - \mathbb{E}[\Delta_j^+] \right) \right. \\
+ \sum_{j=1}^n \sum_{i=1}^m \sum_{k=1}^v c_{jik}^{TO} \left(\sum_{s=1}^S \tilde{g}_{jsik} \right) + \sum_{i=1}^m \tilde{c}_i^S(\tilde{g}_i) + \sum_{i=1}^m c_i^{POST}(\tilde{g}_i) \\
+ \sum_{j=1}^n c_j^W \beta_j \left(\sum_{i=1}^m g_{ij} - \mathbb{E}[\Delta_j^+] \right) + \sum_{i=1}^m \tilde{c}_i^W \cdot \left(\tilde{g}_i + \sum_{j=1}^n \sum_{k=1}^v \gamma_{jik} \cdot \left(\sum_{s=1}^S \tilde{g}_{jsik} \right) \right) \\
\left. + \sum_{p=1}^P \sum_{k=1}^v \sum_{i=1}^m \sum_{j=1}^n \sum_{s=1}^S c_{kijps}^p (y_{kijps}^p) + \sum_{j=1}^n \delta^- \mathbb{E}[\Delta_j^-] + \sum_{j=1}^n \delta^+ \mathbb{E}[\Delta_j^+] \right\} \quad (11)
\end{aligned}$$

subject to the constraints:

$$g_{ijk} \geq 0 \quad i = 1, \dots, m, j = 1, \dots, n, k = 1, \dots, v; \quad (12)$$

$$g_{ij} \geq 0 \quad i = 1, \dots, m, j = 1, \dots, n; \quad (13)$$

$$\tilde{g}_{jsik} \geq 0 \quad j = 1, \dots, n, s = 1, \dots, S, i = 1, \dots, m, k = 1, \dots, v; \quad (14)$$

$$\tilde{g}_i \geq 0 \quad i = 1, \dots, m; \quad (15)$$

$$y_{kijps}^p \geq 0 \quad k = 1, \dots, v, i = 1, \dots, m,$$

$$j = 1, \dots, n, s = 1, \dots, S, p = 1, \dots, P; \quad (16)$$

$$g_{ij} \leq \bar{M} \quad i = 1, \dots, m, j = 1, \dots, n. \quad (17)$$

$$\tilde{g}_i = \sum_{j=1}^n \sum_{k=1}^v \left[(1 - \gamma_{jik}) \left(\sum_{s=1}^S \tilde{g}_{jsik} \right) \right] \quad i = 1, \dots, m; \quad (18)$$

$$\sum_{i=1}^m \sum_{k=1}^v \tilde{g}_{jsik} = (1 - \beta_j) g_{js} \quad j = 1, \dots, n, s = 1, \dots, S; \quad (19)$$

$$t_{jik} \tilde{g}_{jsik} \leq \bar{R}_s \tilde{g}_{jsik}$$

$$j = 1, \dots, n, s = 1, \dots, S, i = 1, \dots, m, k = 1, \dots, v; \quad (20)$$

$$t_{kij_s}^p y_{kij_s}^p - t_{j_s}^{PR} y_{kij_s}^p \leq \bar{R}_s y_{kij_s}^p$$

$$k = 1, \dots, v, i = 1, \dots, m, j = 1, \dots, n, s = 1, \dots, S, p = 1, \dots, P; \tag{21}$$

$$\sum_{j=1}^n \sum_{k=1}^v g_{ijk} \leq \sum_{j=1}^n \sum_{s=1}^S \sum_{k=1}^v \tilde{g}_{jsik} \quad i = 1, \dots, m; \tag{22}$$

$$\sum_{k=1}^v g_{ijk} \leq g_{ij} \quad i = 1, \dots, m, j = 1, \dots, n; \tag{23}$$

$$\sum_{z=1}^v \tilde{g}_{jsiz} = \sum_{p=1}^P \sum_{k=1}^v y_{kij_s}^p \quad i = 1, \dots, m, j = 1, \dots, n, s = 1, \dots, S; \tag{24}$$

3 Variational Inequality

For convenience of expression let

$$a_j = \sum_{i=1}^m g_{ij}, \quad \forall j = 1, \dots, n.$$

Therefore, for each donor hospital,

$$\frac{\partial \mathbb{E}[\Delta_j^-]}{\partial g_{ij}} = \frac{\partial \mathbb{E}[\Delta_j^-]}{\partial a_j} \cdot \frac{\partial a_j}{\partial g_{ij}} = P_j \left(\sum_{i=1}^m g_{ij} \right) - 1,$$

$$\forall i = 1, \dots, m, \forall j = 1, \dots, n;$$

$$\frac{\partial \mathbb{E}[\Delta_j^+]}{\partial g_{ij}} = \frac{\partial \mathbb{E}[\Delta_j^+]}{\partial a_j} \cdot \frac{\partial a_j}{\partial g_{ij}} = P_j \left(\sum_{i=1}^m g_{ij} \right),$$

$$\forall i = 1, \dots, m, \forall j = 1, \dots, n.$$

We associate the Lagrange multiplier λ_j^1 with constraint (17) and we denote the associated optimal Lagrange multiplier by λ_j^{1*} . Similarly, Lagrange multipliers $\lambda_{ij}^2, \lambda_{j_s}^3, \lambda_{jsik}^4, \lambda_{pkij_s}^5, \lambda_i^6, \lambda_{ij}^7$ and λ_{ijs}^8 are associated with constraints (18)–(24).

We group these Lagrange multipliers into the vectors $\lambda^1, \dots, \lambda^8$, respectively. Let \mathbb{K} denote the feasible set such that:

$$\mathbb{K} = \{(G^1, G^2, G^3, G^4, Y) \in R^{mnv+mn+2nSmv+m+PnSvm} : (12)-(24) \text{ hold}\}.$$

Theorem 1 A vector $(G^{1*}, G^{2*}, G^{3*}, G^{4*}, Y^*) \in \mathbb{K}$ is an optimal solution of the minimization problem (11) under constraints (12)-(24) if and only if there exist $\lambda^{1*} \in \mathbb{R}_+^{mn}$, $\lambda^{2*} \in \mathbb{R}^m$, $\lambda^{3*} \in \mathbb{R}^{nS}$, $\lambda^{4*} \in \mathbb{R}_+^{nSmv}$, $\lambda^{5*} \in \mathbb{R}_+^{PnSvm}$, $\lambda^{6*} \in \mathbb{R}_+^m$, $\lambda^{7*} \in \mathbb{R}_+^{mn}$, $\lambda^{8*} \in \mathbb{R}^{nSm}$, such that

$(G^{1*}, G^{2*}, G^{3*}, G^{4*}, Y^*, \lambda^{1*}, \lambda^{2*}, \lambda^{3*}, \lambda^{4*}, \lambda^{5*}, \lambda^{6*}, \lambda^{7*}, \lambda^{8*})$ is a solution to the following variational inequality:

$$\begin{aligned} & \sum_{i=1}^m \sum_{j=1}^n \sum_{k=1}^v \left[2 \frac{\partial c_{ijk}^{TE}(g_{ijk}^*)}{\partial g_{ijk}} + \lambda_i^{6*} + \lambda_{ij}^{7*} \right] \times [g_{ijk} - g_{ijk}^*] \\ & + \sum_{i=1}^m \sum_{j=1}^n \left[\frac{\partial c_j^S(\sum_{i=1}^m \sum_{k=1}^v g_{ij}^* - \mathbb{E}[\Delta_j^+])}{\partial g_{ij}} + c_j^W \beta_j - c_j^W \beta_j \frac{\partial \mathbb{E}[\Delta_j^+]}{\partial g_{ij}} \right. \\ & \quad \left. + \delta^- \frac{\partial \mathbb{E}[\Delta_j^-]}{\partial g_{ij}} + \delta^+ \frac{\partial \mathbb{E}[\Delta_j^+]}{\partial g_{ij}} + \lambda_{ij}^{1*} - \lambda_{ij}^{7*} \right] \times [g_{ij} - g_{ij}^*] \\ & + \sum_{j=1}^n \sum_{s=1}^S \sum_{i=1}^m \sum_{k=1}^v \left[\frac{\partial c_{jik}^{TO}(\tilde{g}_{jsik}^*)}{\partial \tilde{g}_{jsik}} + \gamma_{jik} \tilde{c}_i^W + \lambda_i^{2*} (1 - \gamma_{jik}) \right. \\ & \quad \left. + \lambda_{js}^{3*} + \lambda_{jsik}^{4*} (t_{jik} - \bar{R}_s) - \lambda_i^{6*} - \lambda_{ijs}^{8*} \right] \times [\tilde{g}_{jsik} - \tilde{g}_{jsik}^*] \\ & + \sum_{i=1}^m \left[\frac{\partial \tilde{c}_i^S(\tilde{g}_i^*)}{\partial \tilde{g}_i} + \frac{\partial c_i^{POST}(\tilde{g}_i^*)}{\partial \tilde{g}_i} + \tilde{c}_i^W - \lambda_i^{2*} \right] \times [\tilde{g}_i - \tilde{g}_i^*] \\ & + \sum_{p=1}^P \sum_{k=1}^v \sum_{i=1}^m \sum_{j=1}^n \sum_{s=1}^S \left[\frac{\partial c_{kij s}^p(y_{kij s}^{p*})}{\partial y_{kij s}^p} + \lambda_{pkij s}^{5*} (t_{kij s}^p - t_{js}^{PR} - \bar{R}_s) \right. \\ & \quad \left. + \lambda_{ijs}^{8*} \right] \times [y_{kij s}^p - y_{kij s}^{p*}] \\ & - \sum_{i=1}^m \sum_{j=1}^n \left[g_{ij} - \bar{M} \right] \times [\lambda_{ij}^1 - \lambda_{ij}^{1*}] \\ & - \sum_{i=1}^m \left[\sum_{j=1}^n \sum_{k=1}^v \left[(1 - \gamma_{jik}) \left(\sum_{s=1}^S \tilde{g}_{jsik}^* \right) \right] - \tilde{g}_i^* \right] \times [\lambda_i^2 - \lambda_i^{2*}] \\ & - \sum_{j=1}^n \sum_{s=1}^S \left[\sum_{i=1}^m \sum_{k=1}^v \tilde{g}_{jsik}^* - (1 - \beta_j) g_{js} \right] \times [\lambda_{js}^3 - \lambda_{js}^{3*}] \end{aligned}$$

$$\begin{aligned}
 & - \sum_{j=1}^n \sum_{s=1}^S \sum_{i=1}^m \sum_{k=1}^v \left[t_{jik} \tilde{g}_{jsik}^* - \bar{R}_s g_{jsik}^* \right] \times [\lambda_{jsik}^4 - \lambda_{jsik}^{4*}] \\
 & - \sum_{p=1}^P \sum_{k=1}^v \sum_{i=1}^m \sum_{j=1}^n \sum_{s=1}^S \left[t_{kijps}^p y_{kijps}^{p*} - t_{jps}^{PR} y_{kijps}^{p*} - \bar{R}_s y_{kijps}^{p*} \right] \times [\lambda_{pkijps}^5 - \lambda_{pkijps}^{5*}] \\
 & - \sum_{i=1}^m \left[\sum_{j=1}^n \sum_{k=1}^v g_{ijk}^* - \sum_{j=1}^n \sum_{s=1}^S \sum_{k=1}^v \tilde{g}_{jsik}^* \right] \times [\lambda_i^6 - \lambda_i^{6*}] \\
 & - \sum_{i=1}^m \sum_{j=1}^n \left[\sum_{k=1}^v g_{ijk}^* - g_{ij}^* \right] \times [\lambda_{ij}^7 - \lambda_{ij}^{7*}] \\
 & - \sum_{i=1}^m \sum_{j=1}^n \sum_{s=1}^S \left[\sum_{k=1}^v y_{kijps}^* - \sum_{z=1}^v \tilde{g}_{jsiz}^* \right] \times [\lambda_{ijs}^8 - \lambda_{ijs}^{8*}] \geq 0, \tag{25}
 \end{aligned}$$

$\forall (G^1, G^2, G^3, G^4, Y, \lambda^1, \lambda^2, \lambda^3, \lambda^4, \lambda^5, \lambda^6, \lambda^7, \lambda^8) \in V$, where

$$V = \mathbb{K} \times \mathbb{R}_+^{mn} \times \mathbb{R}^m \times \mathbb{R}^{nS} \times \mathbb{R}_+^{nSmv} \times \mathbb{R}_+^{PnSvm} \times \mathbb{R}_+^m \times \mathbb{R}_+^{mn} \times \mathbb{R}^{nSm}.$$

The following theorem ensures the existence of solutions to (25).

Theorem 2 *A solution $(G^{1*}, G^{2*}, G^{3*}, G^{4*}, Y^*, \lambda^{1*}, \lambda^{2*}, \lambda^{3*}, \lambda^{4*}, \lambda^{5*}, \lambda^{6*}, \lambda^{7*}, \lambda^{8*}) \in V$ to variational inequality (25) is guaranteed to exist.*

Proof The result follows from the classical theory of variational inequalities (see [8]), since the feasible set is compact and the function that enters the variational inequality is continuous.

4 Numerical Examples

In order to further illustrate the above model, in this section we present several simple numerical examples.

In the following examples we consider some quadratic cost functions which in some sense represent the reality, since the marginal cost functions decrease when a large number of transplants is performed.

The costs are in thousands of euros and the organs are in dozens of units.

Further, we assume that if the amount of organs transported via k from the hospital j to the transplant center i is strictly positive, then at least one team has to move from i to j via k , that is:

se $\sum_{s=1}^S \tilde{g}_{jsik} > 0 \Rightarrow g_{ijk} \geq 1$.

Such a constraint can be expressed as:

$$\sum_{s=1}^S \tilde{g}_{jsik} (1 - g_{ijk}) \leq 0, \quad \forall i = 1, \dots, m, \quad \forall j = 1, \dots, n, \quad \forall k = 1, \dots, v. \tag{26}$$

The optimal solutions are calculated using the Matlab program, by applying the projection-contraction method proposed by Solodov and Tseng (see [13]). The algorithm was implemented on a laptop with an AMD A6-9225 Radeon R4, 5 compute cores2C+3G, 2.6 GHz processor and 8 GB RAM. For all the analyzed cases, we have depicted the underlying network and specified the cost functions.

4.1 Example 1

In this example we consider a simple network consisting of one transplant center, one donor hospital with 2 classes of organs and with two different transportation modes, as depicted in Figure 2.

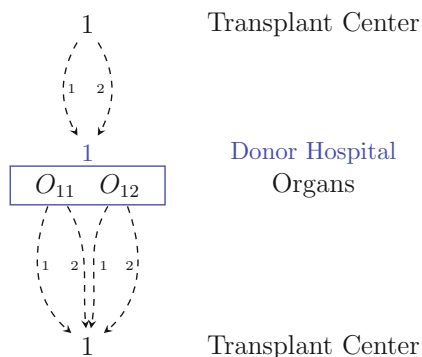
The transportation costs associated with the medical teams from the transplant center 1 to the donor hospital 1 using the first and the second transportation service, respectively, are

$$c_{111}^{TE}(g_{111}) = 0.25 \cdot g_{111}^2 + 0.2 \cdot g_{111} + 8,$$

$$c_{112}^{TE}(g_{112}) = 1.25 \cdot g_{112}^2 + 1.2 \cdot g_{112} + 18.$$

The transportation costs associated with the organs from the donor hospital 1 to the transplant center 1 using the first and the second transportation service, respectively,

Fig. 2 Organs transport network: Example 1



are

$$c_{111}^{TO}(\tilde{g}_{1111} + \tilde{g}_{1211}) = 0.75 \cdot (\tilde{g}_{1111} + \tilde{g}_{1211})^2 + 0.4 \cdot (\tilde{g}_{1111} + \tilde{g}_{1211}) + 5,$$

$$c_{112}^{TO}(\tilde{g}_{1112} + \tilde{g}_{1212}) = 1.75 \cdot (\tilde{g}_{1112} + \tilde{g}_{1212})^2 + 1.4 \cdot (\tilde{g}_{1112} + \tilde{g}_{1212}) + 25.$$

The costs for transporting patients assigned to organs, through the first or the second transportation service, are

$$c_{111s}^p(y_{111s}^p) = (y_{111s}^p)^2 + y_{111s}^p + 0.5,$$

$$c_{211s}^p(y_{211s}^p) = 2 \cdot (y_{211s}^p)^2 + 2 \cdot y_{211s}^p + 1.5.$$

Furthermore, we assume $t_{111} = 6$, $t_{112} = 4$, $\bar{R}_1 = 10$, $\bar{R}_2 = 5$, $t_{1111}^p = 11$, $t_{2111}^p = 6$, $t_{1112}^p = 11$, $t_{2112}^p = 6$, $t_{11}^{PR} = t_{12}^{PR} = 3$.

The solution is as follows:

$$g_{111}^* = 1, \quad g_{112}^* = 1, \quad g_{11}^* = 11,$$

$$\tilde{g}_{1111}^* = \tilde{g}_{1212}^* = 3, \quad \tilde{g}_{1211}^* = \tilde{g}_{1112}^* = 0, \quad \tilde{g}_1^* = 6,$$

$$y_{1111}^{p*} = 2, \quad y_{2111}^{p*} = 1, \quad y_{1112}^{p*} = 0, \quad y_{2112}^{p*} = 3,$$

$$\lambda_1^{1*} = 0, \quad \lambda_1^{2*} = 0.01, \quad \lambda_{11}^{3*} = \lambda_{12}^{3*} = 0.24,$$

$$\lambda_{1111}^{4*} = 0, \quad \lambda_{1211}^{4*} = 6.76, \quad \lambda_{1112}^{4*} = \lambda_{1212}^{4*} = 0,$$

$$\lambda_{p1111}^{5*} = \lambda_{p2111}^{5*} = 0, \quad \lambda_{p1112}^{5*} = 4.53, \quad \lambda_{p2112}^{5*} = 0,$$

$$\lambda_1^{6*} = 0, \quad \lambda_{11}^{7*} = 0, \quad \lambda_{111}^{8*} = \lambda_{112}^{8*} = 0.01.$$

The elapsed time is 13.39 s.

We note that it is more convenient to use the first transportation service for bringing the organs of type 1 and the second transportation service (that is more expensive) for those of type 2 because of the ischemia time.

Instead, it is more suitable to transport patients associated with the second type of organs always with the second transportation service, while, with regard to the first type of organ, two patients with the first transportation service and one with the second.

Example 1.2

This example has the same data as the previous one, except that now we assume that the portion of explanted organs discarded in the donor hospital is $\beta_1 = 0.5$, the portion of organs reaching the transplant center, but which cannot be transplanted, is $\gamma_{111} = \gamma_{112} = 0.1$ and $\tilde{g}_{11}, \tilde{g}_{12} \sim U([0, 8])$.

The solution is

$$\begin{aligned}
 g_{111}^* &= 1, & g_{112}^* &= 1, & g_{11}^* &= 14, \\
 \tilde{g}_{1111}^* &= \tilde{g}_{1212}^* = 2, & \tilde{g}_{1211}^* &= \tilde{g}_{1112}^* = 0, & \tilde{g}_1^* &= 4, \\
 y_{1111}^{p*} &= 1, & y_{2111}^{p*} &= 1, & y_{1112}^{p*} &= 0, & y_{2112}^{p*} &= 2, \\
 \lambda_1^{1*} &= 0, & \lambda_1^{2*} &= 0.0042, & \lambda_{11}^{3*} &= \lambda_{12}^{3*} = 0.16, \\
 \lambda_{1111}^{4*} &= 0, & \lambda_{1211}^{4*} &= 4.76, & \lambda_{1112}^{4*} &= 0.0027, & \lambda_{1212}^{4*} &= 0, \\
 \lambda_{p1111}^{5*} &= \lambda_{p2111}^{5*} = 0, & \lambda_{p1112}^{5*} &= 3.14, & \lambda_{p2112}^{5*} &= 0, \\
 \lambda_1^{6*} &= 0, & \lambda_{11}^{7*} &= 0, & \lambda_{111}^{8*} &= 0.0063, & \lambda_{112}^{8*} &= 0.0072.
 \end{aligned}$$

The elapsed time is 16.10 s.

In Figure 3, we show the trend of quantity of organs that the medical teams in i intend to take from the donor hospital j , when the penalty δ^- changes.

4.2 Example 2

In this example we consider a new network consisting of two transplant centers, five donor hospitals with 2 classes of organs and with two different transportation modes, as depicted in Figure 4.

The new solution is as follows:

$$\begin{aligned}
 g_{111}^* &= 1.00, & g_{112}^* &= 0.98, & g_{121}^* &= 1.00, & g_{122}^* &= 0.79, & g_{131}^* &= 1.00, \\
 g_{132}^* &= 0.98, & g_{141}^* &= 1.00, & g_{142}^* &= 1.00, & g_{151}^* &= 1.00, & g_{152}^* &= 0.99, \\
 g_{211}^* &= 1.00, & g_{212}^* &= 0.98, & g_{221}^* &= 1.00, & g_{222}^* &= 1.00, & g_{231}^* &= 1.00, \\
 g_{232}^* &= 0.98, & g_{241}^* &= 1.00, & g_{242}^* &= 0.90, & g_{251}^* &= 1.00, & g_{252}^* &= 1.00, \\
 g_{11}^* &= 6.96, & g_{12}^* &= 6.96, & g_{13}^* &= 6.96, & g_{14}^* &= 6.96, & g_{15}^* &= 6.96,
 \end{aligned}$$

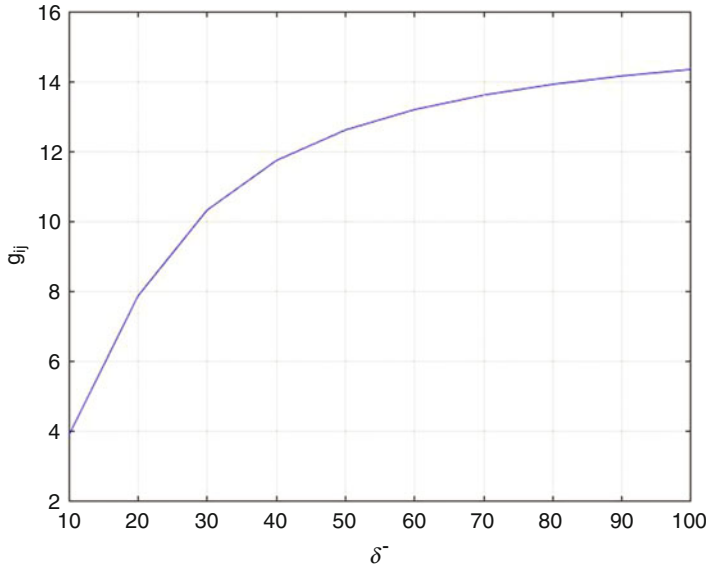


Fig. 3 Values of g_{ij} when δ^- varies

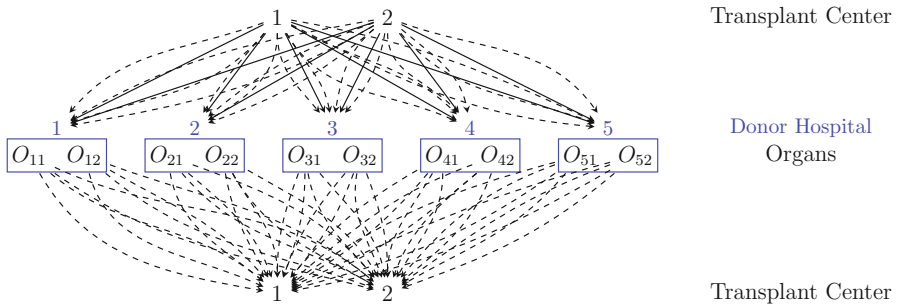


Fig. 4 Organs transport network: Example 2

$$\begin{aligned}
 &g_{21}^* = 6.96, \quad g_{22}^* = 6.96, \quad g_{23}^* = 6.96, \quad g_{24}^* = 6.96, \quad g_{25}^* = 6.96, \\
 &\tilde{g}_{1111}^* = 0.81, \quad \tilde{g}_{1112}^* = 0.20, \quad \tilde{g}_{1121}^* = 0.79, \quad \tilde{g}_{1122}^* = 0.20, \quad \tilde{g}_{1211}^* = 0.81, \\
 &\tilde{g}_{1212}^* = 0.20, \quad \tilde{g}_{1221}^* = 0.79, \quad \tilde{g}_{1222}^* = 0.20, \quad \tilde{g}_{2111}^* = 1.24, \quad \tilde{g}_{2112}^* = 0.27, \\
 &\tilde{g}_{2121}^* = 0.00, \quad \tilde{g}_{2122}^* = 0.49, \quad \tilde{g}_{2211}^* = 0.05, \quad \tilde{g}_{2212}^* = 0.00, \quad \tilde{g}_{2221}^* = 1.95, \\
 &\tilde{g}_{2222}^* = 0.00, \quad \tilde{g}_{3111}^* = 0.81, \quad \tilde{g}_{3112}^* = 0.20, \quad \tilde{g}_{3121}^* = 0.79, \quad \tilde{g}_{3122}^* = 0.20, \\
 &\tilde{g}_{3211}^* = 0.81, \quad \tilde{g}_{3212}^* = 0.20, \quad \tilde{g}_{3221}^* = 0.79, \quad \tilde{g}_{3222}^* = 0.20, \quad \tilde{g}_{4111}^* = 0.94,
 \end{aligned}$$

$$\tilde{g}_{4112}^* = 0.00, \quad \tilde{g}_{4121}^* = 0.80, \quad \tilde{g}_{4122}^* = 0.26, \quad \tilde{g}_{4211}^* = 0.55, \quad \tilde{g}_{4212}^* = 0.80,$$

$$\tilde{g}_{4221}^* = 0.59, \quad \tilde{g}_{4222}^* = 0.05, \quad \tilde{g}_{5111}^* = 0.66, \quad \tilde{g}_{5112}^* = 0.04, \quad \tilde{g}_{5121}^* = 1.30,$$

$$\tilde{g}_{5122}^* = 0.00, \quad \tilde{g}_{5211}^* = 1.01, \quad \tilde{g}_{5212}^* = 0.39, \quad \tilde{g}_{5221}^* = 0.00, \quad \tilde{g}_{5222}^* = 0.60,$$

$$\tilde{g}_1^* = 8.99, \quad \tilde{g}_2^* = 9.00,$$

$$\tilde{y}_{1111}^{1*} = 0.50, \quad \tilde{y}_{1112}^{1*} = 0.50, \quad \tilde{y}_{1121}^{1*} = 0.67, \quad \tilde{y}_{1122}^{1*} = 0.02, \quad \tilde{y}_{1131}^{1*} = 0.50,$$

$$\tilde{y}_{1132}^{1*} = 0.50, \quad \tilde{y}_{1141}^{1*} = 0.47, \quad \tilde{y}_{1142}^{1*} = 0.28, \quad \tilde{y}_{1151}^{1*} = 0.35, \quad \tilde{y}_{1152}^{1*} = 0.63,$$

$$\tilde{y}_{1211}^{1*} = 0.49, \quad \tilde{y}_{1212}^{1*} = 0.49, \quad \tilde{y}_{1221}^{1*} = 0.25, \quad \tilde{y}_{1222}^{1*} = 0.00, \quad \tilde{y}_{1231}^{1*} = 0.49,$$

$$\tilde{y}_{1232}^{1*} = 0.49, \quad \tilde{y}_{1241}^{1*} = 0.52, \quad \tilde{y}_{1242}^{1*} = 0.32, \quad \tilde{y}_{1251}^{1*} = 0.60, \quad \tilde{y}_{1252}^{1*} = 0.30,$$

$$\tilde{y}_{2111}^{1*} = 0.00, \quad \tilde{y}_{2112}^{1*} = 0.00, \quad \tilde{y}_{2121}^{1*} = 0.08, \quad \tilde{y}_{2122}^{1*} = 0.00, \quad \tilde{y}_{2131}^{1*} = 0.00,$$

$$\tilde{y}_{2132}^{1*} = 0.00, \quad \tilde{y}_{2141}^{1*} = 0.00, \quad \tilde{y}_{2142}^{1*} = 0.00, \quad \tilde{y}_{2151}^{1*} = 0.00, \quad \tilde{y}_{2152}^{1*} = 0.07,$$

$$\tilde{y}_{2211}^{1*} = 0.00, \quad \tilde{y}_{2212}^{1*} = 0.00, \quad \tilde{y}_{2221}^{1*} = 0.00, \quad \tilde{y}_{2222}^{1*} = 0.00, \quad \tilde{y}_{2231}^{1*} = 0.00,$$

$$\tilde{y}_{2232}^{1*} = 0.00, \quad \tilde{y}_{2241}^{1*} = 0.01, \quad \tilde{y}_{2242}^{1*} = 0.00, \quad \tilde{y}_{2251}^{1*} = 0.00, \quad \tilde{y}_{2252}^{1*} = 0.00,$$

$$\tilde{y}_{1111}^{2*} = 0.50, \quad \tilde{y}_{1112}^{2*} = 0.50, \quad \tilde{y}_{1121}^{2*} = 0.67, \quad \tilde{y}_{1122}^{2*} = 0.02, \quad \tilde{y}_{1131}^{2*} = 0.50,$$

$$\tilde{y}_{1132}^{2*} = 0.50, \quad \tilde{y}_{1141}^{2*} = 0.47, \quad \tilde{y}_{1142}^{2*} = 0.28, \quad \tilde{y}_{1151}^{2*} = 0.35, \quad \tilde{y}_{1152}^{2*} = 0.63,$$

$$\tilde{y}_{1211}^{2*} = 0.49, \quad \tilde{y}_{1212}^{2*} = 0.49, \quad \tilde{y}_{1221}^{2*} = 0.25, \quad \tilde{y}_{1222}^{2*} = 0.00, \quad \tilde{y}_{1231}^{2*} = 0.49,$$

$$\tilde{y}_{1232}^{2*} = 0.49, \quad \tilde{y}_{1241}^{2*} = 0.52, \quad \tilde{y}_{1242}^{2*} = 0.32, \quad \tilde{y}_{1251}^{2*} = 0.60, \quad \tilde{y}_{1252}^{2*} = 0.30,$$

$$\tilde{y}_{2111}^{2*} = 0.00, \quad \tilde{y}_{2112}^{2*} = 0.00, \quad \tilde{y}_{2121}^{2*} = 0.08, \quad \tilde{y}_{2122}^{2*} = 0.00, \quad \tilde{y}_{2131}^{2*} = 0.00,$$

$$\tilde{y}_{2132}^{2*} = 0.00, \quad \tilde{y}_{2141}^{2*} = 0.00, \quad \tilde{y}_{2142}^{2*} = 0.00, \quad \tilde{y}_{2151}^{2*} = 0.00, \quad \tilde{y}_{2152}^{2*} = 0.07,$$

$$\tilde{y}_{2211}^{2*} = 0.00, \quad \tilde{y}_{2212}^{2*} = 0.00, \quad \tilde{y}_{2221}^{2*} = 0.00, \quad \tilde{y}_{2222}^{2*} = 0.00, \quad \tilde{y}_{2231}^{2*} = 0.00,$$

$$\tilde{y}_{2232}^{2*} = 0.00, \quad \tilde{y}_{2241}^{2*} = 0.01, \quad \tilde{y}_{2242}^{2*} = 0.00, \quad \tilde{y}_{2251}^{2*} = 0.05, \quad \tilde{y}_{2252}^{2*} = 0.00.$$

The elapsed time is 1048.45 s.

5 Conclusions

Thanks to the advances of scientific research and progress in medicine, every year organ transplants save thousands of lives.

The increase in the number of organ transplantations made necessary to improve the organization of the whole network. Particularly, time and cost management was considered fundamental to handle the transplant system and to coordinate all the process phases which are necessary to reach the goal as quickly as possible.

In this paper, we deal with this important problem of optimizing a network of organ transplants with uncertain availability.

We analyze the Italian Health System and introduce the network underlying the organs transplant system. From the *transplant centers*, the medical teams reach the *donor hospitals* which can own different types of *available organs*. At the donor hospitals, the medical teams perform the organ removal and then they, together with the organs, go back to the transplant centers, where the organ transplant is performed.

As we all know, in each region the waiting lists never run out because the demand of organs is greater than the availability; moreover, we consider the quantities of organs available at the donor hospitals as random variables. Therefore, we take into account the penalties to be paid in the case of lack of medical teams ready for transplant or excess of teams.

In addition, the cold ischemia time plays an important role in the transplant process. For this reason we also treat the problem of choosing the appropriate vehicle, even because the fastest means is often the most expensive, while the cheapest one could be so slow that it does not allow us to perform the transplant in time.

As a consequence, the aim of this paper is to present a model, based on networks, consisting of finding the optimal quantities of medical teams that must move from the transplant centers to the donor hospitals, choosing the most appropriate means, the optimal quantities of organs to be taken, to be transported (choosing the vehicle) and to be implanted, in order to minimize transport costs of both teams and organs, as well as those of transplant patients, the costs of removal, of transplantation and of post-transplantation, the costs of disposal of diseased or non-functioning organs and of those damaged, the penalties. Moreover, our model ensures that the constraints on cold ischemia times are respected and that the quantities of organs available at the donor hospitals, which are random variables, are entirely used.

We then determine the optimality conditions for the national health service and derive the variational inequality formulation introducing the Lagrange multipliers associated with the constraints and we also report existence results for the solution. In addition, we also apply the model to some numerical examples.

In a future work we can examine the role of the Lagrange multipliers associated with the constraints (see [6, 11, 12]), which turns out to be very useful for the analysis of the behavior of the decision makers in the network.

References

1. O. Alagoz, A.J. Schaefer, M.S. Roberts, Optimizing organ allocation and acceptance, in *Handbook of Optimization in Medicine* (Springer, Berlin, 2013), pp. 1–24
2. J. Belien, L. De Boeck, J. Colpaert, S. Devesse, F. Van den Bossche, Optimizing the supply chain design for organ transplants, Hub Research Paper, Economics & Management (2011)
3. R.D. Bloom, A.M. Doyle, Evaluation of the kidney transplant candidate and follow-up of the listed patient, in *Kidney Transplantation: A Guide to the Care of Kidney Transplant Recipients*, ed. by D. McKay, S. Steinberg (Springer, Boston, 2010)
4. M.E. Bruni, D. Conforti, N. Sicilia, S. Trotta, A new organ transplantation location-allocation policy: a case study of Italy. *Health Care Manag. Sci.* **9**, 125–142 (2006)
5. V. Caruso, P. Daniele, A network model for minimizing the total organ transplant costs. *Eur. J. Oper. Res.* **266**(2), 652–662 (2018)
6. G. Colajanni, P. Daniele, S. Giuffrè, A. Nagurney, Cybersecurity investments with nonlinear budget constraints and conservation laws: variational equilibrium, marginal expected utilities, and Lagrange multipliers. *Int. Trans. Oper. Res.* **25**(5), 1443–1464 (2018)
7. G.M. Danovitch, S. Hariharan, J.D. Pirsch, D. Rush, D. Roth, E. Ramos, R.C. Starling, C. Cangro, M.R. Weir, Clinical Practice Guidelines Committee of the American Society of Transplantation, Management of the waiting list for cadaveric kidney transplants: report of a survey and recommendations by the Clinical Practice Guidelines Committee of the American Society of Transplantation. *J. Am. Soc. Nephrol.* **13**, 528–535 (2002)
8. D. Kinderlehrer, G. Stampacchia, *Variational Inequalities and Their Applications* (Academic Press, New York, 1980)
9. N. Koizumi, R. Ganesan, M. Gentili, C.-H. Chen, N. Waters, D. DasGupta, D. Nicholas, A. Patel, D. Srinivasan, K. Melancon, *Redesigning Organ Allocation Boundaries for Liver Transplantation in the United States*. Springer Proceedings in Mathematics & Statistics, vol. 61 (Springer, Berlin, 2014)
10. N. Kong, A.J. Schaefer, B. Hunsaker, M.S. Roberts, Maximizing the efficiency of the US liver allocation system through region design. *Manag. Sci.* **56**(12), 2111–2122 (2010)
11. A. Nagurney, P. Daniele, E. Alvarez Flores, V. Caruso, A variational equilibrium network framework for humanitarian organizations in disaster relief: Effective product delivery under competition for financial funds, in *Dynamics of Disasters*, ed. by I.S. Kotsireas, A. Nagurney, P.M. Pardalos. Springer Optimization and Its Applications, vol. 140 (Springer, Berlin, 2018), pp. 109–133
12. A. Nagurney, M. Salarpour, P. Daniele, An integrated financial and logistical game theory model for humanitarian organizations with purchasing costs, multiple freight service providers, and budget, capacity, and demand constraints. *Int. J. Prod. Econ.* **212**, 212–226 (2019)
13. M.V. Solodov, P. Tseng, Modified projection-type methods for monotone variational inequalities. *SIAM J. Control Optim.* **34**(5), 1814–1830 (1996)

Algebraic Based Techniques as Decision Making Tools



M. Couceiro, G. Meletiou, and K. Skouri

Abstract This study explores the use of some well-established algebraic structures as tools in multicriteria decision making. Under a rigorous axiomatic foundation, a complex decision problem is decomposed into a multilevel hierarchic structure of objective, criteria, and alternatives. A priority is derived for each element of the hierarchy, allowing comparisons based on linear rankings, weak orders, and other order structures. Consensus rules are provided for the final ranking of the alternatives.

1 Introduction

Group decision making [23, 27, 29] deals with the consolidation and aggregation of preferences that a set of criteria (or a group of decision-makers) orders a set of alternatives, aiming to determine the best collective alternative solution. So, it is required to establish a consensus reaching process to obtain an acceptable common preference order—collective solution [10, 22]. The consensus has been defined in different ways in the literature [29]: consensus is defined as a consolidate—aggregate preference [3], and regards the approaches, tools, and procedures leading to the final decision [1, 24]. Also, consensus is defined as the full and unanimous agreement of all the criteria (decision-makers) regarding all the feasible alternatives

M. Couceiro
University of Lorraine, CNRS, Inria, LORIA, Nancy, France
e-mail: miguel.couceiro@Loria.fr

G. Meletiou (✉)
Department of Agriculture, University of Ioannina, Arta, Greece
Computational Intelligence Laboratory (CILab), University of Patras, Patras, Greece
e-mail: gmelet@neptune.math.upatras.gr

K. Skouri
Department of Mathematics, University of Ioannina, Ioannina, Greece
e-mail: kskouri@uoi.gr

[8], and according to this definition as measure is used the characteristic function with values 0 (absence consensus) and 1 (full consensus). However, unanimity may be difficult to achieve, in particular with large and diversified groups of criteria (or decision-makers). So, the concept of consensus has been viewed in a more flexible way, which has led to the derivation and use of consensus metrics [14, 16, 18] aiming to achieve [29]: (1) better partial agreement and (2) the derivation of the consensus process until an acceptable high level of agreement is attained. Most of these proposals are closely related to Arrow's conditions [2, 4] that the aggregation procedure should fulfil, namely full rationality of the overall preference relation, unrestricted domain, weak Paretian orderings, independence of irrelevant alternatives, and the non-dictatorship requirement. In this framework, it should be noticed the structural identity of social choice and multicriteria decision problems [15]. It had been established that the basic results of consumer theory could be reproduced assuming nothing stronger than the ability of indication, for any two goods, which is preferable to which (i.e. assuming ordinal utility). While, for Arrow, cardinal utility measure (as it has been established by Von Neumann and Morgenstern [28]) is not allowed in collective decision making processes due to its linking with the individual's attitude toward risk-taking [15].

In this ground, the present study proposes consensus reaching processes, linking algebraic tools and assuming ordinal scale, that lead to a collective solution in a multicriteria decision problem. The processes based on median computations and could be used as an alternative to analytic hierarchy process (AHP) decision tools within an ordinal scale framework requiring less and easiest computations.

2 Preliminaries

Let A be a (finite) set of elements (the set of alternatives). By $X = L(A)$ we denote the set of linear orderings of A . Every element \langle_i of X is a decision criterion in the sense that for $\langle_i \in X$, $a, b \in A$, $a \langle_i b$ means that b is preferred than a , or b dominates a according to the \langle_i criterion.

A consensus procedure is a function $F : X^n \rightarrow X$ where n is the number of criteria (decision-makers). Function F merges the n criteria $(\langle_1, \langle_2, \dots, \langle_n) \mapsto F(\langle_1, \langle_2, \dots, \langle_n) = \langle_T$ into a single one \langle_T .

Following [2] a consensus procedure has to satisfy the following conditions:

- (1) Unanimity or Ranking Preservation: For $a, b \in A$, if $a \langle_i b$, for $i = 1, \dots, n$, then $a \langle_T b$
- (2) Independence of Irrelevant Alternatives: Let $(\langle_1, \dots, \langle_n)$ and $(\langle_1^*, \dots, \langle_n^*)$ be preference profiles. Assume that $a, b \in A$ have the same order in \langle_i and \langle_i^* , for all $i = 1, \dots, n$. Then, a and b have the same order in $\langle_T = F(\langle_1, \dots, \langle_n)$ and $\langle_T^* = F(\langle_1^*, \dots, \langle_n^*)$
- (3) Non-dictatorship: There is no $i : 1 \leq i \leq n$ s.t. for all profiles $(\langle_1, \dots, \langle_n) : F(\langle_1, \dots, \langle_n) = \langle_i$

However, in the case $|A| \geq 2$ the three conditions are incompatible. In other terms (1) and (2) imply dictatorship [2]. This means that the above conditions pose limits to the merge of the different preference orders on a set of options into a single preference order [15]. So, it is natural to pose the questions: Is it just because of the Independence of Irrelevant Alternatives (condition (2)), or does the underlying structure play a role? Can the linear ordering $<_i$ be generalized? What about if \leq become partial orders (that is reflexive antisymmetric and transitive), but a and b are incomparable? What about the relations \leq become total weak orders, that is reflexive, transitive and total, that is for all a, b , a and b are always comparable, that is either $a \leq b$ or $b \leq a$? In this framework, other structures will be examined in order a collective solution to be derived. These structures overcome the dictatorship of the above conditions and lead to oligarchy. Median structures have been already used as tools for study of consensus rules.

3 Median Algebras

In this section a brief introduction on median algebras is presented. A median algebra is a ternary algebraic structure $\langle M, m \rangle$ consisting of a non-empty underlying set M and a ternary operation: $m : M^3 \rightarrow M$ satisfying [5, 19, 25, 26]:

- (1) $m(a, a, b) = a$ (Majority),
- (2) $m(a, b, c) = m(b, a, c) = m(b, c, a) = \dots$ (Symmetry)
- (3) $m(a, b, m(c, d, e)) = m(m(a, b, c), d, m(a, b, e))$ (Distributivity)

Examples of Median Algebras, that could be useful to decision making, include:

- (i) Linear Orders or Chains: Where $m(\cdot, \cdot, \cdot)$ is the betweenness operation
- (ii) More general, every distributive lattice (L, \vee, \wedge) is a median algebra with the self-dual ternary operation $m(a, b, c) = (a \wedge b) \vee (b \wedge c) \vee (a \wedge c) = (a \vee b) \wedge (b \vee c) \wedge (a \vee c)$ [9, 17]
- (iii) Median graphs: Connected simple graphs without loops having the property that for every three vertices $\{a, b, c\}$ there is exactly one vertex which lies on shortest path between each pair of vertices in $\{a, b, c\}$.

Notice that the concept of a median graph is a common generalization of a tree and a hypercube.

For every median algebra and an element $d \in M$, one may consider the binary operation: $\hat{d} : (a, b) \mapsto m(a, d, b)$. This operation is a meet semi-lattice operation and d is its zero, or absorptive element, or lower bound since $a\hat{d}d = m(a, d, d) = m(d, d, a) = d\hat{d}a = d$, for all $a \in M$. Also, for $e, d \in M$ the (binary) operations \hat{e}, \hat{d} are mutually distributive [5, §7.3].

The corresponding semi-lattice order related to the operation \hat{d} is defined as $p \leq_a q$ iff $m(p, d, q) = p\hat{d}q = p$, for $p, q \in M$. The structure $\langle M, \leq_a \rangle$ is a median semi-lattice in the sense that every principal ideal $\downarrow p := \{q \in M : q \leq_a p\}$

is a distributive lattice and such that for any $a, b, c \in M$ the set $\{a, b, c\}$ has an upper bound whenever each of its 2-element subsets $\{a, b\}$, $\{b, c\}$, $\{c, a\}$ has an upper bound [5, 7, 25, 26].

Also, every element $d \in M$ defines a comparison between the elements of M and d is the bound of \leq_d (Optimal element). Therefore, every element of M is both an element and a comparison criterion between elements.

The interest in median semi-lattices and their generalization grew out of the study of taxonomic structures. While, hierarchies, weak hierarchies splits, weak orders are also mentioned.

Every order of the form \leq_d , $d \in A$ is compatible with the median structure. There exist semi-lattice orders that are not of the form \leq_d but they preserve the median operation—for more details see [6]. The median algebra is extended to $\xi(M)$, its zero completion. All compatible semi-lattice orders can be considered as comparison criteria between the elements of M .

4 Median Homomorphisms as Consensus Procedures

In this section consensus procedures are presented that based on median homomorphism. For A, B median algebras, a median homomorphism is a function $f : A \rightarrow B$ satisfying: $f(m(a, b, c)) = m(f(a), f(b), f(c))$. The median homomorphism can be characterized as median preserving function. The main idea is to use median homomorphisms of the form $A_1 \times A_2 \times \dots \times A_n \rightarrow M$ as consensus procedures. We require the score of a median profile to be the median of the scores of the profiles.

However, it has to be mentioned that median preserving maps are not necessarily order-preserving maps. In the special case of the chains (totally ordered sets), a function is a median homomorphism iff it is order preserving or order reversing.

4.1 Consensus Over Trees

Consider A_1, \dots, A_n, B median algebras and $f : A_1 \times \dots \times A_n \rightarrow B$ a median preserving function that is a consensus procedure. According to [12, 13], all homomorphisms f are essentially unary iff B is a tree. That means:

- (1) If B is a tree, then all consensus are dictatorships.
- (2) A non-dictatorship $f : A_1 \times \dots \times A_n \rightarrow B$ exists iff B is not a tree.

Since a tree is a median algebra containing no 2-dimensional hypercubes, the above result motivates for a generalization to median algebras not containing the m -dimensional hypercube as a subalgebra and so avoiding dictatorship.

4.2 Consensus Over Hypercube-Free Median Algebras

Consider A_1, \dots, A_n, B median algebras and $f : A_1 \times \dots \times A_n \rightarrow B$ a consensus as above.

According to [11], f is essentially k -ary iff B does not contain the $k + 1$ dimensional hypercube as a subalgebra. In other terms

1. If B is a hypercube free median algebra, then all consensus are oligarchies
2. A non-oligarchy $f : A_1 \times \dots \times A_n \rightarrow B$ exists iff B is not a hypercube free median algebra.

In this framework the consensus over more than one criterion is allowed.

4.3 Consensus Over Weak Orders

Firstly, the weak order or total preorder will be defined. Let R be a binary relation over the underlying set A . The term weak order or total preorder is often used for a relation R satisfying the conditions:

- (i) R is reflexive: For all $a \in A, aRa$
- (ii) R is transitive: For all $a, b, c \in A, aRb$ and bRc imply aRc
- (iii) R is total (or complete): For all $a, b \in A$, either aRb or bRa

Since every binary relation is a set of ordered pairs of A , the notation aRb means $(a, b) \in R$. Therefore, R is a subset of A^2 . The binary relations on A are partially ordered by implications, that is $R_1 \subseteq R_2$ means that aR_1b implies aR_2b .

Weak orders are used as preference relations. aRb means that a is preferred than b , or a dominates b , or a is a better choice than b .

Also, R is a preorder. The relation $I \subseteq A^2 : aIb$ iff aRb and bRa is an equivalence relation. The order relation among the equivalence classes is a total or linear order. Relation I is called relation of indifference. We can choose either a or b .

The intersection and union of two binary relations faced as subsets of A^2 will be denoted by \cap and \cup , respectively. The intersection of two weak orders is always a transitive and reflexive relation but not always total.

For example, $A = \{a, b, c\}$, R_1 is the linear ordering $a > b > c$ and R_2 is the ordering $b > c > a$, then $R_1 \cap R_2$ is the ordering $b > c$ and a incomparable to both b and c , which is not a total relation.

The union of two weak orders is always a reflexive and total relation. However, it is not always transitive. For example, consider R_1 and R_2 as above. Then, $R_1 \cup R_2$ is not transitive since cR_2aR_1b or $c > a > b$ but the couple (c, b) is not included in $R_1 \cup R_2$. The transitive closure of $R_1 \cup R_2$ is the least weak order containing R_1 and R_2 and concerning the example it is the complete relation A^2 .

Every binary relation on A can be represented by a $n \times n$ square matrix with 0, 1 as entries, $n = |A|$, i.e. $A = (\alpha_{ij}), i, j = 1, \dots, n$ where:

$$\alpha_{ij} = \begin{cases} 1 & \text{if } iRj \\ 0 & \text{else} \end{cases}$$

In other words, the i, j -th entry is 1 iff i is related to j . For example, $|A| = 3, A = \{a, b, c\}$, the relation $b > c > a$ will be represented by the matrix

$$\begin{array}{c} a \quad b \quad c \\ a \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \\ b \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \\ c \begin{pmatrix} 1 & 0 & 1 \end{pmatrix} \end{array}$$

The relation $c < a, c < b, aIb$ will be represented by the matrix

$$\begin{array}{c} a \quad b \quad c \\ a \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \\ b \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \\ c \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \end{array}$$

The complete relation $aIbIc$ is described from the matrix

$$\begin{array}{c} a \quad b \quad c \\ a \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \\ b \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \\ c \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \end{array}$$

Since the entries of the matrices are either 0 or 1 the two elements Boolean algebra as basic set is used. The operation join (\vee) as addition and the operation meet (\wedge) as multiplication are employed. For $\mathbf{X}=(x_{ij})_{1 \leq i, j \leq n}, \mathbf{Y} = (y_{ij}), 1 \leq i, j \leq n, n \times n$ matrices their product is defined as $\mathbf{X} \cdot \mathbf{Y} = \mathbf{Z} := (z_{ij})$, where $(z_{ij}) = \bigvee_{k=1}^n (x_{ik} \wedge y_{kj})$.

The matrix derived from the binary relation R will be denoted by M_R . For $\mathbf{X} = M_R$ the conditions for R being reflexive and transitive are respectively $\mathbf{I} \leq \mathbf{X}$ and $\mathbf{X}^2 \leq \mathbf{X}$, \mathbf{I} is the identity matrix. In the case of being reflexive and transitive the condition becomes $\mathbf{X}^2 = \mathbf{X}$. Finally, the condition for R is a total relation is $\mathbf{X} \vee \mathbf{X}^T = (\mathbf{1})_{1 \leq i, j \leq n} := \mathbf{1}$. The transitive closure of a binary relation is the intersection of all binary relations containing it. It can be shown that for $\mathbf{X} = M_R$ the transitive closure of \mathbf{X} is: $\mathbf{X} \vee \mathbf{X}^2 \vee \mathbf{X}^3 \dots \vee \mathbf{X}^n$. Also, in the case \mathbf{X} reflexive that is $\mathbf{X} = \mathbf{I} \vee \mathbf{X}$ the transitive closure is \mathbf{X}^n since $\mathbf{X} \vee \mathbf{X}^2 = \mathbf{X}(\mathbf{X} \vee \mathbf{I}) = \mathbf{X} \cdot \mathbf{X} = \mathbf{X}^2$, etc.

The set of all weak orders (or total preorders) over A is denoted by $W(A)$ [20, 21]. According to [20] and [21], the set of weak orders over the underlying set A , $W(A)$, is a (join) median semi-lattice. The join operation $R_1 \vee R_2$ of two weak

orders R_1 and R_2 is the transitive closure of $R_1 \cup R_2$. If we use matrix notations $\mathbf{X}_1 = M_{R_1}$, $\mathbf{X}_2 = M_{R_2}$, then the join $R_1 \vee R_2$ can be represented as

$$M_{R_1 \vee R_2} = (\mathbf{X}_1 \vee \mathbf{X}_2)^n$$

Also, the meet of R_1 and R_2 is given as $M_{R_1 \wedge R_2} = \mathbf{X}_1 \wedge \mathbf{X}_2$ The meet operation is a “partial” operation. It is defined only iff $\mathbf{X}_1 \wedge \mathbf{X}_2$ correspond to a total relation.

Operations \vee and \wedge over $W(A)$ satisfy:

- (i) Every principal filter $\{S \in W(A) : S \geq R\} = F(R)$ is a distributive lattice
- (ii) If every pair of the triple $\{R, S, T\}$ of $W(A)$ has a lower bound, then R, S, T have a common lower bound in $W(A)$ (Join median semi-lattice)

Also, $W(A)$ is characterized as a semi-Boolean algebra in the sense that every principal filter $F(R)$ is a Boolean algebra [21].

Since $W(A)$ is a median semi-lattice it is also a median algebra. In the case $|A| = n$, it is easy to see that $W(A)$ contains the $n - 1$ -dimensional hypercube as a subalgebra but it does not contain the n -dimensional hypercube as a subalgebra.

The results provided in Sections 4.2 and 4.3 could be used as consensus rules for multicriteria decision making as the next section presents.

5 Application

In this section, an example is presented for the use of the above-mentioned median structures in choosing a leader for a company. There are 3 competing candidates namely A, B, C and 4 competing criteria, namely: Experience, Education, Harisma, and Age. The order of candidates according to the above-mentioned criteria are: Experience: $B > A > C$, Education: $C > A > B$, Harisma: $A > B > C$, and Age: $B > A > C$. By using median computations of the four criteria the final ordering (and so the suitable leader) will be derived. Since the number of criteria is even, we get the interval $[R_1, R_2]$ where R_1 is the linear ordering: $C < A < B$ and R_2 is the weak ordering: $C < A, C < B, A I B$. Since the interval contains only its end points the decision-maker should reject candidate C and choose either B or $A I B$, so it is reasonable to choose B since B is the candidate appeared in both alternative solutions. Notice that the example is motivated by an example showing the use of the analytic hierarchy process (AHP) (https://en.wikipedia.org/wiki/Analytic_hierarchy_process_-_leader_example), and the solution provided by the proposed technique is the same as with AHP. However, the it should be mentioned the computational simplicity of the proposed technique as the interval scale and the eigenvectors calculations required by AHP are avoided.

6 Conclusions

In this study alternative consensus rules based on algebraic tools are presented in order to make multicriteria decisions. These techniques require only order scale pairwise comparisons (Boolean operations) and avoid numerical ones as for example the well know AHP technique. An example is presented leading to a final decision that coincides with that obtained through AHP process.

References

1. J. Aguarón , M.T. Escobar, J.M. Moreno-Jiménez, The precise consistency consensus matrix in a local AHP-group decision making context. *Ann. Oper. Res.* **34**, 245–259 (2016)
2. K.J. Arrow, A difficulty in the concept of social welfare. *J. Polit. Econ.* **58**, 328–346 (1950)
3. K.J. Arrow, *Individual Values and Social Choice* (Wiley, New York, 1951)
4. K.J. Arrow, *Social Choice and Individual Values*, 2nd edn. (Yale University Press, New Haven, 1963)
5. H.J. Bandelt, J. Hedlikova, Median algebras. *Discret. Math.* **45**, 1–30 (1983)
6. H.J. Bandelt, G.C. Meletiou, The zero completion of a median algebra. *Czech. J. Math.* **43**(118), 409–417 (1993)
7. H.J. Bandelt, M.F. Janowitz, G.C. Meletiou, n -median semilattices. *Order* **8**, 185–195 (1991)
8. J.C. Bezdek, B. Spillman, R. Spillman, A fuzzy relation space for group decision theory. *Fuzzy Sets Syst.* **245**, 255–268 (1978)
9. G. Birkhoff, S.A. Kiss, A ternary operation in distributive lattices. *Bull. Amer. Math. Soc.* **53**, 749–752 (1947)
10. A.K. Choudhury, R. Shankar, M.K. Tiwari, Consensus-based intelligent group decision-making model for the selection of advanced technology. *Decision Supp. Syst.* **42**, 1776–1799 (2006)
11. M. Couceiro, G.C. Meletiou, On the number of essential arguments of homomorphisms between products of median algebras. *Algebra Univers.* **79**, 85 (2018)
12. M. Couceiro, J.-L. Marichal, B. Teheux, Conservative median algebras and semilattices. *Order* **33**, 121–132 (2016)
13. M. Couceiro, S. Foldes, G.C. Meletiou, On homomorphisms between products of median algebras. *Algebra Univers.* **78**, 545–553 (2017)
14. B. Erdamar, J.L. García-Lapresta, D. Pérez-Román, M. Remzi Sanver, A fuzzy relation space for group decision theory. *Inf. Fusion* **17**, 14–21 (2014)
15. M. Franssen, Arrow’s theorem, multi-criteria decision problems and multi-attribute preferences in engineering design. *Res. Eng. Design* (16), 42–56 (2005)
16. T. González-Arteaga, R. De Andrés Calle, F. Chiclana, A new measure of consensus with fuzzy preference relations: the correlation consensus degree. *Knowl.-Based Syst.* **107**, 104–116 (2016)
17. A.A. Grau, Ternary Boolean algebra. *Bull. Amer. Math. Soc.* **53**, 567–572 (1947)
18. E. Herrera-Viedma, F.J. Cabrerizo, J. Kacprzyk, W. Pedrycz, A review of soft consensus models in a fuzzy environment. *Inf. Fusion* **17**, 4–13 (2014)
19. J.R. Isbell, Median algebra. *Trans. Am. Math. Soc.* **260**, 319–362 (1980)
20. M.F. Janowitz, On the semilattice of weak orders of a set. *Math. Soc. Sci.* **8**, 229–239 (1984)
21. M.F. Janowitz, Induced social welfare functions. *Math. Soc. Sci.* **15**, 261–276 (1988)
22. A. Kozierkiewicz-Hetmanska, The analysis of expert opinions’ consensus quality. *Episteme* **34**, 80–86 (2017)

23. Y. Liu, Y. Dong, H. Liang, F. Chiclana, E. Herrera-Viedma, Multiple attribute strategic weight manipulation with minimum cost in a group decision making context with interval attribute weights information. *IEEE Trans. Syst. Man Cyb.* **49**, 1981–1992 (2019)
24. J.M. Moreno-Jiménez, M. Salvador, P. Gargallo, A. Altuzarra, Systemic decision making A Bayesian approach in AHP. *Ann. Oper. Res.* **245**, 261–284 (2016)
25. M. Sholander, Trees, lattices, order, and betweenness. *Proc. Am. Math. Soc.* **3**, 369–381 (1952)
26. M. Sholander, Medians, lattices, and trees. *Proc. Am. Math. Soc.* **5**, 808–812 (1954)
27. J. Wallenius, J.S. Dyer, P.C. Fishburn, R.E. Steuer, S. Zionts, K. Deb, Multiple criteria decision making, multiattribute utility theory: Recent accomplishments and what lies ahead. *Manag. Sci.* **54**, 1336–1349 (2008)
28. J. Von Neumann, O. Morgenstern, *Theory of Games and Economic Behavior*, 2nd rev. edn. (Princeton University Press, Princeton, 1947)
29. H. Zhang, Y. Dong, F. Chiclana, S. Yu, Consensus efficiency in group decision making: a comprehensive comparative study and its optimal design. *Eur. J. Oper. Res.* **275**, 580–598 (2019)

Norm Estimates for the Composite Operators



Shusen Ding, Guannan Shi, and Yuming Xing

Abstract In this paper, we obtain both local and global L^p norm inequalities and imbedding inequalities for the composition of the homotopy operator, differential operator, and Green's operator applied to differential forms. These inequalities can be used to study the integrability of the composition of the operators.

1 Introduction

The purpose of this paper is to obtain norm estimates for the composition $T \circ d \circ G$ of the homotopy operator T , differential operator d , and Green's operator G applied to differential forms. These three operators are key operators in some areas of mathematics, such as analysis and partial differential equations. We all know that any differential form u can be expressed as $u = Td(u) + dT(u)$. Thus, $G(u)$ can be decomposed as $G(u) = TdG(u) + dTG(u)$. However, $dTG(u) = (G(u))_B$ holds for any differential form u defined in a ball $B \in \mathbb{R}^n$, and $(G(u))_B$ is a closed form that has received much investigation in recent years, see [1, 5–7]. Thus, we are motivated to study the composition $TdG(u)$ in this paper. We prove both local and global L^p norm inequalities and imbedding inequalities for the composition $T \circ d \circ G$. Specifically, we find that $T \circ d \circ G(u)$ has higher integral exponent than that of u for a differential form u satisfying certain conditions. Our results will provide efficient ways to estimate the norm of $T \circ d \circ G(u)$ in terms of the norm of u or du .

S. Ding (✉)

Department of Mathematics, Seattle University, Seattle, WA, USA

e-mail: sding@seattleu.edu

G. Shi

School of Mathematics and Statistics, Northeast Petroleum University, Daqing, R. P. China

Y. Xing

Department of Mathematics, Harbin Institute of Technology, Harbin, P. R. China.

e-mail: xyuming@hit.edu.cn

© Springer Nature Switzerland AG 2021

T. M. Rassias, P. M. Pardalos (eds.), *Nonlinear Analysis and Global Optimization*,

Springer Optimization and Its Applications 167,

https://doi.org/10.1007/978-3-030-61732-5_8

As extensions of functions, differential forms have been well studied and used in recent years, see [1, 2, 7]. We continue to use the traditional notations appearing in [1] throughout this paper. Particularly, let Ω denote a domain and B denote a ball in \mathbb{R}^n , $n \geq 2$. Assume that $\wedge^l = \wedge^l(\mathbb{R}^n)$ is the set of all l -forms in \mathbb{R}^n , and

$$\wedge = \wedge(\mathbb{R}^n) = \bigoplus_{l=0}^n \wedge^l(\mathbb{R}^n)$$

is a graded algebra with respect to the exterior products. Let $D'(\Omega, \wedge^l)$ be the space of all differential l -forms in Ω , and $L^p(\Omega, \wedge^l)$ be the space of all l -forms $u(x) = \sum_I u_I(x) dx_I$ in Ω satisfying $\int_{\Omega} |u_I|^p < \infty$ for all ordered l -tuples I , $l = 1, 2, \dots, n$. We will use $d : D'(\Omega, \wedge^l) \rightarrow D'(\Omega, \wedge^{l+1})$, $l = 0, 1, \dots, n - 1$, to denote the exterior derivative. The Hodge star operator $\star : \wedge^k \rightarrow \wedge^{n-k}$ is defined as follows. If

$$\omega = \omega_{i_1 i_2 \dots i_k}(x_1, x_2, \dots, x_n) dx_{i_1} \wedge dx_{i_2} \wedge \dots \wedge dx_{i_k} = \omega_I dx_I, \quad i_1 < i_2 < \dots < i_k,$$

is a differential k -form, then

$$\star \omega = \star(\omega_{i_1 i_2 \dots i_k} dx_{i_1} \wedge dx_{i_2} \wedge \dots \wedge dx_{i_k}) = (-1)^{\sum(I)} \omega_I dx_J,$$

where $I = (i_1, i_2, \dots, i_k)$, $J = \{1, 2, \dots, n\} - I$, and

$$\sum(I) = \frac{k(k+1)}{2} + \sum_{j=1}^k i_j.$$

The Hodge codifferential operator $d^* : D'(\Omega, \wedge^{l+1}) \rightarrow D'(\Omega, \wedge^l)$ is defined by $d^* = (-1)^{nl+1} \star d \star$ on $D'(\Omega, \wedge^{l+1})$, $l = 0, 1, \dots, n - 1$.

For $u \in D'(\Omega, \wedge^l)$, the vector-valued differential form

$$\nabla u = \left(\frac{\partial u}{\partial x_1}, \dots, \frac{\partial u}{\partial x_n} \right)$$

consists of differential forms $\frac{\partial u}{\partial x_i} \in D'(\Omega, \wedge^l)$, where the partial differentiation is applied to the coefficients of u . Let $|E|$ denote the n -dimensional Lebesgue measure of a set $E \subseteq \mathbb{R}^n$. For a function u , the average of u over B is defined by $u_B = \frac{1}{|B|} \int_B u dx$. All integrals involved in this paper are the Lebesgue integrals. Let $C^\infty(\Omega, \wedge^l)$ be the space of smooth l -forms on Ω and the Green's operator G be defined on $C^\infty(\Omega, \wedge^l)$ by assigning $G(u)$ to be a solution of the Poisson's equation

$$\Delta G(u) = u - H(u),$$

where H is the harmonic projection operator, see [1] and [7] for more results about Green’s operator. For any subset $E \subset \mathbb{R}^n$ and $p > 1$, we use $W^{1,p}(E, \wedge^l)$ to denote the Sobolev space of l -forms, which equals $L^p(E, \wedge^l) \cap L^p_1(E, \wedge^l)$ with norm

$$\|u\|_{W^{1,p}(E)} = \|u\|_{W^{1,p}(E, \wedge^l)} = \text{diam}(E)^{-1} \|u\|_{p,E} + \|\nabla u\|_{p,E}. \tag{1}$$

A homotopy operator $T : C^\infty(\Omega, \wedge^l) \rightarrow C^\infty(\Omega, \wedge^{l-1})$ is defined by averaging K_y over all points $y \in \Omega$:

$$Tu = \int_\Omega \phi(y) K_y u dy,$$

where $\phi \in C^\infty_0(\Omega)$ is normalized so that $\int \phi(y) dy = 1$, and the linear operator $K_y : C^\infty(\Omega, \wedge^l) \rightarrow C^\infty(\Omega, \wedge^{l-1})$ is defined by

$$(K_y u)(x; \xi_1, \dots, \xi_{l-1}) = \int_0^1 t^{l-1} u(tx + y - ty; x - y, \xi_1, \dots, \xi_{l-1}) dt.$$

See [1] and [3] for more details for the homotopy operator. It is also known that for each differential form u , we have the decomposition

$$u = d(Tu) + T(du). \tag{2}$$

$$\|\nabla(Tu)\|_{p,\Omega} \leq C|\Omega| \|u\|_{p,\Omega}, \text{ and } \|Tu\|_{p,\Omega} \leq C|\Omega| \text{diam}(\Omega) \|u\|_{p,\Omega}. \tag{3}$$

We know that any open subset Ω in \mathbb{R}^n is the union of a sequence of cubes Q_k , whose sides are parallel to the axes, whose interiors are mutually disjoint, and whose diameters are approximately proportional to their distances from F , where F is the complement of Ω in \mathbb{R}^n . Specifically, (1) $\Omega = \cup_{k=1}^\infty Q_k$, (2) $Q_j^0 \cap Q_k^0 = \emptyset$ if $j \neq k$, (3) there exist two constants $c_1, c_2 > 0$ (we can take $c_1 = 1$, and $c_2 = 4$), so that $c_1 \text{diam}(Q_k) \leq \text{distance}(Q_k, F) \leq c_2 \text{diam}(Q_k)$. Hence, the definition of the homotopy operator T can be generalized to any domain Ω in \mathbb{R}^n : For any $x \in \Omega$, $x \in Q_k$ for some k . Let T_{Q_k} be the homotopy operator defined on Q_k (each cube is bounded and convex). Thus, we can define the homotopy operator T_Ω on any domain Ω by $T_\Omega = \sum_{k=1}^\infty T_{Q_k} \chi_{Q_k}(x)$.

The nonlinear partial differential equation for differential forms

$$d^* A(x, du) = B(x, du) \tag{4}$$

is called non-homogeneous A -harmonic equation, where $A : \Omega \times \wedge^l(\mathbb{R}^n) \rightarrow \wedge^l(\mathbb{R}^n)$ and $B : \Omega \times \wedge^l(\mathbb{R}^n) \rightarrow \wedge^{l-1}(\mathbb{R}^n)$ satisfy the conditions:

$$|A(x, \xi)| \leq a|\xi|^{p-1}, \quad A(x, \xi) \cdot \xi \geq |\xi|^p \quad \text{and} \quad |B(x, \xi)| \leq b|\xi|^{p-1} \tag{5}$$

for almost every $x \in \Omega$ and all $\xi \in \wedge^l(\mathbb{R}^n)$. Here, $a, b > 0$ are constants and $1 < p < \infty$ is a fixed exponent associated with (4). A solution to (4) is an element u of the Sobolev space $W_{loc}^{1,p}(\Omega, \wedge^{l-1})$ such that

$$\int_{\Omega} A(x, du) \cdot d\varphi + B(x, du) \cdot \varphi = 0 \tag{6}$$

for all $\varphi \in W_{loc}^{1,p}(\Omega, \wedge^{l-1})$ with compact support. If u is a function (0-form) in \mathbb{R}^n , Equation (4) reduces to

$$\operatorname{div} Ax, \nabla u = B(x, \nabla u). \tag{7}$$

If the operator $B = 0$, Equation (4) becomes

$$d^*A(x, du) = 0, \tag{8}$$

which is called the (homogeneous) A -harmonic equation. Let $A : \Omega \times \wedge^l(\mathbb{R}^n) \rightarrow \wedge^l(\mathbb{R}^n)$ be defined by $A(x, \xi) = \xi|\xi|^{p-2}$ with $p > 1$. Then, A satisfies the required conditions, and (8) becomes the p -harmonic equation $d^*(du|du|^{p-2}) = 0$ for differential forms.

2 Local Integrability

We first prove the local norm inequalities of the composite operator $T \circ d \circ G$ in this section. The following Lemma 1 (Weak Reverse Hölder’s Inequality) appeared in [4].

Lemma 1 *Let u be a solution of the A -harmonic equation (4) in a domain Ω and $0 < s, t < \infty$. Then, there exists a constant C , independent of u , such that*

$$\|u\|_{s,B} \leq C|B|^{(t-s)/st} \|u\|_{t,\sigma B} \tag{9}$$

for all balls B with $\sigma B \subset \Omega$ for some $\sigma > 1$.

The following Lemma 2 appeared in [3].

Lemma 2 *Let $u \in D^l(Q, \wedge^l)$ and $du \in L^p(Q, \wedge^{l+1})$. Then $u - u_Q$ is in $L^{np/(n-p)}(Q, \wedge^l)$ and*

$$\left(\int_Q |u - u_Q|^{np/(n-p)} dx \right)^{(n-p)/np} \leq C_p(n) \left(\int_Q |du|^p dx \right)^{1/p} \tag{10}$$

for Q a cube or a ball in \mathbb{R}^n , $l = 0, 1, \dots, n - 1$, and $1 < p < n$.

The following Lemma 3 appeared in [5].

Lemma 3 *Let u be a smooth differential form defined in Ω and $1 < s < \infty$. Then, there exists a positive constant $C = C(s)$, independent of u , such that*

$$\begin{aligned} & \|dd^*G(u)\|_{s,B} + \|d^*dG(u)\|_{s,B} + \|dG(u)\|_{s,B} \\ & + \|d^*G(u)\|_{s,B} + \|G(u)\|_{s,B} \leq C(s)\|u\|_{s,B} \end{aligned} \tag{11}$$

for any ball $B \subset \Omega$.

Now, we prove the following local norm inequality, which will be used to prove the global inequality in the next section.

Theorem 1 *Assume that $u \in D^l(\Omega, \wedge^l)$, $l = 1, 2, \dots, n$, $1 < p < n$, and T is the homotopy operator and G is Green's operator. If $u \in L^p_{loc}(\Omega, \wedge^l)$, then $TdG(u) \in L^s_{loc}(\Omega, \wedge^l)$ for any $0 < s < np/(n - p)$. Moreover, there exists a constant C , independent of u , such that*

$$\|TdG(u)\|_{s,B} \leq C|B|^{1+1/s+1/n-1/p}\|u\|_{p,\sigma B} \tag{12}$$

for any ball B with $\sigma B \subset \Omega$ for some $\sigma > 1$.

Proof Note that for any constant $p > 1$ and any differential form u , we have

$$d(TdG(u)) = (dG(u))_B$$

and

$$\|(dG(u))_B\|_{p,B} \leq C_1\|dG(u)\|_{p,B} \leq C_2\|u\|_{p,B}, \tag{13}$$

where C_1 and C_2 are the constants. Applying Lemma 2 to $TdG(u)$ and then using (3) and Lemma 3, we have

$$\begin{aligned} & \|TdG(u) - (TdG(u))_B\|_{np/(n-p),B} \\ & = \left(\int_B |TdG(u) - (TdG(u))_B|^{np/(n-p)} dx\right)^{(n-p)/np} \\ & \leq C_3 \left(\int_B |d(TdG(u))|^p dx\right)^{1/p} \\ & \leq C_4 \left(\int_B |\nabla(TdG(u))|^p dx\right)^{1/p} \\ & = C_5|B| \left(\int_B |dG(u)|^p dx\right)^{1/p} \\ & \leq C_6|B| \left(\int_B |u|^p dx\right)^{1/p}. \end{aligned} \tag{14}$$

We know that $(TdG(u))_B$ is a closed form, so it satisfies the Weak Reverse Hölder Inequality, that is,

$$\begin{aligned}
 \|(T dG(u))_B\|_{np/(n-p), B} &\leq C_7|B|^{-1/n}\|(T dG(u))_B\|_{p, \sigma B} \\
 &\leq C_8|B|^{-1/n}\|T dG(u)\|_{p, \sigma B} \\
 &\leq C_9|B|^{-1/n}|B|^{1+1/n}\|dG(u)\|_{p, \sigma B} \\
 &\leq C_9|B|\|dG(u)\|_{p, \sigma B} \\
 &\leq C_{10}|B|\|u\|_{p, \sigma B},
 \end{aligned} \tag{15}$$

where $\sigma > 1$ is a constant. Using Minkowski’s inequality, (14), and (15), we find that

$$\begin{aligned}
 &\|T dG(u)\|_{np/(n-p), B} \\
 &\leq \|T dG(u) - (T dG(u))_B\|_{np/(n-p), B} + \|(T dG(u))_B\|_{np/(n-p), B} \\
 &\leq C_6|B|\|u\|_{p, B} + C_{10}|B|\|u\|_{p, \sigma B} \\
 &\leq C_{11}|B|\|u\|_{p, \sigma B}.
 \end{aligned} \tag{16}$$

By the monotonic property of the L^p -space, for any s with $0 < s < np/(n - p)$, we obtain

$$\left(\frac{1}{|B|} \int_B |T dG(u)|^s dx\right)^{1/s} \leq \left(\frac{1}{|B|} \int_B |T dG(u)|^{np/(n-p)} dx\right)^{(n-p)/np},$$

that is, it follows that

$$\|T dG(u)\|_{s, B} \leq |B|^{1/s-1/p+1/n}\|T dG(u)\|_{np/(n-p), B}. \tag{17}$$

Combining (16) and (17), we obtain

$$\|T dG(u)\|_{s, B} \leq C_{11}|B|^{1+1/s-1/p+1/n}\|u\|_{p, \sigma B}.$$

We have completed the proof of Theorem 1. □

We should notice that the above inequality (12) can be written as the following symmetric version:

$$\left(\frac{1}{|B|} \int_B |T dG(u)|^s dx\right)^{1/s} \leq C|B|^{1+1/n} \left(\frac{1}{|B|} \int_{\sigma B} |u|^p dx\right)^{1/p}. \tag{12'}$$

Also, from Theorem 1 and Minkowski’s inequality, it follows that

$$\begin{aligned}
 &\|T dG(u) - (T dG(u))_B\|_{s, B} \\
 &\leq \|T dG(u)\|_{s, B} + \|(T dG(u))_B\|_{s, B} \\
 &\leq \|T dG(u)\|_{s, B} + C_1\|T dG(u)\|_{s, B} \\
 &\leq C_2|B|^{1+1/s+1/n-1/p}\|u\|_{p, \sigma B}.
 \end{aligned}$$

Therefore, we obtain the following corollary about the integrability of $TdG(u) - (TdG(u))_B$.

Corollary 1 *Assume that $u \in D'(\Omega, \wedge^l)$, $l = 1, 2, \dots, n$, $1 < p < n$, and T is the homotopy operator and G is Green's operator. If $u \in L^p_{loc}(\Omega, \wedge^l)$, then $TdG(u) - (TdG(u))_B \in L^s_{loc}(\Omega, \wedge^l)$ for any $0 < s < np/(n - p)$. Moreover, there exists a constant C , independent of u , such that*

$$\|TdG(u) - (TdG(u))_B\|_{s,B} \leq C|B|^{1+1/s+1/n-1/p} \|u\|_{p,\sigma B} \tag{18}$$

for any ball B with $\sigma B \subset \Omega$ for some $\sigma > 1$.

Similarly, we can write (18) in the following symmetric version:

$$\left(\frac{1}{|B|} \int_B |TdG(u) - (TdG(u))_B|^s dx\right)^{1/s} \leq C|B|^{1+1/n} \left(\frac{1}{|B|} \int_{\sigma B} |u|^p dx\right)^{1/p} \tag{18'}$$

for all balls B with $\sigma B \subset \Omega$ for some $\sigma > 1$.

Note that in (12) and (18), the integral exponent s on the left side may be much larger than the exponent p on the right side since $np/(n - p) \rightarrow \infty$ as $p \rightarrow n$, which gives the higher integrability of the composite operator $T \circ d \circ G$ for the case $1 < p < n$. Next, we prove the higher integrability of $T \circ d \circ G$ for the case $p \geq n$.

Theorem 2 *Let $u \in D'(\Omega, \wedge^l)$, $l = 1, 2, \dots, n$, $p \geq n$, and T be the homotopy operator and G be Green's operator. If $u \in L^p_{loc}(\Omega, \wedge^l)$, then $TdG(u) \in L^s_{loc}(\Omega, \wedge^l)$ for any $s > 0$. Moreover, there exists a constant C , independent of u , such that*

$$\|TdG(u)\|_{s,B} \leq C|B|^{1+1/s+1/n-1/p} \|u\|_{p,\sigma B} \tag{19}$$

for any ball B with $\sigma B \subset \Omega$ for some $\sigma > 1$.

Proof Let $k = \max\{1, s/p\}$ and choose $q = knp/(n + kp)$. Since $n - p \leq 0$, we have

$$q - p = \frac{p((k(n - p) - n))}{n + kp} < 0, \tag{20}$$

which means $q < p$. We also notice that $1 < q = knp/(n + kp) < n$. Applying Lemma 2 to $TdG(u)$, and then using (3) and the monotonic property of the L^p -space, we find that

$$\begin{aligned}
& \left(\int_B |TdG(u) - (TdG(u))_B|^{nq/(n-q)} dx \right)^{(n-q)/nq} \\
& \leq C_1 \left(\int_B |d(TdG(u))|^q dx \right)^{1/q} \\
& \leq C_1 \left(\int_B |\nabla(TdG(u))|^q dx \right)^{1/q} \\
& \leq C_2 |B| \|dG(u)\|_{q,B} \\
& \leq C_3 |B| \|u\|_{q,B} \\
& \leq C_3 |B|^{1+1/q-1/p} \|u\|_{p,B}.
\end{aligned} \tag{21}$$

The remaining part of the proof of Theorem 2 is similar to that of Theorem 1. For the completing purpose, we continue the proof as follows. We know that $(TdG(u))_B$ is a closed form, so it satisfies the Weak Reverse Hölder Inequality, that is

$$\begin{aligned}
& \|(TdG(u))_B\|_{nq/(n-q),B} \\
& \leq C_4 |B|^{-1/n} \|(TdG(u))_B\|_{q,\sigma B} \\
& \leq C_5 |B|^{-1/n} \|TdG(u)\|_{q,\sigma B} \\
& \leq C_6 |B|^{-1/n} |B|^{1+1/n} \|dG(u)\|_{q,\sigma B} \\
& \leq C_6 |B| \|dG(u)\|_{q,\sigma B} \\
& \leq C_7 |B| \|u\|_{q,\sigma B} \\
& \leq C_7 |B|^{1+1/q-1/p} \|u\|_{p,\sigma B},
\end{aligned} \tag{22}$$

where $\sigma > 1$ is a constant. Using Minkowski's inequality again, (21), and (22), we have

$$\begin{aligned}
& \|TdG(u)\|_{nq/(n-q),B} \\
& \leq \|TdG(u) - (TdG(u))_B\|_{nq/(n-q),B} + \|(TdG(u))_B\|_{nq/(n-q),B} \\
& \leq C_3 |B|^{1+1/q-1/p} \|u\|_{p,B} + C_7 |B|^{1+1/q-1/p} \|u\|_{p,\sigma B} \\
& \leq C_3 |B|^{1+1/q-1/p} \|u\|_{p,\sigma B} + C_7 |B|^{1+1/q-1/p} \|u\|_{p,\sigma B} \\
& \leq C_8 |B|^{1+1/q-1/p} \|u\|_{p,\sigma B}.
\end{aligned} \tag{23}$$

From the choice of k , we know that $nq/(n-q) = kp > s$, using the monotonic property of the L^p -space again and (23),

$$\begin{aligned}
\|TdG(u)\|_{s,B} & \leq |B|^{1/s+1/n-1/q} \|TdG(u)\|_{nq/(n-q),B} \\
& \leq C_8 |B|^{1+1/s+1/n-1/p} \|u\|_{p,\sigma B},
\end{aligned}$$

which is equivalent to (19). We have completed the proof of Theorem 2. \square

By the same method, we can obtain the local higher integrability for Green's operator G and the composition of the homotopy operator T , differential operator d , and projection operator H for the case $1 < p < n$ or $p \geq n$, respectively. That is, using the similar methods, we can prove the following Theorems 3 and 4.

Theorem 3 *Let $u \in D'(\Omega, \wedge^l)$ be a differential form, $l = 1, 2, \dots, n$, $1 < p < n$, T be the homotopy operator, G be Green's operator, and H be the projection operator. If $du \in L^p_{loc}(\Omega, \wedge^l)$, then $TdH(u) \in L^s_{loc}(\Omega, \wedge^l)$ for any $0 < s <$*

$np/(n - p)$. Moreover, there exist constants C_1 and C_2 , independent of u , such that

$$\|TdH(u)\|_{s,B} \leq C_1|B|^{1+1/s+1/n-1/p} \|du\|_{p,\sigma B} \tag{24}$$

and

$$\|G(u)\|_{s,B} \leq C_1|B|^{1/s+1/n-1/p} \|u\|_{p,\sigma B} \tag{25}$$

for any ball B with $\sigma B \subset \Omega$ for some $\sigma > 1$.

Theorem 4 Let $u \in D'(\Omega, \wedge^l)$ be a differential form, $l = 1, 2, \dots, n$, $p \geq n$, and T be the homotopy operator and H be the projection operator. If $du \in L^p_{loc}(\Omega, \wedge^l)$, then $TdH(u) \in L^s_{loc}(\Omega, \wedge^l)$ for any $s > 0$. Moreover, there exist constants C_1 and C_2 , independent of u , such that

$$\|TdH(u)\|_{s,B} \leq C_1|B|^{1+1/s+1/n-1/p} \|du\|_{p,\sigma B} \tag{26}$$

and

$$\|G(u)\|_{s,B} \leq C_1|B|^{1/s+1/n-1/p} \|u\|_{p,\sigma B} \tag{27}$$

for all balls B with $\sigma B \subset \Omega$ for some $\sigma > 1$.

Now, we are ready to prove the higher order imbedding inequality for the composition $T \circ d \circ G$ in the case $1 < p < n$ and the case $p \geq n$ in the following Theorems 5 and 6, respectively.

Theorem 5 Let $u \in D'(\Omega, \wedge^l)$ be a differential form, $l = 1, 2, \dots, n$, $1 < p < n$, and T be the homotopy operator and G be Green's operator. If $du \in L^p_{loc}(\Omega, \wedge^l)$, then $TdG(u) \in W^{1,s}_{loc}(\Omega, \wedge^l)$ for any $0 < s < np/(n - p)$. Specifically, there exists a constant C , independent of u , such that

$$\|TdG(u)\|_{W^{1,s}(B)} \leq C|B|^{1+1/s+1/n-1/p} \|du\|_{p,\sigma B} \tag{28}$$

and

$$\|TdG(u) - (TdG(u))_B\|_{W^{1,s}(B)} \leq C|B|^{1+1/s+1/n-1/p} \|du\|_{p,\sigma B} \tag{29}$$

for any ball B with $\sigma B \subset \Omega$ for some $\sigma > 1$.

Proof By Definitions (1) and (25), we have

$$\begin{aligned}
& \|TdG(u)\|_{W^{1,s}(B)} \\
&= \text{diam}(B)^{-1} \|TdG(u)\|_{s,B} + \|\nabla TdG(u)\|_{s,B} \\
&\leq \text{diam}(B)^{-1} C_1 |B|^{1+1/n} \|dG(u)\|_{s,B} + C_2 |B| \|dG(u)\|_{s,B} \\
&= \text{diam}(B)^{-1} C_1 |B|^{1+1/n} \|G(du)\|_{s,B} + C_2 |B| \|G(du)\|_{s,B} \\
&\leq C_3 |B| \|G(du)\|_{s,B} \\
&\leq C_4 |B|^{1+1/s+1/n-1/p} \|du\|_{p,\sigma B},
\end{aligned} \tag{30}$$

which indicates that (28) holds. Next, we prove that the inequality (29) is also true. Using (1), (3), and (27) and the fact that $dTdG(u) = (dG(u))_B$, we obtain

$$\begin{aligned}
& \|TdG(u) - (TdG(u))_B\|_{W^{1,s}(B)} \\
&\leq \|TdTdG(u)\|_{W^{1,s}(B)} \\
&= (\text{diam}(B))^{-1} \|TdTdG(u)\|_{s,B} + \|\nabla TdTdG(u)\|_{s,B} \\
&\leq (\text{diam}(B))^{-1} C_5 |B|^{1+1/n} \|dTdG(u)\|_{s,B} + C_6 |B| \|dTdG(u)\|_{s,B} \\
&\leq C_7 |B| \|dTdG(u)\|_{s,B} \\
&= C_7 |B| \|(dG(u))_B\|_{s,B} \\
&\leq C_8 |B| \|dG(u)\|_{s,B} \\
&\leq C_8 |B| \|G(du)\|_{s,B} \\
&\leq C_9 |B| |B|^{1/s+1/n-1/p} \|du\|_{p,\sigma B} \\
&\leq C_{10} |B|^{1+1/s+1/n-1/p} \|du\|_{p,\sigma B}.
\end{aligned} \tag{31}$$

Thus, (29) also holds. The proof of Theorem 5 has been completed. \square

By the same method as we developed in the proof of Theorem 5, we obtain the following higher order imbedding for the case $p \geq n$.

Theorem 6 *Let $u \in D'(\Omega, \wedge^l)$ be a differential form, $l = 1, 2, \dots, n$, $p \geq n$, and T be the homotopy operator and G be Green's operator. If $du \in L_{loc}^p(\Omega, \wedge^l)$, then $TdG(u) \in W_{loc}^{1,s}(\Omega, \wedge^l)$ for any $s > 0$. Moreover, there exists a constant C , independent of u , such that*

$$\|TdG(u)\|_{W^{1,s}(B)} \leq C |B|^{1+1/s+1/n-1/p} \|du\|_{p,\sigma B} \tag{32}$$

and

$$\|TdG(u) - (TdG(u))_B\|_{W^{1,s}(B)} \leq C |B|^{1+1/s+1/n-1/p} \|du\|_{p,\sigma B} \tag{33}$$

for any ball B with $\sigma B \subset \Omega$ for some $\sigma > 1$.

3 Global Integrability

In this section, we prove the global higher integrability and the higher order imbedding inequalities for the composition $T \circ d \circ G$. We need the following Covering Lemma.

Lemma 4 *Each domain Ω has a modified Whitney cover of cubes $\mathcal{V} = \{Q_i\}$ such that*

$$\cup_i Q_i = \Omega, \quad \sum_{Q_i \in \mathcal{V}} \chi_{\sqrt{\frac{5}{4}}Q_i} \leq N \chi_\Omega, \tag{34}$$

and some $N > 1$, and if $Q_i \cap Q_j \neq \emptyset$, then there exists a cube R (this cube need not be a member of \mathcal{V}) in $Q_i \cap Q_j$ such that $Q_i \cup Q_j \subset NR$. Moreover, if Ω is δ -John, then there is a distinguished cube $Q_0 \in \mathcal{V}$, which can be connected with every cube $Q \in \mathcal{V}$ by a chain of cubes $Q_0, Q_1, \dots, Q_k = Q$ from \mathcal{V} and such that $Q \subset \rho Q_i$, $i = 0, 1, 2, \dots, k$, for some $\rho = \rho(n, \delta)$.

We first prove the following global L^p norm inequality that indicates that $TdG(u)$ has a higher order integrability compared with u .

Theorem 7 *Let $u \in D^l(\Omega, \wedge^l)$ be a differential form, $l = 1, 2, \dots, n$, $1 < p < n$ and T be the homotopy operator and G be Green's operator. If $u \in L^p(\Omega, \wedge^l)$, then $TdG(u) \in L^s(\Omega, \wedge^l)$ for any $0 < s < np/(n - p)$. Moreover, there exists a constant C , independent of u , such that*

$$\|TdG(u)\|_{s,\Omega} \leq C|\Omega|^{1+1/s+1/n-1/p} \|u\|_{p,\Omega} \tag{35}$$

for any convex domain $\Omega \subset \mathbb{R}^n$ with $|\Omega| < \infty$.

Proof From Lemma 4 and Theorem 1 and noticing $1 + 1/s + 1/n - 1/p > 0$, we find that

$$\begin{aligned} \|TdG(u)\|_{s,\Omega} &\leq \sum_{B \in \mathcal{V}} \|TdG(u)\|_{s,B} \\ &\leq \sum_{B \in \mathcal{V}} (C_1|B|)^{1+1/s+1/n-1/p} \|u\|_{p,\sigma B} \\ &\leq \sum_{B \in \mathcal{V}} (C_1|\Omega|)^{1+1/s+1/n-1/p} \|u\|_{p,\sigma B} \\ &\leq C_2|\Omega|^{1+1/s+1/n-1/p} N \|u\|_{p,\Omega} \\ &\leq C_3|\Omega|^{1+1/s+1/n-1/p} \|u\|_{p,\Omega}. \end{aligned} \tag{36}$$

The proof of Theorem 7 has been completed. □

We already proved the global higher integrability for $T \circ d \circ G$ for the case $1 < p < n$ above. Using Theorem 2 and the same method as we developed above, we can prove the following global integrability for the case $p \geq n$.

Theorem 8 *Let $u \in D^l(\Omega, \wedge^l)$ be a differential form, $l = 1, 2, \dots, n$, $p \geq n$, and T be the homotopy operator and G be Green's operator. If $u \in L^p(\Omega, \wedge^l)$,*

then $TdG(u) \in L^s(\Omega, \wedge^l)$ for any $s > 0$, and inequality (35) holds for any convex domain $\Omega \subset \mathbb{R}^n$ with $|\Omega| < \infty$.

Theorem 9 Let $u \in D^l(\Omega, \wedge^l)$ be a differential form, $l = 1, 2, \dots, n$, $1 < p < n$, and T be the homotopy operator and G be Green's operator. If $du \in L^p(\Omega, \wedge^l)$, then $TdG(u) \in W^{1,s}(\Omega, \wedge^l)$ for any $0 < s < np/(n - p)$, and, furthermore,

$$\|TdG(u)\|_{W^{1,s}(\Omega)} \leq C|\Omega|^{1+1/s+1/n-1/p} \|du\|_{p,\Omega} \tag{37}$$

and

$$\|TdG(u) - (TdG(u))_\Omega\|_{W^{1,s}(\Omega)} \leq C|\Omega|^{1+1/s+1/n-1/p} \|du\|_{p,\Omega} \tag{38}$$

for any convex domain $\Omega \subset \mathbb{R}^n$ with $|\Omega| < \infty$.

Proof From (25) and Lemma 4 and noticing $1/s + 1/n - 1/p > 0$ since $0 < s < np/(n - p)$, we have

$$\|dG(u)\|_{s,\Omega} = \|G(du)\|_{s,\Omega} \leq C_1|\Omega|^{1/s+1/n-1/p} \|du\|_{p,\Omega} \tag{39}$$

and

$$\|G(u)\|_{s,\Omega} \leq C_2|\Omega|^{1/s+1/n-1/p} \|u\|_{p,\Omega}. \tag{40}$$

Using Definitions (1), (3), and (39), we obtain

$$\begin{aligned} & \|T(G(u))\|_{W^{1,s}(\Omega)} \\ &= (\text{diam}(\Omega))^{-1} \|TdG(u)\|_{s,\Omega} + \|\nabla TdG(u)\|_{s,\Omega} \\ &\leq (\text{diam}(\Omega))^{-1} C_3|\Omega|^{1+1/n} \|dG(u)\|_{s,\Omega} + C_4|\Omega| \|dG(u)\|_{s,\Omega} \\ &\leq C_5|\Omega| \|dG(u)\|_{s,\Omega} \\ &= C_5|\Omega| \|G(du)\|_{s,\Omega} \\ &\leq C_6|\Omega| |\Omega|^{1/s+1/n-1/p} \|du\|_{p,\Omega} \\ &\leq C_7|\Omega|^{1+1/s+1/n-1/p} \|du\|_{p,\Omega}. \end{aligned} \tag{41}$$

Thus, (37) holds. Next, we prove (38). Using (1), (3), and (39) and Lemma 3, we find that

$$\begin{aligned}
 & \|TdG(u) - (TdG(u))_\Omega\|_{W^{1,s}(\Omega)} \\
 &= \|Td(TdG(u))\|_{W^{1,s}(\Omega)} \\
 &= (\text{diam}(\Omega))^{-1} \|Td(TdG(u))\|_{s,\Omega} + \|\nabla Td(TdG(u))\|_{s,\Omega} \\
 &\leq (\text{diam}(\Omega))^{-1} C_6 |\Omega|^{1+1/n} \|dTdG(u)\|_{s,\Omega} + C_7 |\Omega| \|dTdG(u)\|_{s,\Omega} \\
 &\leq C_8 |\Omega| \|dT(dG(u))\|_{s,\Omega} \\
 &\leq C_8 |\Omega| \|(dG(u))_\Omega\|_{s,\Omega} \\
 &\leq C_9 |\Omega| \|G(du)\|_{s,\Omega} \\
 &\leq C_{10} |\Omega|^{1/s+1/n-1/p} \|du\|_{p,\Omega}.
 \end{aligned} \tag{42}$$

So, the inequality (38) is true. We have completed the proof of Theorem 9. \square

We all know that the imbedding inequality is stronger than the L^p norm inequality. From the imbedding inequality, we can prove the following L^p norm inequality with the higher integral exponent on the left hand side.

Corollary 2 *Let $u \in D'(\Omega, \wedge^l)$ be a differential form, $l = 1, 2, \dots, n$, $1 < p < n$, and T be the homotopy operator and G be Green's operator. If $du \in L^p(\Omega, \wedge^l)$, then $T(G(u)) - (T(G(u)))_\Omega \in L^s(\Omega, \wedge^l)$ for any $0 < s < np/(n - p)$ and*

$$\|TdG(u) - (TdG(u))_\Omega\|_{s,\Omega} \leq C \|du\|_{p,\Omega} \tag{43}$$

for any convex domain $\Omega \subset \mathbb{R}^n$ with $|\Omega| < \infty$.

Proof First, note that $\text{diam}(\Omega) \|\nabla(TdG(u) - (TdG(u))_\Omega)\|_{s,\Omega} \geq 0$. Applying Theorem 9 and Definition (1), we obtain

$$\begin{aligned}
 & \|TdG(u) - (TdG(u))_\Omega\|_{s,\Omega} \\
 &\leq \|TdG(u) - (TdG(u))_\Omega\|_{s,\Omega} + \text{diam}(\Omega) \|TdG(u) - (TdG(u))_\Omega\|_{s,\Omega} \\
 &= \text{diam}(\Omega) ((\text{diam}(\Omega))^{-1} \|TdG(u) - (TdG(u))_\Omega\|_{s,\Omega} \\
 &\quad + \|\nabla(TdG(u) - (TdG(u))_\Omega)\|_{s,\Omega}) \\
 &= \text{diam}(\Omega) \|TdG(u) - (TdG(u))_\Omega\|_{W^{1,s}(\Omega)} \\
 &= C_1 \text{diam}(\Omega) |\Omega|^{1+1/s+1/n-1/p} \|du\|_{p,\Omega} \\
 &\leq C_2 |\Omega|^{1+1/s+2/n-1/p} \|du\|_{p,\Omega} \\
 &\leq C_3 \|du\|_{p,\Omega}.
 \end{aligned}$$

We have completed the proof of Corollary 2. \square

Using the same method as we did in the proof of Theorem 9, we can prove the global result for the case $p \geq n$.

Theorem 10 *Let $u \in D'(\Omega, \wedge^l)$ be a differential form, $l = 1, 2, \dots, n$, $p \geq n$, and T be the homotopy operator and G be Green's operator. If $u \in L^p(\Omega, \wedge^l)$, then $T(G(u)) \in W^{1,s}(\Omega, \wedge^l)$ for any $s > 0$ and*

$$\|TdG(u)\|_{W^{1,s}(\Omega)} \leq C |\Omega|^{1+1/s+1/n-1/p} \|du\|_{p,\Omega} \tag{44}$$

and

$$\|TdG(u) - (TdG(u))_{\Omega}\|_{W^{1,s}(\Omega)} \leq C|\Omega|^{1+1/s+1/n-1/p}\|du\|_{p,\Omega} \quad (45)$$

for any convex domain $\Omega \subset \mathbb{R}^n$ with $|\Omega| < \infty$.

By the same method as we did for the case $1 < p < n$ above, we prove the following L^p norm inequality with the higher integral exponent on the left side for the case $p \geq n$.

Corollary 3 Let $u \in D'(\Omega, \wedge^l)$ be a differential form, $l = 1, 2, \dots, n$, $p \geq n$, and T be the homotopy operator and G be Green's operator. If $du \in L^p(\Omega, \wedge^l)$, then $T(G(u)) \in W^{1,s}(\Omega, \wedge^l)$ for any $s > 0$ and

$$\|T(G(u)) - (T(G(u)))_{\Omega}\|_{s,\Omega} \leq C\|du\|_{p,\Omega} \quad (46)$$

for any convex domain $\Omega \subset \mathbb{R}^n$ with $|\Omega| < \infty$.

Remark (i) The global inequalities can be proved in more general domains, such as the L^p -averaging domains and $L^{\varphi}(\mu)$ -averaging domains, see [1] for more properties of these two kinds of domains. (ii) The method developed in this paper can be used to prove the norm inequalities for other operators.

References

1. R.P. Agarwal, S. Ding, C. Nolder, *Inequalities for Differential Forms* (Springer, New York, 2009). MR 2552910
2. H. Cartan, *Differential Forms* (Translated from the French, Houghton Mifflin, Boston, 1970). MR 0267477
3. T. Iwaniec, A. Lutoborski, Integral estimates for null Lagrangians. Arch. Rational Mech. Anal. **125**(1), 25–79 (1993). MR 1241286
4. C.A. Nolder, Hardy-Littlewood theorems for A -harmonic tensors. Illinois J. Math. **43**(4), 613–632 (1999). MR 1712513
5. C. Scott, L^p theory of differential forms on manifolds. Trans. Amer. Math. Soc. **347**(6), 2075–2096 (1995). MR 1297538
6. Y. Wang, C. Wu, Global Poincaré inequalities for Green's operator applied to the solutions of the nonhomogeneous A -harmonic equation. Comput. Math. Appl. **47**(10–11), 1545–1554 (2004). MR 2079864
7. F.W. Warner, *Foundations of Differentiable Manifolds and Lie Groups*. Graduate Texts in Mathematics, vol. 94 (Springer, New York, 1983). Corrected reprint of the 1971 edition. MR 722297

A Variational Inequality Based Stochastic Approximation for Inverse Problems in Stochastic Partial Differential Equations



Rachel Hawks, Baasansuren Jadamba, Akhtar A. Khan, Miguel Sama, and Yidan Yang

Abstract The primary objective of this work is to study the inverse problem of identifying a parameter in partial differential equations with random data. We explore the nonlinear inverse problem in a variational inequality framework. We propose a projected-gradient-type stochastic approximation scheme for general variational inequalities and give a complete convergence analysis under weaker conditions on the random noise than those commonly imposed in the available literature. The proposed iterative scheme is tested on the inverse problem of parameter identification. We provide a derivative characterization of the solution map, which is used in computing the derivative of the objective map. By employing a finite element based discretization scheme, we derive the discrete formulas necessary to test the developed stochastic approximation scheme. Preliminary numerical results show the efficacy of the developed framework.

2010 Mathematics Subject Classification 35R30, 49N45, 65J20, 65J22, 65M30

1 Introduction

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, that is, Ω is a nonempty set whose elements are termed as elementary events, \mathcal{F} is a σ -algebra of subsets of Ω , and \mathbb{P} a probability measure. Let $D \subset \mathbb{R}^n$ be a sufficiently smooth bounded domain and ∂D be the boundary of Ω . Given two random fields $a : \Omega \times D \mapsto \mathbb{R}$ and $f : \Omega \times D \rightarrow \mathbb{R}$, we consider the stochastic partial differential equation (SPDE) of

R. Hawks · B. Jadamba · A. A. Khan (✉) · Y. Yang
School of Mathematical Sciences, Rochester Institute of Technology, Rochester, NY, USA
e-mail: rh9784@rit.edu; bxjsma@rit.edu; aakmsma@rit.edu; yy2513@rit.edu

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

finding a random field $u : \Omega \times D \rightarrow \mathbb{R}$ that almost surely satisfies the following:

$$-\nabla \cdot (a(\omega, x)\nabla u(\omega, x)) = f(\omega, x), \text{ in } D, \quad (1a)$$

$$u(\omega, x) = 0, \text{ on } \partial D. \quad (1b)$$

The above SPDE models interesting real-world phenomena and has been studied in great detail in the deterministic setting. For example, in (1), u may represent the steady-state temperature at a given point of a body; then a would be a variable thermal conductivity coefficient, and f an external heat source. The system (1) also models underground steady-state aquifers in which the parameter a is the aquifer transmissivity coefficient, u is the hydraulic head, and f is the recharge. Some details on inverse problems can be found in [6–8, 10, 12, 16, 18, 20–22].

A natural interpretation of (1) is that realizations of the data lead to deterministic PDEs. That is, for a fixed $\omega \in \Omega$, SPDE (1), under appropriate conditions, admits a weak solution $u(\omega, \cdot) \in H_0^1(D)$.

The objective of this work is to study the nonlinear inverse problem of identifying the parameter a from a measurement of the solution u of (1) by solving a stochastic optimization problem of the following form:

$$\min_{a \in \mathbb{A}} \mathbb{J}(a) := \mathbb{E}[J(a, \omega)]. \quad (2)$$

Here \mathbb{A} is the set of feasible parameters, which is a subset of a real Hilbert space H , $J(a, \omega)$ is a misfit function, which we will define shortly, and \mathbb{E} is the expectation with respect to the probability measure.

If the expected value $\mathbb{E}[J(a, \omega)]$ is readily assessable, either analytically or numerically, then (2) is practically a deterministic optimization problem that can be solved by a wide variety of available numerical methods. However, the evaluation of $\mathbb{E}[J(a, \omega)]$ is a challenging task. For instance, even when the random vector ω has a known probability distribution, the computation of the expected value $\mathbb{E}[J(a, \omega)]$ could involve computationally expensive multi-dimensional integral evaluations. A likely scenario is when the function $J(a, \omega)$ is known, but the distribution of ω is unknown, and any information on ω can only be gathered using sampling. Another challenging situation occurs when the expected value $\mathbb{E}[J(a, \omega)]$ is not observable, and it must be evaluated through a simulation process. In all such situations, the existing methods for deterministic optimization problems are not applicable and ought to be appropriately modified.

Our objective is to employ the stochastic approximation approach (SAA) in a Hilbert space setting for solving the nonlinear inverse problem of parameter identification in stochastic PDEs. The SAA has a long history and has been used for a wide variety of problems. On the other hand, SPDEs have also received a great deal of attention in recent years. To the best of our knowledge, this is the first time that the SAA approach is being used for nonlinear inverse problems. Before describing the main contributions and our strategy, let us briefly discuss the key ideas that have been used in these two disciplines.

During the past several decades, the dynamic field of stochastic approximation, initiated by the seminal paper by Robbins and Monro [32], witnessed an explosive growth in numerous directions. To give a brief review of some work relevant to this research, we note that Kiefer and Wolfowitz [25] extended the stochastic approximation approach to finding a unique minimum/maximum of the one-dimensional unknown regression function. An informative survey of the earlier developments in the stochastic approximation is by Lai [27]. Many authors have studied stochastic approximation in general space inspired by applications in control theory and related areas. A small sample of such contributions includes the research by Barty, Roy, and Strugarek [3], Goldstein [17], Kushner and Shwartz [26], Salov [35], Yin and Zhu [37], where more references can be found. Interesting related results have been given by Bertsekas and Tsitsiklis [5], Culioli and Cohen [9], and others.

On the other hand, the stochastic PDE-constrained optimization also attracted a great deal of attention in recent years. Such problems emerged from two sources, the inverse problems of parameter or source identification and optimal control problems. For example, Narayanan and Zabaras [2] studied the inverse problem in the presence of uncertainties in the material data and developed an adjoint approach based identification process using the spectral stochastic finite element method. In [38], the authors developed a scalable methodology for the stochastic inverse problem using a sparse grid collocation approach. In [36], the authors developed a robust and efficient method by employing generalized polynomial chaos expansion to identifying uncertain elastic parameters from experimental modal data. In [30], the authors presented an implicit sampling for parameter identification. In [34], the authors studied the parameter identification in a Bayesian setting for the elastoplastic problem. In [31], the authors explored the optimal control problem for the stochastic diffusion equation. In [24], the authors focused on determining the optimal thickness subjected to stochastic force. In [1], the authors investigated the impact of errors and uncertainties of the conductivity on the electrocardiography imaging solution.

Since the stochastic approximation approach is designed for problems with either noisy experimental values or noisy samples of the function, it seems ideal for solving inverse and ill-posed problems. However, the use of the stochastic approximation approach is mostly non-existent. Note that Bertran [4], who studied a stochastic projected gradient algorithm in a Hilbert space, gave an application to optimal control problems where the data was uncertain. A formal study of the stochastic approximation approach for optimal control in stochastic PDEs was initiated independently by Martin, Krumschield, and Nobile [29] and Geiersbach and Pflug [11]. Since the control-to-state map is linear, these problems involve a convex objective function. On the other hand, the inverse problem we consider in the present work is nonlinear, and the commonly used output least-squares (OLS) objective functional is nonconvex, in general. Therefore, the classical results of convex optimization cannot be combined with the SAA approach. We circumvent this difficulty by employing a modified output least-squares (MOLS) objective functional that uses the energy norm and is always convex. The use of the MOLS

functional permits us to use the stochastic approximation to the inverse problem of identifying a parameter in stochastic PDEs.

The contents of this paper are organized into five sections. Section 2 presents a projected gradient scheme for variational inequalities in the general stochastic approximation framework. We provide complete convergence analysis for the proposed iterative scheme under weaker conditions on the random noise. In Section 3, we focus on the inverse problem and present three optimization formulations, namely, the OLS functional, the MOLS, functional, and the equation error (EE) approach. The primary focus, however, remains on the MOLS approach. Besides providing technical details on the three functionals in a continuous setting, we also provide a discretization scheme, including discrete formulas for the objective functionals and the gradient for the MOLS functional. Two numerical examples, given in Section 4, demonstrate the feasibility and the efficacy of the developed framework. The paper concludes with some remarks and a discussion of future objectives.

2 Stochastic Approximation for Variational Inequalities

Let H be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$, $K \subset H$ be nonempty, closed, and convex, and $F : H \mapsto H$ be a given map.

We consider the variational inequality of finding $u \in K$ such that

$$\langle F(u), v - u \rangle \geq 0, \quad \text{for all } v \in K. \quad (3)$$

We aim to develop iterative methods for (3) in the general framework of stochastic approximation, that is, when the map F can only be accessed with some random noise. As a particular case, we will explore the variational inequality of finding $u \in K$ such that

$$\langle \mathbb{E}[G(u, \omega)], v - u \rangle \geq 0, \quad \text{for every } v \in K, \quad (4)$$

where $G(\cdot, \cdot) : \Omega \times H \mapsto H$, and $\mathbb{E}[G(u, \omega)]$ is the expected value of $G(u, \omega)$.

Our focus is on the following projected stochastic approximation scheme for (3):

$$u_{n+1} = P_K[u_n - \alpha_n(F(u_n) + \omega_n)], \quad \alpha_n > 0. \quad (5)$$

Here $F(u_n)$ is the true value of F at u_n , $F(u_n) + \omega_n$ is an approximation of F at u_n , and ω_n is a stochastic error. In the context of (4), $F(u_n) + \omega_n = G(u_n, \omega_n)$, where ω_n is a sample of the random variable ω . To be specific, here at iteration n , we use a sample ω_n of ω to calculate $G(u_n, \omega)$ and treat it as an approximation of $\mathbb{E}[G(u_n, \omega)] = F(u_n)$. Evidently, $F(u_n)$ can be approximated without any information on the probability distribution of ω .

We recall that, given the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a filtration $\{\mathcal{F}_n\} \subset \mathcal{F}$ is an increasing sequence of σ -algebras. A sequence of random variable $\{\omega_n\}$ is termed to be adapted to a filtration \mathcal{F}_n , if and only if, $\omega_n \in \mathcal{F}_n$ for all $n \in \mathbb{N}$, that is, ω_n is \mathcal{F}_n -measurable. Moreover, the natural filtration is the filtration generated by the sequence $\{\omega_n\}$ and is given by $\mathcal{F}_n = \sigma(\omega_m : m \leq n)$.

The following result by Robbins and Siegmund [33] will be used shortly:

Lemma 1 *Let \mathcal{F}_n be an increasing sequence of σ -algebras, and V_n, a_n, b_n , and c_n be nonnegative random variables adapted to \mathcal{F}_n . Assume that $\sum_{n=1}^{\infty} a_n < \infty$ and*

$$\sum_{n=1}^{\infty} b_n < \infty, \text{ almost surely, and}$$

$$\mathbb{E}[V_{n+1} | \mathcal{F}_n] \leq (1 + a_n)V_n - c_n + b_n.$$

Then $\{V_n\}$ is almost surely convergent and $\sum_{n=1}^{\infty} c_n < \infty$, almost surely.

We also need the following notions of monotonicity and continuity:

Definition 1 Given the Hilbert space H , let $F : X \mapsto X^*$ be a nonlinear map. The map F is called:

1. **monotone**, if

$$\langle Fu - Fv, u - v \rangle \geq 0, \text{ for all } u, v \in X. \tag{6}$$

2. **m -strongly monotone**, if there exists a constant $m > 0$ such that

$$\langle Fu - Fv, u - v \rangle \geq m\|u - v\|^2, \text{ for all } u, v \in X. \tag{7}$$

3. **L -Lipschitz continuous**, if there exists a constant $L > 0$ such that

$$\|Fu - Fv\| \leq L\|u - v\|, \text{ for all } u, v \in X. \tag{8}$$

4. **hemicontinuous**, if the real function $t \mapsto \langle F(u + tv), w \rangle$ is continuous on $[0, 1]$, for all $u, v, w \in X$.

The following result provides the convergence analysis for the scheme (5):

Theorem 1 *Let H be a real Hilbert space, $K \subset H$ be nonempty, closed, and convex, and $F : H \mapsto H$ be given. Let $\{\omega_n\}$ be an H -valued sequence of random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\{u_n\}$ be the sequence generated by (5) and $\mathcal{F}_n := \sigma(u_0, \dots, u_n)$ be a filtration on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\{u_n\}$ is \mathcal{F}_n -measurable. Assume that the following conditions hold:*

(A₁) *There is a constant $c > 0$ such that $\|F(u)\| \leq c(1 + \|u\|)$, for every $u \in K$.*

(A₂) F is m -strongly monotone and hemicontinuous.

(A₃) There are constants $c_1 \geq 0$ and $c_2 > 0$ such that

$$\|\mathbb{E}[\omega_n | \mathcal{F}_n]\| \leq c_1 \beta_n (1 + \|F(u_n)\|), \quad \beta_n > 0, \quad (9)$$

$$\mathbb{E}\left[\|\omega_n\|^2 | \mathcal{F}_n\right] \leq c_2 \left(1 + \frac{1}{\delta_n} \|F(u_n)\|^2\right), \quad \delta_n > 0. \quad (10)$$

(A₄) The sequences $\{\alpha_n\}$, $\{\beta_n\}$, and $\{\delta_n\}$ of positive reals satisfy:

$$\sum_{n \in \mathbb{N}} \alpha_n = \infty, \quad \sum_{n \in \mathbb{N}} \alpha_n^2 < \infty, \quad \sum_{n \in \mathbb{N}} \frac{\alpha_n^2}{\delta_n} < \infty, \quad \sum_{n \in \mathbb{N}} \alpha_n \beta_n < \infty. \quad (11)$$

Then, $\{u_n\}$ converges, almost surely, to the unique solution \bar{u} of (3).

Proof Due to the strong monotonicity of F , variational inequality (3) has a unique solution $\bar{u} \in K$. Then, we have $\bar{u} = P_K(\bar{u})$, and by (5) and the m -strong monotonicity of F , we get

$$\begin{aligned} \|u_{n+1} - \bar{u}\|^2 &= \|P_K(u_n - \alpha_n(F(u_n) + \omega_n)) - P_K(\bar{u})\|^2 \\ &\leq \|u_n - \bar{u} - \alpha_n(F(u_n) + \omega_n)\|^2 \\ &= \|u_n - \bar{u}\|^2 + \alpha_n^2 \|F(u_n) + \omega_n\|^2 - 2\alpha_n \langle F(u_n) + \omega_n, u_n - \bar{u} \rangle \\ &\leq (1 - 2m\alpha_n) \|u_n - \bar{u}\|^2 + 2\alpha_n^2 \|F(u_n)\|^2 + 2\alpha_n^2 \|\omega_n\|^2 \\ &\quad - 2\alpha_n \langle \omega_n, u_n - \bar{u} \rangle, \end{aligned}$$

where we used the m -strong monotonicity of F to deduce that

$$\langle F(u_n), u_n - \bar{u} \rangle \geq m \|u_n - \bar{u}\|^2 + \langle F(\bar{u}), u_n - \bar{u} \rangle \geq m \|u_n - \bar{u}\|^2.$$

Next, by taking conditional expectation with respect to \mathcal{F}_n , we deduce

$$\begin{aligned} \mathbb{E}[\|u_{n+1} - \bar{u}\|^2 | \mathcal{F}_n] &\leq (1 - 2m\alpha_n) \|u_n - \bar{u}\|^2 + 2\alpha_n^2 \|F(u_n)\|^2 \\ &\quad + 2\alpha_n^2 \mathbb{E}\left[\|\omega_n\|^2 | \mathcal{F}_n\right] + 2\alpha_n \|u_n - \bar{u}\| \|\mathbb{E}[\omega_n | \mathcal{F}_n]\|. \end{aligned} \quad (12)$$

To find bounds on the terms in (12), we begin by noticing that

$$\begin{aligned} \|F(u_n)\| &\leq c(1 + \|u_n\|) \\ &\leq c(1 + \|\bar{u}\|) + c\|u_n - \bar{u}\| \\ &\leq k_1(1 + \|u_n - \bar{u}\|), \end{aligned} \quad (13)$$

where $k_1 := c(1 + \|\bar{u}\|)$, and hence by setting $k_2 := 4k_1^2$, we obtain

$$2\alpha_n^2 \|F(u_n)\|^2 \leq k_2 \alpha_n^2 (1 + \|u_n - \bar{u}\|^2). \quad (14)$$

Moreover, by the inequality $a \leq 1 + a^2$, which holds for all $a \in \mathbb{R}$, and (13), we get

$$\begin{aligned} \|u_n - \bar{u}\| \|\mathbb{E}[\omega_n | \mathcal{F}_n]\| &\leq \beta_n \|u_n - \bar{u}\| (1 + \|F(u_n)\|) \\ &\leq \beta_n \|u_n - \bar{u}\| (1 + k_1 + k_1 \|u_n - \bar{u}\|) \\ &\leq \beta_n (1 + k_1) \|u_n - \bar{u}\| + k_1 \beta_n \|u_n - \bar{u}\|^2 \\ &\leq \beta_n (1 + k_1) (1 + \|u_n - \bar{u}\|^2) + k_1 \beta_n \|u_n - \bar{u}\|^2 \\ &\leq \beta_n (1 + k_1) + (1 + 2k_1) \beta_n \|u_n - \bar{u}\|^2, \end{aligned}$$

and hence setting $k_3 := 2(1 + 2k_1)$, we obtain

$$2\alpha_n \|u_n - \bar{u}\| \|\mathbb{E}[\omega_n | \mathcal{F}_n]\| \leq k_3 \alpha_n \beta_n (1 + \|u_n - \bar{u}\|^2). \quad (15)$$

Finally, using (13) again, we obtain

$$\begin{aligned} \mathbb{E}[\|\omega_n\|^2 | \mathcal{F}_n] &\leq c_2 \left(1 + \frac{\|F(u_n)\|^2}{\delta_n}\right) \\ &\leq c_2 \left(1 + \frac{2k_1^2 (1 + \|u_n - \bar{u}\|^2)}{\delta_n}\right) \\ &\leq c_2 + \frac{2c_2 k_1^2}{\delta_n} (1 + \|u_n - \bar{u}\|^2), \end{aligned}$$

and hence, setting $k_4 := 4c_2 k_1^2$, we obtain

$$2\alpha_n^2 \mathbb{E}[\|\omega_n\|^2 | \mathcal{F}_n] \leq 2c_2 \alpha_n^2 + \frac{k_4 \alpha_n^2}{\delta_n} + \frac{k_4 \alpha_n^2}{\delta_n} \|u_n - \bar{u}\|^2. \quad (16)$$

Summarizing, due to (12), (14), (15), and (16), there is a constant $k > 0$ with

$$\begin{aligned} \mathbb{E}[\|u_{n+1} - \bar{u}\|^2 | \mathcal{F}_n] &\leq \left(1 - 2m\alpha_n + k\alpha_n^2 + k\alpha_n \beta_n + \frac{k\alpha_n^2}{\delta_n}\right) \|u_n - \bar{u}\|^2 \\ &\quad + k\alpha_n^2 + k\alpha_n \beta_n + \frac{k\alpha_n^2}{\delta_n}, \end{aligned} \quad (17)$$

which can be written as

$$\mathbb{E}[\|u_{n+1} - \bar{u}\|^2 | \mathcal{F}_n] \leq (1 + a_n) \|u_n - \bar{u}\|^2 - c_n + b_n,$$

where

$$\begin{aligned} a_n &:= k\alpha_n^2 + k\alpha_n\beta_n + \frac{k\alpha_n^2}{\delta_n}, \\ b_n &:= k\alpha_n^2 + k\alpha_n\beta_n + \frac{k\alpha_n^2}{\delta_n}, \\ c_n &:= 2m\alpha_n\|u_n - \bar{u}\|^2. \end{aligned}$$

Since $\sum_{n \in \mathbb{N}} a_n < \infty$ and $\sum_{n \in \mathbb{N}} b_n < \infty$, as a consequence of Theorem 1, it follows that $\|u_n - \bar{u}\|^2$ converges, almost surely, and

$$\sum_{n \in \mathbb{N}} 2m\alpha_n\|u_n - \bar{u}\|^2 < +\infty,$$

which, due to $\sum_{n \in \mathbb{N}} \alpha_n = \infty$, confirms that $\|u_n - \bar{u}\| \rightarrow 0$, almost surely. The proof is complete. □

We shall now discuss two special cases of the above result:

Corollary 1 *Let H be a real Hilbert space, $K \subset H$ be nonempty, closed, and convex, and $F : H \mapsto H$ be given. Let $\{\omega_n\}$ be an H -valued sequence of random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\{u_n\}$ be the sequence generated by (5) and $\mathcal{F}_n := \sigma(u_0, \dots, u_n)$ be a filtration on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\{u_n\}$ is \mathcal{F}_n -measurable. Assume that the following conditions hold:*

- (C₁) *F is m -strongly monotone and L -Lipschitz continuous.*
- (C₂) *$\mathbb{E}[\omega_n | \mathcal{F}_n] = 0$, and $\sum_n \alpha_n^2 \mathbb{E}[\|\omega_n\|^2 | \mathcal{F}_n] < \infty$.*
- (C₃) *$\alpha_n \in (0, 2m/L^2)$.*

Then, $\{u_n\}$ converges, almost surely, to the unique solution \bar{u} of (3).

Proof Note that $\bar{u} = P_K(\bar{u} - \alpha_n F(\bar{u}))$, and hence

$$\begin{aligned} \|u_{n+1} - \bar{u}\|^2 &= \|P_K(u_n - \alpha_n(F(u_n) + \omega_n)) - P_K(\bar{u} - \alpha_n F(\bar{u}))\|^2 \\ &\leq \|u_n - \bar{u} - \alpha_n(F(u_n) - F(\bar{u}) + \omega_n)\|^2 \\ &\leq \|u_n - \bar{u}\|^2 + \alpha_n^2\|F(u_n) - F(\bar{u}) + \omega_n\|^2 \\ &\quad - 2\alpha_n\langle F(u_n) - F(\bar{u}) + \omega_n, u_n - \bar{u} \rangle \\ &\leq (1 - 2m\alpha_n + 2\alpha_n^2 L^2)\|u_n - \bar{u}\|^2 + 2\alpha_n\|\omega_n\|^2 - 2\alpha_n\langle \omega_n, u_n - \bar{u} \rangle, \end{aligned}$$

and by taking the expectation past \mathcal{F}_n , we deduce

$$\mathbb{E} [\|u_{n+1} - \bar{u}\|^2 | \mathcal{F}_n] \leq (1 - 2m\alpha_n + 2\alpha_n^2 L^2) \|u_n - \bar{u}\|^2 + \alpha_n^2 \mathbb{E} [\|\omega_n\|^2 | \mathcal{F}_n],$$

which can be written as

$$\mathbb{E} [\|u_{n+1} - \bar{u}\|^2 | \mathcal{F}_n] \leq (1 + a_n) \|u_n - \bar{u}\|^2 - c_n + b_n,$$

where for a positive constant $k > 0$, we have

$$\begin{aligned} a_n &:= 0, \\ b_n &:= \alpha_n^2 \mathbb{E} [\|\omega_n\|^2 | \mathcal{F}_n], \\ s_n &:= 2(\alpha_n m - \alpha_n^2 L^2), \\ c_n &:= s_n \|u_n - \bar{u}\|^2. \end{aligned}$$

Due to imposed conditions, we have $\sum_{n \in \mathbb{N}} a_n < \infty$, and $\sum_{n \in \mathbb{N}} b_n < \infty$, almost surely.

As a consequence, $\|u_n - \bar{u}\|^2$ converges almost surely, and $\sum_{n=1}^{\infty} c_n < \infty$, almost surely. Furthermore, since s_n is bounded away from zero, we infer that the sequence $\{u_n\}$ converges strongly to \bar{u} , almost surely. The proof is complete. \square

Corollary 2 *Let H be a real Hilbert space, $K \subset H$ be nonempty, closed, and convex, and $F : H \mapsto H$ be given. Let $\{\omega_n\}$ be an H -valued sequence of random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\{u_n\}$ be the sequence generated by (5) and $\mathcal{F}_n := \sigma(u_0, \dots, u_n)$ be a filtration on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\{u_n\}$ is \mathcal{F}_n -measurable.*

(H₁) *There is a constant $c > 0$ such that $\|F(u)\| \leq c(1 + \|u\|)$, for every $u \in K$.*

(H₂) *F is m -strongly monotone and hemicontinuous.*

(H₃) $\mathbb{E}[\omega_n | \mathcal{F}_n] = 0$, and $\sum_n \alpha_n^2 \mathbb{E} [\|\omega_n\|^2 | \mathcal{F}_n] < \infty$.

(H₄) *The sequence $\{\alpha_n\}$ of positive reals satisfies:*

$$\sum_{n \in \mathbb{N}} \alpha_n = \infty, \quad \sum_{n \in \mathbb{N}} \alpha_n^2 < \infty. \tag{18}$$

Then, $\{u_n\}$ converges, almost surely, to the unique solution \bar{u} of (3).

Proof The proof is based on the arguments used above. \square

Remark 1 Hiriart-Urruty [19] extended the stochastic approximation approach to nonlinear variational inequalities when some random noise contaminated the data. He proposed a variety of projection-type iterative methods in Hilbert spaces, even considered variational inequalities with multi-valued maps, and provided several

convergence theorems in quadratic mean and almost certain sense. Theorem 1 is given under the same condition F as in [19]; however, we have more general conditions on random noise, which were inspired by Barty, Roy, and Strugarek [3]. Jiang and Xu [23] initiated a detailed study of the stochastic approximation framework for the expected value formulation of variational inequalities. Corollary 1 is similar to the results [23], given for the particular case of an expected value formulation of a variational inequality.

3 Stochastic Approximation for Inverse Problems

In this section, we will study the inverse problem of identifying a deterministic parameter in a stochastic partial differential equation. In the final section, we will discuss the extension of the present framework to the case of a stochastic parameter.

3.1 Optimization Formulation of the Inverse Problem

We recall that given a real Banach space X , a measure space $(\Omega, \mathcal{F}, \mu)$, and an integer $p \in [1, \infty)$, the Bochner space $L^p(\Omega, X)$ consists of Bochner integrable functions $u : \Omega \rightarrow X$ with finite p th moment, that is,

$$\|u\|_{L^p(\Omega, X)} := \left(\int_{\Omega} \|u(\omega)\|_X^p d\mu(\omega) \right)^{1/p} = \mathbb{E} [\|u(\omega)\|_X^p]^{1/p} < \infty.$$

If $p = \infty$, then $L^\infty(\Omega, X)$ is the space of Bochner measurable functions $u : \Omega \rightarrow X$ such that

$$\text{ess sup}_{\omega \in \Omega} \|u(\omega)\|_X < \infty.$$

For $\omega \in \Omega$, variational formulation of (1) seeks $u_\omega \in V := H_0^1(D)$ such that

$$\int_D a(\omega, x) \nabla u_\omega(a) \cdot \nabla v dx = \int_D f(\omega, x) v dx, \text{ for all } v \in V. \tag{19}$$

Assume that there are constants k_0 and k_1 such that

$$0 < k_0 \leq a(\omega, x) \leq k_1 < \infty, \text{ a.e. in } D \times \Omega.$$

The following is a well-known result for (19):

Lemma 2 *Let $f \in L^2(\Omega, H^{-1}(D))$. Then, there is a positive constant c such that*

$$\begin{aligned} \|u_\omega(a)\|_{H_0^1(D)} &\leq c\|f\|_{H^{-1}(D)} \quad \text{for a.e. } \omega \in \Omega, \\ \|u(a)\|_{L^2(\Omega, H_0^1(D))} &\leq c\|f\|_{L^2(\Omega, H_0^{-1}(D))}. \end{aligned}$$

In the following, we shall assume that a is deterministic. Moreover, for positive k_0 and k_1 , we define the set of admissible parameters:

$$\mathbb{A} := \{a \in L^\infty(D) : 0 < k_0 \leq a(x) \leq k_1 < \infty, x \in D\}. \quad (20)$$

We now state some technical results. Since these results are stated for realizations, their proofs are natural generalizations of the results given in [15] for the corresponding deterministic case.

Theorem 2 For $\omega \in \Omega$, the map $\mathbb{A} \ni a \mapsto u_\omega(a)$ is Lipschitz continuous.

Theorem 3 For $\omega \in \Omega$, and a in the interior of \mathbb{A} , the map $a \mapsto u_\omega(a)$ is differentiable at a . The derivative $\delta u_\omega := Du_\omega(a)(\delta a)$ of $u_\omega(a)$ at a in the direction δa is the unique solution of the stochastic variational problem: Find $\delta u_\omega \in V$ such that

$$\int_D a(x)\nabla\delta u_\omega \cdot \nabla v dx = - \int_D \delta a \nabla u_\omega(a) \cdot \nabla v dx, \quad \text{for all } v \in V. \quad (21)$$

One of the most commonly used optimization formulations is the following output least-squares (OLS) objective functional:

$$\widehat{\mathbb{J}}(a) = \frac{1}{2}\mathbb{E}\left[\|u_\omega(a) - z_\omega\|_{L^2(D)}^2\right], \quad (22)$$

where $u_\omega(a)$ is the solution of (19) for a and z_ω is the measured data.

One of the shortcomings of the above functional is that it is nonconvex, in general. Although it is known that the gradient of the OLS functional, with the aid of a regularization, can be made strongly monotone, it runs into the risk of over-regularizing the identification process, see [14].

We now define the modified output least-squares (MOLS) objective functional:

$$\mathbb{J}(a) = \frac{1}{2}\mathbb{E}\left[\int_D a(x)\nabla(u_\omega(a) - z_\omega) \cdot \nabla(u_\omega(a) - z_\omega)dx\right], \quad (23)$$

where $u_\omega(a)$ is the solution of (19) for a and z_ω is the measured data.

The following result summarizes some properties of the MOLS objective:

Theorem 4 Let a be in the interior of \mathbb{A} . Then:

1. The first derivative of \mathbb{J} at a is given by

$$D\mathbb{J}(a)(\delta a) = -\frac{1}{2}\mathbb{E}\left[\int_D \delta a \nabla(u_\omega(a) + z_\omega) \nabla(u_\omega(a) - z_\omega) dx\right].$$

2. The second derivative of \mathbb{J} at a is given by

$$D^2\mathbb{J}(a)(\delta a, \delta a) = \mathbb{E}\left[\int_D a(x) \nabla u_\omega(a) \nabla u_\omega(a) dx\right].$$

Consequently, the MOLS functional is convex in the interior of the set \mathbb{A} .

For the sake of a comparison, we would also describe another commonly used method, the so-called equation error approach (see [13]), which consists of minimizing, for $\omega \in \Omega$, and for the data $z_\omega \in H_0^1(D)$, the quadratic objective functional:

$$\min_{a \in \mathbb{A}} \tilde{\mathbb{J}}(a) = \frac{1}{2}\mathbb{E}\left[\|e_\omega(a, z_\omega)\|_{H_0^1}^2\right], \quad (24)$$

where $e_\omega(a, u_\omega) \in H_0^1(D)$ satisfies the following variational problem:

$$\langle e_\omega(a, u_\omega), v \rangle_{H_0^1(D)} = \int_D a \nabla u_\omega \cdot \nabla v - \int_D f(\omega, x)v, \quad \text{for all } v \in H_0^1(D).$$

Since the inverse problem of identifying parameters in partial differential equations is ill-posed, and for a stable identification process, some regularization is needed. For this, we assume that the set of admissible parameters \mathbb{A} belongs to a Hilbert space that is compactly embedded into $L^\infty(D)$.

Therefore, we consider the following regularized analogues of the three functionals described above:

$$\min_{a \in \mathbb{A}} \widehat{\mathbb{J}}_\kappa(a) := \frac{1}{2}\mathbb{E}\left[\|u_\omega(a) - z_\omega\|_{L^2(D)}^2\right] + \frac{\kappa}{2}\|a\|_H^2, \quad (25)$$

$$\min_{a \in \mathbb{A}} \mathbb{J}_\kappa(a) := \frac{1}{2}\mathbb{E}\left[\int_D a(x) \nabla(u_\omega(a) - z_\omega) \cdot \nabla(u_\omega(a) - z_\omega) dx\right] + \frac{\kappa}{2}\|a\|_H^2, \quad (26)$$

$$\min_{a \in \mathbb{A}} \tilde{\mathbb{J}}_\kappa(a) := \frac{1}{2}\mathbb{E}\left[\|e_\omega(a, z_\omega)\|_{H_0^1}^2\right] + \frac{\kappa}{2}\|a\|_H^2. \quad (27)$$

Here $u_\omega(a)$ is the solution of (19) for $a(x)$, z_ω is the measured data, $\kappa > 0$ is a fixed regularization parameter, and $\|\cdot\|_H^2$ is the regularizer.

Since \mathbb{J} is convex and \mathbb{A} is closed and convex, the following variational inequality is a necessary and sufficient optimality condition for (26): Find $a \in \mathbb{A}$ such that

$$\langle \nabla \mathbb{J}(a), b - a \rangle + \kappa \langle a, b - a \rangle \geq 0, \quad \text{for every } b \in \mathbb{A}. \quad (28)$$

Note that by defining

$$J(a, \omega) = \int_D a(x) \nabla(u_\omega(a) - z_\omega) \cdot \nabla(u_\omega(a) - z_\omega) dx,$$

we can show that

$$\nabla \mathbb{J}(a, \omega)(\delta a) = -\frac{1}{2} \int_D \delta a(x) \nabla(u_\omega(a) + z_\omega) \nabla(u_\omega(a) - z_\omega) dx,$$

and, consequently,

$$\nabla \mathbb{J}(a) = \nabla \mathbb{E}[J(a, \omega)] = \mathbb{E}[\nabla J(a, \omega)].$$

Therefore, it follows that

$$\nabla \mathbb{J}_\kappa(a) = \nabla \mathbb{E}[J(a, \omega) + \kappa a] = \mathbb{E}[\nabla J(a, \omega) + \kappa a] = \mathbb{E}[G(a, \omega)], \quad (29)$$

where we set $G(a, \omega) = \nabla J(a, \omega) + \kappa a$.

Analogous statements can be made for the OLS objective and the EE objective.

3.2 Discrete Formulas

We will use a standard finite element discretization of the spaces V and H . We begin, therefore, with a triangulation \mathcal{T}_h on D . Let V_h and H_h be the spaces of piecewise linear continuous polynomials relative to \mathcal{T}_h . Let $\{\phi_1, \phi_2, \dots, \phi_m\}$ and $\{\varphi_1, \varphi_2, \dots, \varphi_l\}$ be the corresponding bases for V_h and H_h , respectively. The space H_h is then isomorphic to \mathbb{R}^l , and for any $a \in H_h$, we define $A \in \mathbb{R}^l$ by $A_i = a(x_i)$, $i = 1, 2, \dots, l$, where the nodal basis $\{\varphi_1, \varphi_2, \dots, \varphi_l\}$ corresponds to the nodes $\{x_1, x_2, \dots, x_l\}$. Conversely, each $A \in \mathbb{R}^l$ corresponds to $a \in H_h$ defined by $a(x) = \sum_{i=1}^l A_i \varphi_i$. Analogously, $u \in V_h$ will correspond to $U \in \mathbb{R}^m$, where

$U_i = u(y_i)$, $i = 1, 2, \dots, m$, and $u = \sum_{i=1}^m U_i \phi_i$, where y_1, y_2, \dots, y_m are the interior nodes of the finite element mesh (triangulation) \mathcal{T}_h .

Given a realization/sample $\omega \in \Omega$, the discrete version of variational problem (19) seeks $U = U(\omega, A) \in \mathbb{R}^m$ by solving

$$K(A)U(\omega, A) = F(\omega),$$

where $K(A) \in \mathbb{R}^{m \times m}$ and $F(\omega_n) \in \mathbb{R}^m$ are the stiffness matrix and the load vector defined by

$$K(A)_{i,j} = \int_D a_h(x) \nabla \phi_j \cdot \nabla \phi_i dx, \quad \text{for } i, j = 1, \dots, m,$$

$$F(\omega)_i = \int_D f_h(\omega, x) \phi_i dx, \quad \text{for } i = 1, \dots, m.$$

To compute the gradient of the MOLS objective, it is convenient to define the so-called adjoint stiffness matrix $L(\cdot) \in \mathbb{R}^{m \times l}$ by the condition

$$L(V)A = K(A)V, \quad \text{for every } V \in \mathbb{R}^m, \quad A \in \mathbb{R}^l.$$

Then,

$$\begin{aligned} \nabla J(A, \omega)(\delta A) &= -\frac{1}{2}(U(\omega, A) + Z(\omega))^\top K(\delta A)(U(\omega, A) - Z(\omega)) \\ &= -\frac{1}{2}\delta A^\top L(U(\omega, A) + Z(\omega))^\top (U(\omega, A) - Z(\omega)), \end{aligned}$$

which yields

$$\nabla J_\kappa(A, \omega) = -\frac{1}{2}L(U(\omega, A) + Z(\omega))^\top (U(\omega, A) - Z(\omega)) + \kappa(\mathbb{M} + \mathbb{K})A,$$

where $\mathbb{M}, \mathbb{K} \in \mathbb{R}^{m \times m}$ are the corresponding mass and stiffness matrices in H_h :

$$\begin{aligned} \mathbb{M}_{i,j} &= \int_D \varphi_j \varphi_i dx, \quad \text{for } i, j = 1, \dots, l, \\ \mathbb{K}_{i,j} &= \int_D \nabla \varphi_j \cdot \nabla \varphi_i dx, \quad \text{for } i, j = 1, \dots, l. \end{aligned}$$

The above preparation permits to define the following stochastic approximation scheme for computing a solution of the discrete variant of (28):

In the classical stochastic gradient, a single sampling is done at each iterative step. However, in the above algorithm, instead of sampling the random variable at each step once, at step n , we sample a predetermined number s_n times, called the sample rate, and use the empirical average to approximate the expected value.

4 Computational Experiments

In this section, we present results from our numerical computations. We consider the domain $D = (0, 1)$ and choose functions $a(x)$ and $u(\omega, x) = u(Y_1(\omega), Y_2(\omega), x)$ and compute the corresponding $f(\omega, x)$ by $f(\omega, x) = -(a(x)u_x(\omega, x))_x$. We choose a uniform mesh on $(0, 1)$ with mesh size $h = 1/(N + 1)$, where N stands

Algorithm 1 Stochastic approximation for parameter identification

- 1: Choose an initial guess $A_0 \in \mathbb{R}^m$, positive step-lengths $\{\alpha_n\}$ satisfying (18), the sample rate $\{s_n\} \subset \mathbb{N}$, and initial samples $\{\omega_j^0\}_{j=1}^{s_0}$ of the random variable ω .
- 2: Given $A_n \in A$, generate samples $\{\omega_j^n\}_{j=1}^{s_n}$ of ω and define

$$A_{n+1} = P_{\mathbb{A}} \left[A_n - \frac{\alpha_n}{s_n} \sum_{j=1}^{s_n} G \left(\omega_j^n, A_n \right) \right], \tag{30}$$

where G is the discrete variant of gradient of the regularized MOLS objective (see (29)).

- 3: Stop if some stopping criteria are met.
-

for the number of interior nodes. The same set of piecewise linear finite element basis functions is used for the representations of $a(x)$ and $u(\omega, x)$; therefore, $U(\omega, A) \in \mathbb{R}^N$ (for a fixed ω) and $A \in \mathbb{R}^{N+2}$ (i.e., $m = N + 2$). The constraint set K is defined by

$$\mathbb{A} = \{a \in H^1(\Omega) : a_0 \leq a(x) \leq a_1\}.$$

Example 1 For this example, we choose

$$\begin{aligned} a(x) &= 1 + x^2, \\ u(\omega, x) &= Y_1(\omega)x(1 - x) + Y_2(\omega) \sin(3\pi x), \end{aligned}$$

where $Y_1(\omega), Y_2(\omega) \sim U[0, 1]$, i.e., random variables Y_1 and Y_2 are uniformly distributed over interval $[0, 1]$. We choose $a_0 = 0.5$ and $a_1 = 3$ and use $N = 99$, $s_n = 5$, $\alpha_n = 0.5\alpha_0/n$ with $\alpha_0 = 10^4$ in Algorithm 1 for this example. Iterations are terminated once the L^2 norm of the expected value of the gradient drops below $\gamma = 10^{-7}$. Results of this computation using the MOLS method are shown in Figure 1. Regularization parameter $\kappa = 10^{-6}$ is used to produce these figures.

Example 2 In this example, we choose the same $u(\omega, x)$ as in Example 1, but with a slightly more interesting function $a(x)$ defined by

$$a(x) = 2 \sin(\pi(x - 0.2)) - 2 \tanh(20x - 8) + 4.$$

Figure 2 shows results of a run using parameters $N = 159$, $s_n = 10$, $\alpha_0 = 10^5$, and $\kappa = 10^{-7}$ using the MOLS method. For the constraints, we use $a_0 = 1$ and $a_1 = 8$. Figure 3 shows some realizations of the random fields $u(\omega, x)$ and $f(\omega, x)$. Note that Figures 1 and 2 represent results of a typical simulation. Regularization parameter κ is chosen after we do several test runs for a particular set of parameter values. The method gives us a very stable reconstruction of the coefficient $a(x)$ in each case regardless of the choice of the initial approximate $A^{(0)}$.

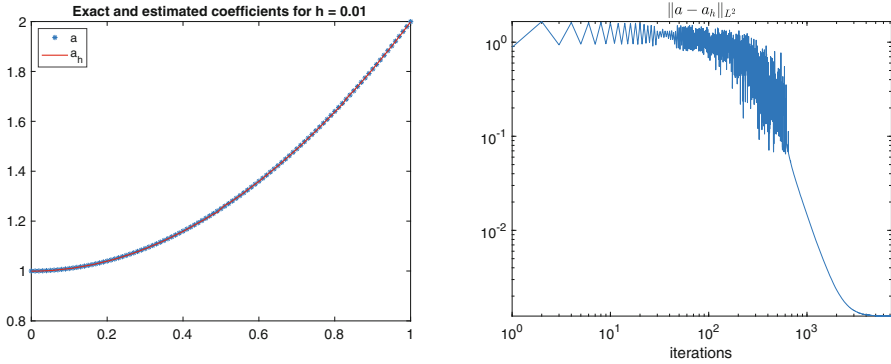


Fig. 1 Example 1: Comparison of exact coefficient a and the approximated coefficient a_h using MOLS method (left) and loglog plot of the error $\|a - a_h\|_{L^2}$ versus iterations (right)

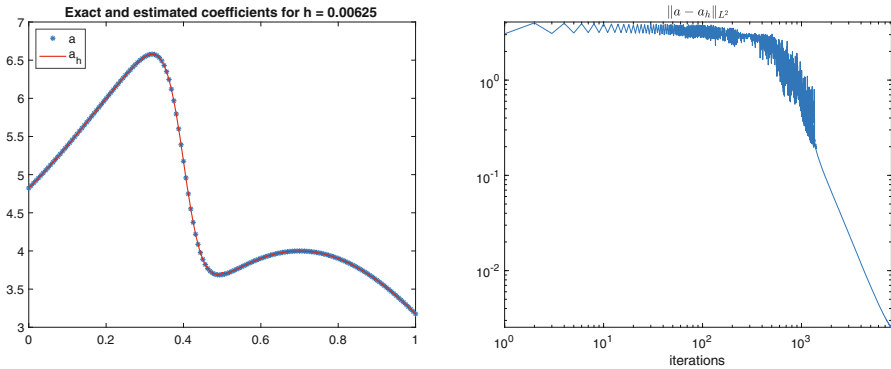


Fig. 2 Example 2: Comparison of exact coefficient a and the approximated coefficient a_h using MOLS method (left) and loglog plot of the error $\|a - a_h\|_{L^2}$ versus iterations (right)

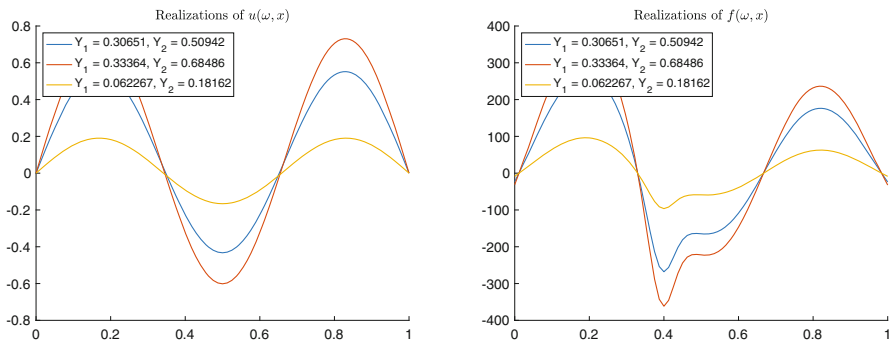


Fig. 3 Typical realizations of the random fields u and f from Example 2

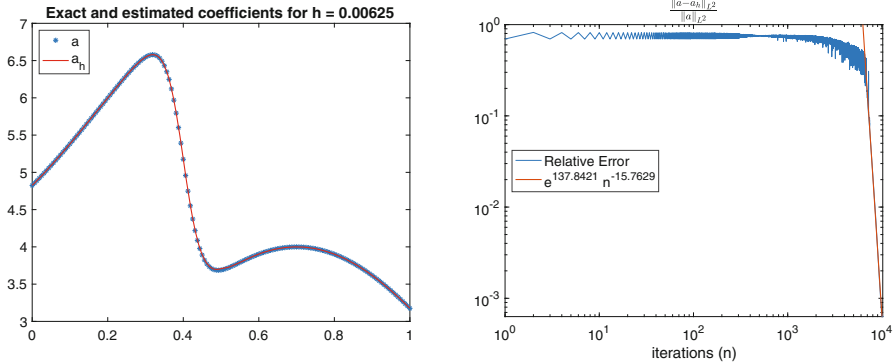


Fig. 4 Example 2: Comparison of exact coefficient a and the approximated coefficient a_h using EE method (left) and loglog plot of the relative error $\|a - a_h\|_{L^2} / \|a\|_{L^2}$ versus iterations (right)

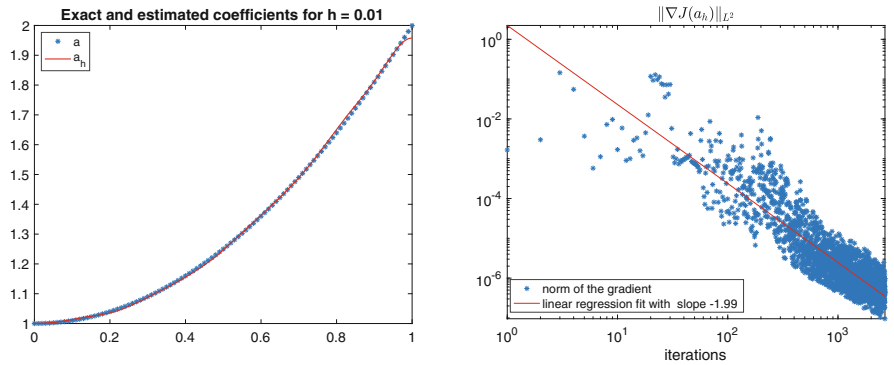


Fig. 5 Example 1: Comparison of exact coefficient a and the approximated coefficient a_h using the OLS method (left) and loglog plot of the norm of the gradient $\|\nabla J(a_h)\|_{L^2}$ versus iterations (right)

Results of Computations by EE and OLS Methods We compare the performance of the MOLS method with those of the OLS and EE methods (see equations (27) and (25) for regularized objective functional definitions). Figure 4 shows the results of a run with parameters $N = 159$, $s_n = 20$, $\alpha_0 = 10^6$, and $\kappa = 5 \cdot 10^{-7}$ for Example 2 using EE method. The quality of the estimation is excellent and the results are comparable with those of the MOLS method. No significant gain in the computational cost for the EE method is observed as our examples are in 1D (these computations take only a minute or two in MATLAB). However, the EE method is expected to have considerable computational cost advantage for problems in two or three space dimensions compared to MOLS and OLS methods. Figure 5 shows results of a simulation using OLS method for Example 1. Parameter values used in the computation are $N = 99$, $s_n = 1$, $\alpha_0 = 10^5$, and $\kappa = 5 \cdot 10^{-6}$. Tolerance for the L^2 norm of the gradient is set to $\gamma = 10^{-7}$ (see the right plot in the figure

referenced above which shows the decrease of this norm as iterations progress). The quality of the estimation seems to be not as good as the ones we obtained from the MOLS and EE methods, and there is a mismatch close to the right boundary of the domain. While applying the OLS method to both examples, we observed that the method requires a more careful tuning of the parameters compared to the other two methods we used in our experiments.

5 Concluding Remarks

We developed a stochastic approximation approach for identifying a deterministic parameter in a stochastic partial differential equation. Besides considering more general stochastic PDEs such as linear elasticity or fourth-order plate models, a desirable extension of this work is to identify a stochastic parameter $a(\omega, x)$. A natural approach would be to separate the deterministic and stochastic components by using the so-called finite-dimensional noise assumption (see [28]). The deterministic components can then be identified by extending the stochastic approximation framework. We aim to pursue this approach in future work.

Acknowledgments Contributions of R. Hawks, B. Jadamba, A. A. Khan, and Y. Yang are supported by the National Science Foundation under Award No. 1720067. M. Sama's work is partially supported by the Ministerio de Ciencia, Innovacion y Universidades (MCIU), Agencia Estatal de Investigacion (AEI) (Spain), and Fondo Europeo de Desarrollo Regional (FEDER) under project PGC2018-096899-B-I00 (MCIU/AEI/FEDER, UE).

References

1. R. Aboulaich, N. Fikal, E. El Guarmah, N. Zemzemi, Stochastic finite element method for torso conductivity uncertainties quantification in electrocardiography inverse problem. *Math. Model. Nat. Phenom.* **11**(2), 1–19 (2016)
2. V.A. Badri Narayanan, N. Zabaras, Stochastic inverse heat conduction using a spectral approach. *Internat. J. Numer. Methods Eng.* **60**(9), 1569–1593 (2004)
3. K. Barty, J.-S. Roy, C. Strugarek, Hilbert-valued perturbed subgradient algorithms. *Math. Oper. Res.* **32**(3), 551–562 (2007)
4. J.-P. Bertran, Optimisation stochastique dans un espace de Hilbert. *Méthode de gradient*. C. R. Acad. Sci. Paris Sér. A-B **276**, A613–A616 (1973)
5. D.P. Bertsekas, J.N. Tsitsiklis, Gradient convergence in gradient methods with errors. *SIAM J. Optim.* **10**(3), 627–642 (2000)
6. N. Cahill, B. Jadamba, A.A. Khan, M. Sama, B. Winkler, A first-order adjoint and a second-order hybrid method for an energy output least squares elastography inverse problem of identifying tumor location. *Boundary Value Prob.* **263**, 1–14 (2013)
7. 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.

8. 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)
9. J.-C. Culioli, G. Cohen, Decomposition/coordination algorithms in stochastic optimization. *SIAM J. Control Optim.* **28**(6), 1372–1403 (1990)
10. 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)
11. C. Geiersbach, G.C. Pflug, Projected stochastic gradients for convex constrained problems in Hilbert spaces. *SIAM J. Optim.* **29**(3), 2079–2099 (2019)
12. A. Gibali, B. Jadamba, A.A. Khan, F. Raciti, B. Winkler, Gradient and extragradient methods for the elasticity imaging inverse problem using an equation error formulation: a comparative numerical study, in *Nonlinear Analysis and Optimization*. Contemporary Mathematics, vol. 659 (American Mathematical Society, Providence, 2016), pp. 65–89
13. 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)
14. 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)
15. M.S. Gockenbach, A.A. Khan, An abstract framework for elliptic inverse problems: part I. an output least-squares approach. *Math. Mech. Solids* **12**(3), 259–276 (2007)
16. 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)
17. L. Goldstein, Minimizing noisy functionals in Hilbert space: an extension of the Kiefer–Wolfowitz procedure, *J. Theoret. Probab.* **1**(2), 189–204 (1988)
18. J. Gwinner, B. Jadamba, A.A. Khan, M. Sama, Identification in variational and quasi-variational inequalities. *J. Convex Anal.* **25**(2), 545–569 (2018)
19. J.-B. Hiriart-Urruty, Algorithmes stochastiques de résolution d'équations et d'inéquations variationnelles. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **33**(3), 167–186 (1975/1976)
20. B. Jadamba, R. Kahler, A.A. Khan, F. Raciti, B. Winkler, Identification of flexural rigidity in a Kirchhoff plates model using a convex objective and continuous Newton method. *Math. Probl. Eng.* **2015**, 290301 (2015)
21. 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**(12), 1768–1787 (2017)
22. 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)
23. H. Jiang, H. Xu, Stochastic approximation approaches to the stochastic variational inequality problem. *IEEE Trans. Automat. Control* **53**(6), 1462–1475 (2008)
24. M. Keyanpour, A.M. Nehrani, Optimal thickness of a cylindrical shell subject to stochastic forces. *J. Optim. Theory Appl.* **167**(3), 1032–1050 (2015)
25. J. Kiefer, J. Wolfowitz, Stochastic estimation of the maximum of a regression function. *Ann. Math. Stat.* **23**, 462–466 (1952)
26. H.J. Kushner, A. Shwartz, Stochastic approximation in Hilbert space: identification and optimization of linear continuous parameter systems. *SIAM J. Control Optim.* **23**(5), 774–793 (1985)
27. T.L. Lai, Stochastic approximation. *Ann. Stat.* **31**(2), 391–406 (2003)
28. G.J. Lord, C.E. Powell, T. Shardlow, in *An Introduction to Computational Stochastic PDEs*. Cambridge Texts in Applied Mathematics (Cambridge University Press, New York, 2014)
29. M. Martin, S. Krumrich, F. Nobile, Analysis of stochastic gradient methods for PDE-constrained optimal control problems with uncertain parameters. Preprint 1–39 (2018)
30. M. Morzfeld, X. Tu, J. Wilkening, A.J. Chorin, Parameter estimation by implicit sampling. *Commun. Appl. Math. Comput. Sci.* **10**(2), 205–225 (2015)

31. R. Naseri, A. Malek, Numerical optimal control for problems with random forced SPDE constraints. *ISRN Appl. Math.* 2014, 974305 (2014)
32. H. Robbins, S. Monro, A stochastic approximation method. *Ann. Math. Stat.* **22**, 400–407 (1951)
33. H. Robbins, D. Siegmund, A convergence theorem for non negative almost supermartingales and some applications, in *Optimizing Methods in Statistics* (Academic Press, Cambridge, 1971) pp. 233–257
34. B. V. Rosić, H.G. Matthies, Identification of properties of stochastic elastoplastic systems, in *Computational Methods in Stochastic Dynamics*, vol. 2 (Springer, Dordrecht, 2013), pp. 237–253
35. G.I. Salov, Stochastic approximation in a Hilbert space in the problem of the detection of the appearance of an object in a sequence of noisy images. *Sib. Zh. Ind. Mat.* **12**(1), 127–135 (2009)
36. K. Sepahvand, S. Marburg, On construction of uncertain material parameter using generalized polynomial chaos expansion from experimental data. *Proc. IUTAM* **6**, 4–17 (2013)
37. G. Yin, Y.M. Zhu, On H -valued Robbins–Monro processes. *J. Multivar. Anal.* **34**(1), 116–140 (1990)
38. N. Zabaras, B. Ganapathysubramanian, A scalable framework for the solution of stochastic inverse problems using a sparse grid collocation approach. *J. Comput. Phys.* **227**(9), 4697–4735 (2008)

An Iterative Method for a Common Solution of Split Generalized Equilibrium Problems and Fixed Points of a Finite Family of Nonexpansive Mapping



Ihssane Hay, Abdellah Bnouhachem, and Themistocles M. Rassias

Abstract In this paper, we introduce and analyze a general iterative algorithm for finding an approximate element of the common set of solutions of the split generalized equilibrium problem and the set of common fixed points of a finite family of nonexpansive mapping in the setting of real Hilbert space. Under appropriate conditions, we derive the strong convergence results for this method. Preliminary numerical experiments are included to verify the theoretical assertions of the proposed method. The results presented in this paper extend and improve some well-known results in the literature.

1 Introduction

In the present paper, we always assume that H is a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$, and norm $\| \cdot \|$. Let C be the nonempty closed convex subset of Hilbert space H . Given two bifunctions $F : C \times C \rightarrow H$, and $\varphi : C \times C \rightarrow H$, then the generalized problem is formulated as follows:

$$\begin{cases} \text{Find } x^* \in C \\ F(x^*, x) + \varphi(x^*, x) \geq 0, \quad \forall x \in C, \end{cases} \quad (1)$$

and the solution set of the generalized equilibrium problems is denoted by $GEP(F, \varphi)$.

I. Hay · A. Bnouhachem (✉)

Equipe MAISI, Ibn Zohr University, ENSA, Agadir, Morocco

T. M. Rassias

Department of Mathematics, National Technical University of Athens Zografou Campus, Athens, Greece

e-mail: trassias@math.ntua.gr

© Springer Nature Switzerland AG 2021

T. M. Rassias, P. M. Pardalos (eds.), *Nonlinear Analysis and Global Optimization*,

Springer Optimization and Its Applications 167,

https://doi.org/10.1007/978-3-030-61732-5_10

This class of problems includes numerous important problems such as optimization, classical or mixed equilibrium problems, variational inequalities problems, Nash equilibrium problems, and others, see, for example, [9, 17, 18, 24, 25, 30, 33]. Many algorithms have been proposed and studied in the current literature, see, for example, [5, 10, 19, 23, 28, 29, 31]. Recently, motivated by the Moudafi's work [8], Kazmi and Rivzi [16] introduced and studied a new form of generalized equilibrium problem called the split generalized equilibrium problem formulated as follows: let C and Q be nonempty closed convex subsets of Hilbert spaces H_1 and H_2 , respectively. Given four nonlinear bifunctions $F, \varphi : C \times C \rightarrow H_1$, and $G, \psi : Q \times Q \rightarrow H_2$, and a bounded linear operator $A : H_1 \rightarrow H_2$, then the split generalized equilibrium problem is defined as follows:

$$\begin{cases} \text{Find } x^* \in C \\ F(x^*, x) + \varphi(x^*, x) \geq 0, \quad \forall x \in C, \end{cases} \tag{2}$$

and

$$\begin{cases} \text{Find } y^* = Ax^* \in Q \\ G(y^*, y) + \psi(y^*, y) \geq 0, \quad \forall y \in Q. \end{cases} \tag{3}$$

Inequalities (2) and (3) constitute a pair of the generalized equilibrium problems which aim to find a solution x^* of a generalized equilibrium problem (2) such that its image $y^* = Ax^*$ under a given bounded linear operator A also solves another generalized equilibrium problem (3). The solution set of the split generalized equilibrium problems is denoted by

$$\Omega = \{z \in C; z \in GEP(F, \varphi) \text{ such that } Az \in GEP(G, \psi)\}.$$

This class of problems includes several special cases.

For example, if $\varphi = 0, \psi = 0$, and $G = 0$, it reduces to the classical equilibrium problem [6], which we denote by $EP(F)$ its solution set.

On the other hand, if $\varphi = 0$ and $\psi = 0$, then the split generalized equilibrium problem reduces to the following split equilibrium problem:

$$\begin{cases} \text{Find } x^* \in C \\ F(x^*, x) \geq 0, \quad \forall x \in C, \end{cases} \tag{4}$$

and

$$\begin{cases} \text{Find } y^* = Ax^* \in Q \\ G(y^*, y) \geq 0, \quad \forall y \in Q. \end{cases} \tag{5}$$

The solution set of the split equilibrium problem [13] is denoted by

$$SEP(F, G) = \{z \in C; z \in EP(F) \text{ such that } Az \in EP(G)\}.$$

If $\psi = 0$ and $G = 0$, it reduces to (1). In general, the split generalized equilibrium problem has had a great influence in the development of various branches of pure and applied sciences, and attracted the attention of the majority of authors, such as Deepho et al. [15] who proved a strong convergence to a common solution set of the split generalized equilibrium problems and the set of solutions of the general system of the variational inequality problem for two inverse strongly monotone mappings in real Hilbert spaces. Some strategies have been studied for the split generalized equilibrium problem, for more details, one can refer [14, 15, 20, 22, 27].

Throughout this paper, motivated by several ongoing works in this direction, we present an iterative algorithm to find an approximate element of the common set of solutions of the split generalized equilibrium problem and the set of common fixed points of a finite family of nonexpansive mappings in the setting of real Hilbert spaces. We establish a strong convergence theorem for the sequence generated by the proposed method. In order to verify the theoretical assertions, some numerical examples are given. Our main result presented in this paper is very general, and it extends and improves some well-known results in the literature [26, 27], and others.

2 Preliminaries

Let H_1 and H_2 be two real Hilbert spaces with inner product $\langle \cdot, \cdot \rangle$, and norm $\| \cdot \|$. Let C and Q be nonempty closed convex subsets of Hilbert spaces H_1 and H_2 , respectively.

For every $i \in \{1, \dots, N\}$, let $F_i, \varphi_i : C \times C \rightarrow H_1$, and $G_i, \psi_i : Q \times Q \rightarrow H_2$, be four bifunctions, and $A_i : H_1 \rightarrow H_2$, be a finite family of bounded linear operators.

For each $i \in \{1, \dots, N\}$, the split generalized equilibrium problem (SGEP) is formulated as follows:

$$\begin{cases} \text{Find } x^* \in C \\ F_i(x^*, x) + \varphi_i(x^*, x) \geq 0, \quad \forall x \in C, \end{cases} \tag{6}$$

and

$$\begin{cases} \text{Find } y^* = A_i x^* \in Q \\ G_i(y^*, y) + \psi_i(y^*, y) \geq 0, \quad \forall y \in Q. \end{cases} \tag{7}$$

The solution set of the SGEP is denoted by $\Omega = \{z \in C; z \in GEP(F, \varphi) \text{ such that } A_i z \in GEP(G, \psi)\}$.

Definition 1 Let C be the nonempty closed convex subsets of \mathbb{R}^n , and $v \in \mathbb{R}^n$, then the projection of v onto C is denoted by $P_C(v)$, that is,

$$P_C(v) := \arg \min \{ \|v - u\| \mid u \in C \}. \quad (8)$$

Since C is convex and closed, the projection onto C is unique.

Definition 2 The mapping $T : C \rightarrow H$ is said to be

(a) monotone if

$$\langle Tx - Ty, x - y \rangle \geq 0, \quad \forall x, y \in C;$$

(b) strongly monotone if there exists an $\alpha > 0$ such that

$$\langle Tx - Ty, x - y \rangle \geq \alpha \|x - y\|^2, \quad \forall x, y \in C;$$

(c) α -inverse strongly monotone if there exists an $\alpha > 0$ such that

$$\langle Tx - Ty, x - y \rangle \geq \alpha \|Tx - Ty\|^2, \quad \forall x, y \in C;$$

(d) nonexpansive if

$$\|Tx - Ty\| \leq \|x - y\|, \quad \forall x, y \in C;$$

(e) k -Lipschitz continuous if there exists a constant $k > 0$ such that

$$\|Tx - Ty\| \leq k \|x - y\|, \quad \forall x, y \in C;$$

(f) contraction on C if there exists a constant $0 \leq k < 1$ such that

$$\|Tx - Ty\| \leq k \|x - y\|, \quad \forall x, y \in C.$$

It is well known that every nonexpansive operator $T : H \rightarrow H$ satisfies, for all $(x, y) \in H \times H$, the inequality

$$\langle (x - Tx) - (y - Ty), Ty - Tx \rangle \leq \frac{1}{2} \|(Tx - x) - (Ty - y)\|^2, \quad (9)$$

and, therefore, we get, for all $(x, y) \in H \times F(T)$,

$$\langle x - Tx, y - Tx \rangle \leq \frac{1}{2} \|Tx - x\|^2. \quad (10)$$

Throughout this paper, we always assume that T is a nonexpansive operator on C . The fixed point problem for the mapping T is to find $x \in C$ such that

$$Tx = x. \tag{11}$$

For the recent applications, numerical techniques see [1–4, 7]. The fixed point set of T is denoted by $F(T)$, and it is well known that $F(T)$ is closed and convex (see [32]).

We denote $x_n \rightarrow q$ to symbolize strong convergence of the sequence x_n to q . And we denote $x_n \rightharpoonup q$ to indicate weak convergence of the sequence x_n .

The following results are very useful to prove the convergence of our method.

Lemma 1 *Let H be a real Hilbert space, then*

- (i) $\langle x, y \rangle = \frac{1}{2}(\|x\|^2 + \|y\|^2 - \|x - y\|^2)$,
- (ii) $\|\alpha x + (1 - \alpha)y\|^2 = \alpha\|x\|^2 + (1 - \alpha)\|y\|^2 - \alpha(1 - \alpha)\|x - y\|^2$, hold for all $\alpha \in [0, 1]$ and $x, y \in H$, such that $x \neq y$.

Assumption 2.1 ([9]) *Let $F, \varphi : C \times C \rightarrow \mathbb{R}$ be two bifunctions satisfying the following assumptions:*

- (A₁) $F(x, x) = 0 \quad \forall x \in C$;
- (A₂) F is monotone, i.e., $F(x, y) + F(y, x) \leq 0 \quad \forall x, y \in C$;
- (A₃) F is upper hemicontinuous, i.e.,

$$\forall x, y, z \in C \quad \lim_{t \rightarrow 0} F(tz + (1 - t)x, y) \leq F(x, y);$$

- (A₄) for each $x \in C, y \rightarrow F(x, y)$ is convex and lower semicontinuous.
- (B₁) $\varphi(x, x) \geq 0$, for all $x \in C$;
- (B₂) φ is monotone;
- (B₃) for each $x \in C, y \rightarrow \varphi(x, y)$ is convex and lower semicontinuous;
- (B₄) for each $y \in C, x \rightarrow \varphi(x, y)$ is upper semicontinuous.

Lemma 2 ([11]) *Let C be nonempty closed convex subset of H , and let F, φ be two bifunctions satisfying Assumption 2.1, then for each $x \in H$, for $r > 0$, there exists $z \in C$ such that*

$$F(z, y) + \varphi(z, y) + \frac{1}{r}\langle y - z, z - x \rangle \geq 0 \quad \forall y \in C.$$

Moreover, define a mapping $T_r^{(F, \varphi)} : H \rightarrow C$ as follows:

$$T_r^{(F, \varphi)}(x) = \left\{ z \in C : F(z, y) + \varphi(z, y) + \frac{1}{r}\langle y - z, z - x \rangle \geq 0 \quad \forall y \in C \right\}.$$

Then for all $x \in H$, we have the following:

- (i) $T_r^{(F, \varphi)}$ is single valued;
- (ii) $T_r^{(F, \varphi)}$ is firmly nonexpansive, i.e., for all $x, y \in H$

$$\| T_r^{(F,\varphi)}(x) - T_r^{(F,\varphi)}(y) \|^2 \leq \langle T_r^{(F,\varphi)}(x) - T_r^{(F,\varphi)}(y), x - y \rangle;$$

- (iii) $F(T_r^{(F,\varphi)}) = GEP(F, \varphi)$;
- (iv) $GEP(F, \varphi)$ is compact and convex.

Lemma 3

- (i) If T is nonexpansive, then $I - T$ is 1-inverse strongly monotone;
- (ii) If $T : C \rightarrow C$ is β -inverse strongly monotone, then for all $\lambda \in]0, 2\beta[$, $I - \lambda T$ is nonexpansive.

Lemma 4 ([32]) Let C be a nonempty closed convex subset of a real Hilbert space H . If $T : C \rightarrow C$ is a nonexpansive mapping with $Fix(T) \neq \emptyset$, then the mapping $I - T$ is demiclosed at 0, i.e., if $\{x_n\}$ is a sequence in C and weakly converges to x and if $\{(I - T)x_n\}$ converges strongly to 0, then $(I - T)x = 0$.

Lemma 5 ([12]) Assume $\{a_n\}$ is a sequence of nonnegative real numbers such that

$$a_{n+1} \leq (1 - v_n)a_n + \delta_n,$$

where $\{v_n\}$ is a sequence in $]0, 1[$ and $\{\delta_n\}$ is a sequence such that

- (1) $\sum_{n=1}^{\infty} v_n = \infty$;
- (2) $\sum_{n=1}^{\infty} \delta_n < \infty$.

Then, $\lim_{n \rightarrow \infty} a_n = 0$.

3 The Proposed Method and Some Properties

In this section, we suggest and analyze an iterative method for finding an approximate element of the common set of solutions of the split generalized equilibrium problem and the set of common fixed points of a finite family of nonexpansive mappings.

Let H_1 and H_2 be the two real Hilbert spaces. Let C (respectively, Q) be the nonempty closed convex subset of H_1 (respectively, H_2). Let $F_i, \varphi_i : C \times C \rightarrow H_1$, and $G_i, \psi_i : Q \times Q \rightarrow H_2$, be the four finite family bifunctions satisfying Assumption 2.1 such that G_i, ψ_i is upper semicontinuous in the first argument. Let $A_i : H_1 \rightarrow H_2$, be a finite family of bounded linear operators, and let $S_i : C \rightarrow C$, be a finite family of nonexpansive mappings. Setting $\Gamma = \left(\bigcap_{i=1}^N F(S_i) \right) \cap \Omega$, where $\Omega = \left\{ p \in C : p \in \bigcap_{i=1}^N GEP(F_i, \varphi_i) \text{ and } A_i p \in GEP(G_i, \psi_i), \text{ for all } i \in \{1, \dots, N\} \right\}$.

Algorithm 3.1 For a given $x_1 \in C_1 = C$, arbitrarily, let the iterative sequences $u_{n,i}, y_{n,i}$ and x_n be generated by the iterative algorithm

$$\begin{cases} u_{n,i} = T_{r_{n,i}}^{(F_i, \varphi_i)}(I - \gamma A_i^*(I - T_{s_{n,i}}^{(G_i, \psi_i)})A_i)x_n, \\ y_{n,i} = \alpha_{n,i}S_i x_n + (1 - \alpha_{n,i})u_{n,i}, \\ C_{n+1} = \{p \in C_n : \|y_{n,i} - p\| \leq \|x_n - p\|\} \\ x_{n+1} = \gamma_n P_{C_{n+1}}x_1 + (1 - \gamma_n)y_{n,i}, \quad n \geq 1. \end{cases} \tag{12}$$

Let $\gamma \in]0, \frac{1}{L}]$, where $L = \max(L_1, L_2, \dots, L_N)$ such that L_i is the spectral radius of the operator $A_i^*A_i$, where A_i^* is the adjoint of $A_i, \forall i \in \{1, 2, \dots, N\}$. Let $s_{n,i}$ and $r_{n,i}$ be two positive real sequences, and let $\alpha_{n,i}$ and γ_n be the two sequences in $]0, 1[$, satisfying the following conditions:

- (C1) $0 < a \leq \alpha_{n,i}, \gamma_n \leq b < 1$;
- (C2) $\lim_{n \rightarrow \infty} \alpha_{n,i} = 0$;
- (C3) $\lim_{n \rightarrow \infty} \gamma_n = 0$ and $\sum_{n=1}^{\infty} \gamma_n = \infty$;
- (C4) $\sum_{n=1}^{\infty} |\gamma_{n+1} - \gamma_n| < \infty$ and $\sum_{n=1}^{\infty} |\alpha_{n+1,i} - \alpha_{n,i}| < \infty$;
- (C5) $\liminf_{n \rightarrow \infty} r_{n,i} > 0$ and $\limsup_{n \rightarrow \infty} s_{n,i} > 0$.

Lemma 6 *Let $\{x_n\}$ be the sequence generated by Algorithm 3.1. Then*

- (a) $\{x_n\}$ is well defined for every $n \in \mathbb{N}^*$ and bounded,
- (b) $\Gamma \subset C_{n+1}$.

Proof We show that the sequence $\{x_n\}$ is well defined for every $n \in \mathbb{N}^*$.

To prove that, we will show that \mathcal{C}_n is a closed convex subset for all $n \geq 1$.

Clearly, $\mathcal{C}_1 = \mathcal{C}$ is closed convex. Suppose that \mathcal{C}_k is closed convex for $k \geq 1$; we prove that so is C_{k+1} .

Let $p_m \in C_{k+1} \subset C_k$ such that $p_m \rightarrow p$ then $p \in C_k$ (because C_k is closed); thus, $\|y_{k,i} - p_m\| \leq \|x_k - p_m\|$, which implies that

$$\begin{aligned} \|y_{k,i} - p\| &\leq \|y_{k,i} - p_m\| + \|p_m - p\| \\ &\leq \|x_k - p_m\| + \|p_m - p\|. \end{aligned}$$

Taking $\lim_{m \rightarrow \infty}$ on both sides of the above estimate, we get

$$\begin{aligned} \lim_{m \rightarrow \infty} \|y_{k,i} - p\| &= \|y_{k,i} - p\| \leq \lim_{m \rightarrow \infty} (\|x_k - p_m\| + \|p_m - p\|) \\ &\leq \|x_k - p\|. \end{aligned}$$

Then $p \in C_{k+1}$, and it follows that C_{k+1} is closed.

Now we set $p = \lambda x + (1 - \lambda)y$, for every $x, y \in C_{k+1}$ and $\lambda \in [0, 1]$, then $p \in C_k$ (because C_k is convex).

By using Lemma 1, we have

$$\begin{aligned}
 \|y_{k,i} - p\|^2 &= \|y_{k,i} - \lambda x - (1 - \lambda)y\|^2 \\
 &= \|\lambda(y_{k,i} - x) + (1 - \lambda)(y_{k,i} - y)\|^2 \\
 &= \lambda \|y_{k,i} - x\|^2 + (1 - \lambda) \|y_{k,i} - y\|^2 - \lambda(1 - \lambda) \|y - x\|^2 \\
 &\leq \lambda \|x_k - x\|^2 + (1 - \lambda) \|x_k - y\|^2 - \lambda(1 - \lambda) \|y - x_k + x_k - x\|^2 \\
 &= \|\lambda(x_k - x) + (1 - \lambda)(x_k - y)\|^2 \\
 &= \|x_k - p\|^2,
 \end{aligned}$$

thus $p \in C_{k+1}$ then C_{k+1} is convex. Therefore, C_n is closed convex for all $n \geq 1$.

Since $P_{C_{n+1}x_1}$ is well defined for every $x_1 \in C$, x_n is well defined.

Obviously, $\Gamma \subset C_1$. If $p \in \Gamma$, we have $p = T_{r_n,i}^{F_i,\varphi_i} p$ and $(I - \gamma A_i^*(I - T_{s_n,i}^{(G_i,\psi_i)})A_i)p = p$, then $p \in C = C_1$.

Assume that $\Gamma \subset C_k$, and we show that $\Gamma \subset C_{k+1}$.

We have

$$\begin{aligned}
 \|A_i^*(I - T_{s_n,i}^{(G_i,\psi_i)})A_i x - A_i^*(I - T_{s_n,i}^{(G_i,\psi_i)})A_i y\|^2 &= \|A_i^*(I - T_{s_n,i}^{(G_i,\psi_i)})(A_i x - A_i y)\|^2 \\
 &= \langle A_i^*(I - T_{s_n,i}^{(G_i,\psi_i)})(A_i x - A_i y), \\
 &\quad A_i^*(I - T_{s_n,i}^{(G_i,\psi_i)})(A_i x - A_i y) \rangle \\
 &= \langle (I - T_{s_n,i}^{(G_i,\psi_i)})(A_i x - A_i y), \\
 &\quad A_i A_i^*(I - T_{s_n,i}^{(G_i,\psi_i)})(A_i x - A_i y) \rangle \\
 &\leq L \| (I - T_{s_n,i}^{(G_i,\psi_i)})(A_i x - A_i y) \|^2.
 \end{aligned}$$

Then

$$\begin{aligned}
 &\| (I - T_{s_n,i}^{(G_i,\psi_i)})(A_i x - A_i y) \|^2 \\
 &\geq \frac{1}{L} \| A_i^*(I - T_{s_n,i}^{(G_i,\psi_i)})A_i x - A_i^*(I - T_{s_n,i}^{(G_i,\psi_i)})A_i y \|^2. \tag{13}
 \end{aligned}$$

Since T is nonexpansive, it follows from Lemma 3 that $I - T$ is 1-inverse strongly monotone, then

$$\begin{aligned}
 \| (I - T_{s_n,i}^{(G_i,\psi_i)})(A_i x - A_i y) \|^2 &\leq \langle (I - T_{s_n,i}^{(G_i,\psi_i)})(A_i x - A_i y), (A_i x - A_i y) \rangle \\
 &= \langle A_i^*(I - T_{s_n,i}^{(G_i,\psi_i)})(A_i x - A_i y), x - y \rangle, \tag{14}
 \end{aligned}$$

and hence using (13) and (14), we obtain

$$\langle A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)})(A_i x - A_i y), x - y \rangle \geq \frac{1}{L} \| A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)})(A_i x - A_i y) \|^2;$$

this implies that

$$\begin{aligned} & \langle A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x - A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i y, x - y \rangle \\ & \geq \frac{1}{L} \| A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x - A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i y \|^2 \end{aligned}$$

thus $A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i$ is $\frac{1}{L}$ -inverse strongly monotone, and by using Lemma 3, we get $I - \gamma A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i$, is nonexpansive for each $\gamma \in]0, \frac{1}{L}[$. Therefore, we obtain

$$\begin{aligned} \| u_{n,i} - p \| &= \| T_{r_{n,i}}^{(F_i, \varphi_i)}(I - \gamma A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i)x_n \\ &\quad - T_{r_{n,i}}^{(F_i, \varphi_i)}(I - \gamma A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i)p \| \\ &\leq \| x_n - p \| . \end{aligned} \tag{15}$$

Let $p \in \Gamma$, then the following results can be immediately obtained from Lemma 1 ii), nonexpansiveness of S_i , and (15)

$$\begin{aligned} \| y_{k,i} - p \|^2 &= \| \alpha_{k,i} S_i x_k + (1 - \alpha_{k,i})u_{k,i} - p \|^2 \\ &= \alpha_{k,i} \| S_i x_k - p \|^2 + (1 - \alpha_{k,i}) \| u_{k,i} - p \|^2 - \alpha_{k,i}(1 - \alpha_{k,i}) \| S_i x_k - u_{k,i} \|^2 \\ &= \alpha_{k,i} \| S_i x_k - S_i p \|^2 + (1 - \alpha_{k,i}) \| u_{k,i} - p \|^2 - \alpha_{k,i}(1 - \alpha_{k,i}) \| S_i x_k - u_{k,i} \|^2 \end{aligned} \tag{16}$$

$$\begin{aligned} &\leq \alpha_{k,i} \| x_k - p \|^2 + (1 - \alpha_{k,i}) \| u_{k,i} - p \|^2 \\ &\leq \alpha_{k,i} \| x_k - p \|^2 + (1 - \alpha_{k,i}) \| x_k - p \|^2 \\ &= \| x_k - p \|^2, \end{aligned} \tag{17}$$

and then we get

$$\| y_{k,i} - p \| \leq \| x_k - p \|, \tag{18}$$

consequently $p \in C_{k+1}$, thus $\Gamma \in C_{k+1}$. Hence for every $n \in \mathbb{N}^*$, the following inclusion $\Gamma \subset C_{n+1}$ is always satisfied.

Next, we show that the sequence x_n is bounded.

Note that $\| P_{C_{n+1}}x_1 - x_1 \|^2 \leq \| x^* - x_1 \|^2$ for all $x^* \in C_{n+1}$. In particular, we have $\| P_{C_{n+1}}x_1 - x_1 \|^2 \leq \| P_{\Gamma}x_1 - x_1 \|^2$. Then, we get

$$\begin{aligned} \| x_{n+1} - x_1 \|^2 &= \| \gamma_n(P_{C_{n+1}}x_1 - x_1) + (1 - \gamma_n)(y_{n,i} - x_1) \|^2 \\ &\leq \gamma_n \| P_{C_{n+1}}x_1 - x_1 \|^2 + (1 - \gamma_n) \| y_{n,i} - x_1 \|^2 \end{aligned}$$

$$\begin{aligned}
&\leq \gamma_n \| P_{\Gamma}x_1 - x_1 \|^2 + (1 - \gamma_n) \| x_n - x_1 \|^2 \\
&\leq \max\{\| P_{\Gamma}x_1 - x_1 \|^2, \| x_n - x_1 \|^2\} \\
&\leq \max\{\| P_{\Gamma}x_1 - x_1 \|^2, \| x_{n-1} - x_1 \|^2\} \\
&\quad \vdots \\
&\leq \max\{\| P_{\Gamma}x_1 - x_1 \|^2, \| x_1 - x_1 \|^2\} \\
&= \| P_{\Gamma}x_1 - x_1 \|^2 .
\end{aligned}$$

Therefore, it follows from the above inequalities that $\| x_{n+1} - x_1 \| < \infty$, hence $\{x_n\}$ is bounded, and so are $\{y_{n,i}\}$ and $\{u_{n,i}\}$. \square

Lemma 7 *Let $\{x_n\}$ be the sequence generated by Algorithm 3.1. Then, for every $i \in \{1, \dots, N\}$, we have*

- (a) $\lim_{n \rightarrow \infty} \| x_{n+1} - x_n \| = 0$;
- (b) $\lim_{n \rightarrow \infty} \| y_{n,i} - x_n \| = 0$;
- (c) $\lim_{n \rightarrow \infty} \| (I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n \| = 0$;
- (d) $\lim_{n \rightarrow \infty} \| u_{n,i} - x_n \| = 0$;
- (e) $\lim_{n \rightarrow \infty} \| S_i x_n - u_{n,i} \| = 0$.

Proof We have

$$\begin{aligned}
&\| y_{n,i} - y_{n-1,i} \| = \| \alpha_{n,i} S_i x_n + (1 - \alpha_{n,i}) u_{n,i} - \alpha_{n-1,i} S_i x_{n-1} - (1 - \alpha_{n-1,i}) u_{n-1,i} \| \\
&= \| \alpha_{n,i} (S_i x_n - S_i x_{n-1}) + (\alpha_{n,i} - \alpha_{n-1,i}) S_i x_{n-1} + (1 - \alpha_{n,i}) (u_{n,i} - u_{n-1,i}) \\
&\quad + (\alpha_{n-1,i} - \alpha_{n,i}) u_{n-1,i} \| \\
&\leq \alpha_{n,i} \| S_i x_n - S_i x_{n-1} \| + |\alpha_{n,i} - \alpha_{n-1,i}| (\| S_i x_{n-1} \| + \| u_{n-1,i} \|) \\
&\quad + (1 - \alpha_{n,i}) \| u_{n,i} - u_{n-1,i} \| \\
&\leq \alpha_{n,i} \| x_n - x_{n-1} \| + |\alpha_{n,i} - \alpha_{n-1,i}| (\| S_i x_{n-1} \| + \| u_{n-1,i} \|) \\
&\quad + (1 - \alpha_{n,i}) (\| u_{n,i} \| + \| u_{n-1,i} \|). \tag{19}
\end{aligned}$$

By using (19) and condition C1), we obtain

$$\begin{aligned}
&\| x_{n+1} - x_n \| \\
&= \| \gamma_n (P_{C_{n+1}} x_1 - P_{C_n} x_1) + (\gamma_n - \gamma_{n-1}) P_{C_n} x_1 + \gamma_n (y_{n-1,i} - y_{n,i}) + (\gamma_{n-1} - \gamma_n) y_{n-1,i} \\
&\quad + y_{n,i} - y_{n-1,i} \| \\
&\leq \gamma_n \| P_{C_{n+1}} x_1 - P_{C_n} x_1 \| + |\gamma_n - \gamma_{n-1}| (\| P_{C_n} x_1 \| + \| y_{n-1,i} \|) \\
&\quad + (1 - \gamma_n) \| y_{n,i} - y_{n-1,i} \|
\end{aligned}$$

$$\begin{aligned}
 &\leq \gamma_n \| P_{C_{n+1}}x_1 - P_{C_n}x_1 \| + |\gamma_n - \gamma_{n-1}|(\| P_{C_n}x_1 \| + \| y_{n-1,i} \|) \\
 &\quad + (1 - \gamma_n) \left(\alpha_{n,i} \| x_n - x_{n-1} \| + |\alpha_{n,i} - \alpha_{n-1,i}|(\| S_i x_{n-1} \| + \| u_{n-1,i} \|) \right. \\
 &\quad \left. + (1 - \alpha_{n,i})(\| u_{n,i} \| + \| u_{n-1,i} \|) \right) \\
 &\leq \| P_{C_{n+1}}x_1 - P_{C_n}x_1 \| + |\gamma_n - \gamma_{n-1}|(\| P_{C_n}x_1 \| + \| y_{n-1,i} \|) \\
 &\quad + (1 - \gamma_n) \left(\| x_n - x_{n-1} \| + |\alpha_{n,i} - \alpha_{n-1,i}|(\| S_i x_{n-1} \| + \| u_{n-1,i} \|) \right. \\
 &\quad \left. + \| u_{n,i} \| + \| u_{n-1,i} \| \right) \\
 &\leq (1 - \gamma_n) \| x_n - x_{n-1} \| + (\| P_{C_{n+1}}x_1 \| + \| P_{C_n}x_1 \|) \\
 &\quad + |\gamma_n - \gamma_{n-1}|(\| P_{C_n}x_1 \| + \| y_{n-1,i} \|) + |\alpha_{n,i} - \alpha_{n-1,i}|(\| S_i x_{n-1} \| + \| u_{n-1,i} \|) \\
 &\quad + \| u_{n,i} \| + \| u_{n-1,i} \| \\
 &\leq (1 - \gamma_n) \| x_n - x_{n-1} \| + M(1 + |\gamma_n - \gamma_{n-1}| + |\alpha_{n,i} - \alpha_{n-1,i}| + 1),
 \end{aligned}$$

where $M = \max\{ \sup_{n \geq 1}(\| P_{C_{n+1}}x_1 \| + \| P_{C_n}x_1 \|), \sup_{n \geq 1}(\| P_{C_n}x_1 \| + \| y_{n-1,i} \|), \sup_{n \geq 1}(\| S_i x_{n-1} \| + \| u_{n-1,i} \|), \sup_{n \geq 1}(\| u_{n,i} \| + \| u_{n-1,i} \|) \}$; setting $\delta_n = M(2 + |\gamma_n - \gamma_{n-1}| + |\alpha_{n,i} - \alpha_{n-1,i}|)$, by using condition C4), it follows that $\sum_{n=1}^{\infty} \delta_n < \infty$, and by condition C3), we get $\sum_{n=1}^{\infty} \gamma_n = \infty$. Hence from Lemma 5, we conclude $\lim_{n \rightarrow \infty} \| x_{n+1} - x_n \| = 0$, which proves the result (a).

On the other hand, we have

$$\begin{aligned}
 \| y_{n,i} - x_n \| &\leq \| y_{n,i} - x_{n+1} \| + \| x_{n+1} - x_n \| \\
 &= \| y_{n,i} - \gamma_n P_{C_{n+1}}x_1 - (1 - \gamma_n)y_{n,i} \| + \| x_{n+1} - x_n \| \\
 &\leq \gamma_n \| y_{n,i} - P_{C_{n+1}}x_1 \| + \| x_{n+1} - x_n \| \\
 &\leq \gamma_n \| x_n - P_{C_{n+1}}x_1 \| + \| x_{n+1} - x_n \| \\
 &\leq \gamma_n \| x_n - P_{\Gamma}x_1 \| + \| x_{n+1} - x_n \| \\
 &\leq \gamma_n(\| x_n \| + \| P_{\Gamma}x_1 \|) + \| x_{n+1} - x_n \|.
 \end{aligned}$$

This implies by condition (C3), and (a), that $\lim_{n \rightarrow \infty} \| y_{n,i} - x_n \| = 0$; thus (b) is proved.

Next, we show the assertion (c). Observe that

$$\begin{aligned}
 &\| u_{n,i} - p \|^2 \\
 &= \| T_{r_{n,i}}^{(F_i, \varphi_i)}(I - \gamma A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i)x_n - T_{r_{n,i}}^{(F_i, \varphi_i)}(I - \gamma A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i)p \|^2 \\
 &\leq \| x_n - p - \gamma A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n \|^2 \\
 &= \| x_n - p \|^2 + \gamma^2 \| A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n \|^2 - 2\gamma \langle x_n - p, A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n \rangle \\
 &= \| x_n - p \|^2 + \gamma^2 \langle (I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n, A_i A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n \rangle
 \end{aligned}$$

$$\begin{aligned}
 & -2\gamma \langle A_i(x_n - p), (I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n \rangle \\
 & \leq \|x_n - p\|^2 + \gamma^2 L \|(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n\|^2 \\
 & \quad - 2\gamma \langle A_i(x_n - p) + (I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n - (I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n, (I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n \rangle \\
 & = \|x_n - p\|^2 + \gamma^2 L \|(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n\|^2 - 2\gamma \|(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n\|^2 \\
 & \quad - 2\gamma \langle T_{S_{n,i}}^{(G_i, \psi_i)}A_i x_n - A_i p, (I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n \rangle.
 \end{aligned}$$

Applying (10), we get

$$\begin{aligned}
 & \|u_{n,i} - p\|^2 \\
 & \leq \|x_n - p\|^2 + \gamma^2 L \|(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n\|^2 - 2\gamma \|(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n\|^2 \\
 & \quad + 2\gamma \frac{1}{2} \|(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n\|^2 \\
 & = \|x_n - p\|^2 + \gamma(\gamma L - 1) \|(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n\|^2. \tag{20}
 \end{aligned}$$

Further, observe that

$$\begin{aligned}
 & \|y_{n,i} - p\|^2 \\
 & = \|\alpha_{n,i} S_i x_n + (1 - \alpha_{n,i})u_{n,i} - p\|^2 \\
 & = \|\alpha_{n,i}(S_i x_n - S_i p) + (1 - \alpha_{n,i})(u_{n,i} - p)\|^2 \\
 & = \alpha_{n,i} \|S_i x_n - S_i p\|^2 + (1 - \alpha_{n,i}) \|u_{n,i} - p\|^2 - \alpha_{n,i}(1 - \alpha_{n,i}) \|S_i x_n - u_{n,i}\|^2 \\
 & \leq \alpha_{n,i} \|x_n - p\|^2 + (1 - \alpha_{n,i}) \|u_{n,i} - p\|^2 \\
 & \leq \alpha_{n,i} \|x_n - p\|^2 + (1 - \alpha_{n,i})(\|x_n - p\|^2 + \gamma(\gamma L - 1) \|(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n\|^2) \\
 & = \|x_n - p\|^2 + (1 - \alpha_{n,i})\gamma(\gamma L - 1) \|(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n\|^2.
 \end{aligned}$$

Hence

$$\begin{aligned}
 & (1 - \alpha_{n,i})(\gamma(1 - \gamma L) \|(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n\|^2) \leq \|x_n - p\|^2 - \|y_{n,i} - p\|^2 \\
 & = \|x_n - y_{n,i}\| (\|x_n - p\| + \|y_{n,i} - p\|).
 \end{aligned}$$

Since $\gamma(1 - \gamma L) > 0$, using (b) and condition C2), and by letting $n \rightarrow \infty$, we obtain the desired result.

Next, we show the assertion (d). From Lemma 2 (ii), for every $p \in \Gamma$, we obtain

$$\begin{aligned}
 \|u_{n,i} - p\|^2 & = \|T_{r_{n,i}}^{(F_i, \varphi_i)}(x_n - \gamma A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n) - T_{r_{n,i}}^{(F_i, \varphi_i)} p\|^2 \\
 & \leq \langle u_{n,i} - p, x_n - p - \gamma A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)})A_i x_n \rangle
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{2} \left(\| u_{n,i} - p \|^2 + \| x_n - p - \gamma A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)}) A_i x_n \|^2 \right. \\
 &\quad \left. - \| u_{n,i} - x_n + \gamma A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)}) A_i x_n \|^2 \right) \\
 &= \frac{1}{2} \left(\| u_{n,i} - p \|^2 + \| x_n - p \|^2 - 2\gamma \langle A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)}) A_i x_n, u_{n,i} - p \rangle - \| u_{n,i} - x_n \|^2 \right).
 \end{aligned}$$

Then

$$\| u_{n,i} - p \|^2 \leq \| x_n - p \|^2 - \| u_{n,i} - x_n \|^2 - 2\gamma \langle A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)}) A_i x_n, u_{n,i} - p \rangle. \tag{21}$$

Substituting (21) in (17), we obtain

$$\begin{aligned}
 \| y_{n,i} - p \|^2 &\leq \alpha_{n,i} \| x_n - p \|^2 + (1 - \alpha_{n,i}) \left(\| x_n - p \|^2 - \| u_{n,i} - x_n \|^2 \right. \\
 &\quad \left. - 2\gamma \langle A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)}) A_i x_n, u_{n,i} - p \rangle \right) \\
 &= \| x_n - p \|^2 - (1 - \alpha_{n,i}) \| u_{n,i} - x_n \|^2 \\
 &\quad - 2\gamma (1 - \alpha_{n,i}) \langle A_i^*(I - T_{S_{n,i}}^{(G_i, \psi_i)}) A_i x_n, u_{n,i} - p \rangle.
 \end{aligned}$$

Furthermore,

$$\begin{aligned}
 (1 - \alpha_{n,i}) \| u_{n,i} - x_n \|^2 &\leq \| x_n - p \|^2 - \| y_{n,i} - p \|^2 \\
 &\quad + 2\gamma (1 - \alpha_{n,i}) \langle (I - T_{S_{n,i}}^{(G_i, \psi_i)}) A_i x_n, A_i(p - u_{n,i}) \rangle \\
 &\leq (\| x_n - p \| + \| y_{n,i} - p \|) \| y_{n,i} - x_n \| \\
 &\quad + 2\gamma (1 - \alpha_{n,i}) \| (I - T_{S_{n,i}}^{(G_i, \psi_i)}) A_i x_n \| \| A_i(p - u_{n,i}) \|.
 \end{aligned}$$

Letting $n \rightarrow \infty$ and using (b), (c), and both conditions (C1) and (C2), then we get the desired result, i.e., $\lim_{n \rightarrow \infty} \| u_{n,i} - x_n \| = 0$.

Next, we show the assertion (e). Let $p \in \Gamma$, it follows from (15) and (16) that

$$\begin{aligned}
 &\| y_{n,i} - p \|^2 \\
 &\leq \alpha_{n,i} \| x_n - p \|^2 + (1 - \alpha_{n,i}) \| u_{n,i} - p \|^2 - \alpha_{n,i} (1 - \alpha_{n,i}) \| S_i x_n - u_{n,i} \|^2 \\
 &\leq \| x_n - p \|^2 - \alpha_{n,i} (1 - \alpha_{n,i}) \| S_i x_n - u_{n,i} \|^2. \tag{22}
 \end{aligned}$$

Consequently,

$$\begin{aligned}
 \alpha_{n,i} (1 - \alpha_{n,i}) \| S_i x_n - u_{n,i} \|^2 &\leq \| x_n - p \|^2 - \| y_{n,i} - p \|^2 \\
 &= (\| x_n - p \| + \| y_{n,i} - p \|) \| x_n - y_{n,i} \|.
 \end{aligned}$$

By condition (C1), it can be easily seen that

$$a(1 - b) \| S_i x_n - u_{n,i} \|^2 \leq (\| x_n - p \| + \| y_{n,i} - p \|) \| x_n - y_{n,i} \| .$$

Using (b), we conclude that $\lim_{n \rightarrow \infty} \| S_i x_n - u_{n,i} \| = 0$. □

Theorem 1 *The sequence $\{x_n\}$ generated by Algorithm 3.1 converges strongly to $q \in \Gamma$, where $\Gamma = \left(\bigcap_{i=1}^N F(S_i) \right) \cap \Omega$, and $\Omega = \left\{ p \in C : p \in \bigcap_{i=1}^N GEP(F_i, \varphi_i) \text{ and } A_i p \in GEP(G_i, \psi_i), \text{ for all } i \in \{1, \dots, N\} \right\}$.*

Proof Since $\{x_n\}$ is bounded, then there exist a subsequence x_{n_j} such that $x_{n_j} \rightharpoonup q$. Next, we show that $q \in \bigcap_{i=1}^N F(S_i)$. Since for all $i \in \{1, \dots, N\}$, we have

$\| S_i x_{n_j} - x_{n_j} \| \leq \| S_i x_{n_j} - u_{n_j,i} \| + \| u_{n_j,i} - x_{n_j} \|$. It follows from (d) and (e) that $\lim_{j \rightarrow \infty} \| S_i x_{n_j} - x_{n_j} \| = 0$. Using Lemma 4, we get $S_i q - q = 0$, then $q \in F(S_i)$ for every $i \in \{1, \dots, N\}$. Hence, $q \in \bigcap_{i=1}^N F(S_i)$.

Furthermore, we show that $q \in \Omega = \left\{ p \in C : p \in \bigcap_{i=1}^N GEP(F_i) \text{ and } A_i p \in GEP(G_i) \text{ for all } i \text{ be } [1, N] \right\}$. First, we will show that for every $i \in \{1, \dots, N\}$, we have $p \in \bigcap_{i=1}^N GEP(F_i)$. For all $i \in \{1, \dots, N\}$, we have $u_{n,i} = T_{r_{n,i}}^{(F_i, \varphi_i)} (I - \gamma A_i^* (I - T_{s_{n,i}}^{(G_i, \psi_i)}) A_i) x_n, \quad n \geq 1$.

Then

$$F_i(u_{n,i}, y) + \varphi_i(u_{n,i}, y) + \frac{1}{r_{n,i}} \langle y - u_{n,i}, u_{n,i} - x_n + \gamma A_i^* (I - T_{s_{n,i}}^{(G_i, \psi_i)}) A_i x_n \rangle \geq 0 \quad \forall y \in C,$$

which implies

$$-F_i(u_{n,i}, y) - \varphi_i(u_{n,i}, y) \leq \frac{1}{r_{n,i}} \left(\langle y - u_{n,i}, u_{n,i} - x_n \rangle + \gamma \langle y - u_{n,i}, A_i^* (I - T_{s_{n,i}}^{(G_i, \psi_i)}) A_i x_n \rangle \right).$$

By using the monotonicity of F_i and φ_i , we can write the last inequality as follows:

$$F_i(y, u_{n,i}) + \varphi_i(y, u_{n,i}) \leq \frac{1}{r_{n,i}} \left(\langle y - u_{n,i}, u_{n,i} - x_n \rangle + \gamma \langle y - u_{n,i}, A_i^* (I - T_{s_{n,i}}^{(G_i, \psi_i)}) A_i x_n \rangle \right).$$

Therefore,

$$\begin{aligned} & F_i(y, u_{n_j,i}) + \varphi_i(y, u_{n_j,i}) \\ & \leq \frac{1}{r_{n_j,i}} \left(\| y - u_{n_j,i} \| \| u_{n_j,i} - x_{n_j} \| + \gamma \| y - u_{n_j,i} \| \| A_i^* (I - T_{s_{n_j,i}}^{(G_i, \psi_i)}) A_i x_{n_j} \| \right). \end{aligned}$$

It follows from (c), (d), and the condition (C5) that for every $i \in \{1, \dots, N\}$, $\lim_{j \rightarrow \infty} F_i(y, u_{n_j,i}) + \varphi_i(y, u_{n_j,i}) \leq 0$, and since F_i and φ_i are lower semicontinuous,

we get

$$\forall i \in \{1, \dots, N\}, \quad F_i(y, q) + \varphi_i(y, q) \leq 0, \quad \forall y \in C. \tag{23}$$

Setting $y_t = ty + (1 - t)q$ for some $0 < t < 1$, then $y_t \in C$. Since F_i satisfies (A_1) – (A_4) and φ_i satisfies (B_1) – (B_4) , it follows from (23) that

$$\begin{aligned} 0 &\leq F_i(y_t, y_t) + \varphi_i(y_t, y_t) \\ &= F_i(y_t, ty + (1 - t)q) + \varphi_i(y_t, ty + (1 - t)q) \\ &\leq t(F_i(y_t, y) + \varphi_i(y_t, y)) + (1 - t)(F_i(y_t, q) + \varphi_i(y_t, q)) \\ &\leq t(F_i(y_t, y) + \varphi_i(y_t, y)), \end{aligned}$$

which implies that for all $i \in \{1, \dots, N\}$, $F_i(ty+(1-t)q, y)+\varphi_i(ty+(1-t)q, y) \geq 0$. Letting $t \rightarrow 0_+$, we obtain that $F_i(q, y) + \varphi_i(q, y) \geq 0, \forall i \in \{1, \dots, N\}$. Then $q \in GEP(F_i, \varphi_i), \forall i \in \{1, \dots, N\}$; therefore, $q \in \bigcap_{i=1}^N GEP(F_i, \varphi_i)$.

Next, we show that $A_iq \in GEP(G_i, \psi_i), \forall i \in \{1, \dots, N\}$.

For all $i \in \{1, \dots, N\}$, we have A_i that is bounded, and since $x_{n_j} \rightharpoonup q$ then $A_i x_{n_j} \rightharpoonup A_i q$, it follows from (c) that $T_{s_{n_j,i}}^{(G_i, \psi_i)} A_i x_{n_j} \rightharpoonup A_i q$. From Lemma 2, we have

$$\begin{aligned} &G_i(T_{s_{n_j,i}}^{(G_i, \psi_i)} A_i x_{n_j}, z) + \psi_i(T_{s_{n_j,i}}^{(G_i, \psi_i)} A_i x_{n_j}, z) \\ &+ \frac{1}{s_{n_j,i}} \left\langle z - T_{s_{n_j,i}}^{(G_i, \psi_i)} A_i x_{n_j}, T_{s_{n_j,i}}^{(G_i, \psi_i)} A_i x_{n_j} - A_i x_{n_j} \right\rangle \geq 0 \quad \forall z \in Q. \end{aligned}$$

Taking the limit sup on both sides of the above inequality, and using the fact that G_i, ψ_i are upper semicontinuous in the first argument and using the condition (C5), we conclude that for every $i \in \{1, \dots, N\}$,

$$G_i(A_iq, z) + \psi_i(A_iq, z) \geq 0 \quad \forall z \in Q,$$

which implies that $A_iq \in GEP(G_i, \psi_i)$ and hence, $q \in \Omega$. Thus, we have $q \in \Gamma$.

Next, we show that $\{x_{n_j}\}$ converges strongly to $q \in \Gamma$. It follows from (18) that

$$\begin{aligned} \|x_{n+1} - q\|^2 &= \|\gamma_n(P_{C_{n+1}}x_1 - q) + (1 - \gamma_n)(y_{n,i} - q)\|^2 \\ &\leq \gamma_n \|P_{C_{n+1}}x_1 - q\|^2 + (1 - \gamma_n) \|y_{n,i} - q\|^2 \\ &\leq \|P_{C_{n+1}}x_1 - q\|^2 + (1 - \gamma_n) \|x_n - q\|^2 \\ &= (1 - \gamma_n) \|x_n - q\|^2 + \delta_n, \end{aligned}$$

where $\delta_n = \|P_{C_{n+1}}x_1 - q\|^2$. Since

$$\begin{cases} \sum_{n=1}^{\infty} \gamma_n = \infty; \\ \sum_{n=1}^{\infty} \delta_n = \sum_{n=1}^{\infty} \| P_{C_{n+1}}x_1 - q \|^2 < \infty. \end{cases}$$

Thus all the conditions of Lemma 5 are satisfied. Hence, we deduce that $\lim_{n \rightarrow \infty} x_n = q \in \Gamma$. And the conclusion of this theorem is proved. \square

Remark 1 Our method is quite general, and it can be considered as an improvement of several existing algorithms [15, 21, 26, 27].

4 Numerical Example

In order to analyze and better understand the convergence of the new algorithm, we present in this section a numerical example to prove the performance of our theoretical results. All codes were written in Matlab.

Let $H_1 = H_2 = \mathbb{R}$, and let $C = [0, 20]$ and $Q =]-\infty, 0]$ be the two closed convex subsets of \mathbb{R} .

For all $i \in \{1, \dots, N\}$, we define A_i and S_i as follows:

$$\begin{aligned} A_i : H_1 &\longrightarrow H_1 \\ x &\longrightarrow 3x \end{aligned}$$

and

$$\begin{aligned} S_i : C &\longrightarrow C \\ x &\longrightarrow \frac{x}{10i} \end{aligned}$$

and it is obvious that A_i is linear bounded and S_i is nonexpansive. Since $S_i(0) = 0$, then $F(S_i) = 0$.

We define the bifunctions F_i and G_i by

$$\begin{aligned} F_i : C \times C &\longrightarrow H_1 \\ u, v &\longrightarrow F_i(u, v) = i(u + 1)(v - u) \end{aligned}$$

and

$$\begin{aligned} G_i : Q \times Q &\longrightarrow H_2 \\ x, y &\longrightarrow G_i(x, y) = i(x - 10)(y - x), \end{aligned}$$

and we define φ_i and ψ_i as follows:

$$\begin{aligned} \varphi_i &: C \times C \longrightarrow H_1 \\ u, v &\longrightarrow \varphi_i(u, v) = iu(v - u) \end{aligned}$$

and

$$\begin{aligned} \psi_i &: Q \times Q \longrightarrow H_2 \\ x, y &\longrightarrow \psi_i(x, y) = i(x - y). \end{aligned}$$

Let $r_{n,i} = \frac{n}{i(n+1)}$, $s_{n,i} = \frac{n}{i(2n+3)}$, $\alpha_{n,i} = \frac{1}{in}$, $\gamma_n = \frac{1}{11(n+1)}$, and $\gamma = \frac{1}{11}$. It is easy to see that $F_i, G_i, \varphi_i, \psi_i, r_{n,i}, s_{n,i}, \alpha_{n,i}, \gamma_n$, and γ are satisfying all conditions of the proposed method.

Assume that $\mathbf{N} = \mathbf{1}$; then all mappings and sequences become

$F_1, \varphi_1 : C \times C \longrightarrow \mathbb{R}$ such that $F_1(u, v) = (u + 1)(v - u)$, and $\varphi_1(u, v) = u(v - u)$, also $G_1, \psi_1 : Q \times Q \longrightarrow \mathbb{R}$ such that $G_1(x, y) = (x - 10)(y - x)$, and $\psi_1(x, y) = x - y$, and $r_{n,1} = \frac{n}{n+1}$, $s_{n,1} = \frac{n}{2n+3}$, $\alpha_{n,1} = \frac{1}{n}$.

In order to facilitate our calculus, we use r_1 and s_1 instead of $r_{n,1}$ and $s_{n,1}$.

First, we estimate $z \in Q$ such that $z = T_{s_1}^{G_1, \psi_1} Ax$, for every $x \in C$.

Indeed $G_1(z, y) + \psi_1(z, y) + \frac{1}{s_1} \langle y - z, z - Ax \rangle \geq 0 \quad (\star)$.

Clearly,

$$\begin{aligned} (\star) &\iff (z - 10)(y - z) - (y - z) + \frac{1}{s_1} \langle y - z, z - Ax \rangle \geq 0, \\ &\iff s_1(z - 11)(y - z) + (y - z)(z - 3x) \geq 0, \\ &\iff (y - z)(s_1(z - 11) + (z - 3x)) \geq 0, \\ &\iff (y - z)(z(s_1 + 1) - (11s_1 + 3x)) \geq 0, \end{aligned}$$

and from Lemma 2, we have $T_{s_1}^{(G_1, \psi_1)}$ that is single valued, then $z = \frac{11s_1 + 3x}{s_1 + 1}$, which implies that

$$T_{s_1}^{(G_1, \psi_1)} Ax = \frac{11s_1 + 3x}{s_1 + 1}.$$

Now, we determine $w \in C$, such that $w = (I - \gamma A^*(I - T_{s_1}^{(G_1, \psi_1)})A)x$, then

$$w = x - \gamma(9x - 3\frac{11s_1 + 3x}{s_1 + 1}).$$

Now, to compute $u = T_{r_1}^{(F_1, \varphi_1)} w$, we will find $u \in C$, which satisfies

$$F_1(u, v) + \varphi_1(u, v) + \frac{1}{r_1} \langle v - u, u - w \rangle \geq 0, \quad (**)$$

which is equivalent to the following assertions:

$$\begin{aligned} (**) &\iff r_1((u + 1)(v - u) + u(v - u)) + (v - u)(u - w) \geq 0, \\ &\iff (v - u)(r_1(2u + 1) + (u - w)) \geq 0, \\ &\iff (v - u)(u(2r_1 + 1) - (w - r_1)) \geq 0, \end{aligned}$$

and utilizing Lemma 2, we get $u = \frac{w - r_1}{2r_1 + 1}$, which implies that for each $n \in \mathbb{N}$

$$u_{n,1} = \frac{1 - 9\gamma}{2r_{n,1} + 1} x_n + \frac{3\gamma(11s_{n,1} + 3x_n)}{(s_{n,1} + 1)(2r_{n,1} + 1)} - \frac{r_{n,1}}{2r_{n,1} + 1}.$$

And

$$y_{n,1} = \frac{\alpha_{n,1}}{10} x_n + (1 - \alpha_{n,1})u_{n,i}.$$

Then

$$x_{n+1} = \gamma_n P_{C_{n+1}} x_1 + (1 - \gamma_n)y_{n,1}.$$

Since for $x_1 \in C = C_1$, we get $0 \leq y_{1,1} \leq x_1 \leq 20$, then $C_2 = \{p \in C : |y_{1,1} - p| \leq |x_1 - p|\} = \left[0, \frac{y_{1,1} + x_1}{2}\right]$, and it can be clearly seen that $\frac{y_{1,1} + x_1}{2} \leq x_1$, which implies $x_2 = P_{C_2} x_1 = \frac{y_{n,1} + x_1}{2}$. Therefore easily get $C_{n+1} = \left[0, \frac{y_{n,1} + x_n}{2}\right]$ and, consequently, $P_{C_{n+1}} x_1 = \frac{y_{n,1} + x_n}{2}$. Thus x_{n+1} can be rewritten as follows:

$$x_{n+1} = \gamma_n \frac{y_{n,1} + x_n}{2} + (1 - \gamma_n)y_{n,1}.$$

Hence, the new form of Algorithm 3.1 is

$$\begin{cases} u_{n,1} = \frac{1 - 9\gamma}{2r_{n,i} + 1} x_n + \frac{3\gamma(11s_{n,i} + 3x_n)}{(s_{n,i} + 1)(2r_{n,i} + 1)} - \frac{r_{n,i}}{2r_{n,i} + 1}, \\ y_{n,1} = \frac{\alpha_{n,i}}{10} x_n + (1 - \alpha_{n,i})u_{n,i}, \\ x_{n+1} = \gamma_n \frac{y_{n,i} + x_n}{2} + (1 - \gamma_n)y_{n,i}, \quad n \geq 1. \end{cases} \quad (24)$$

Table 1 and Figure 1 clearly show the behavior of the sequence x_n generated by the Algorithm 3.1. which converges to the same solution. i.e. $0 \in \Gamma$:

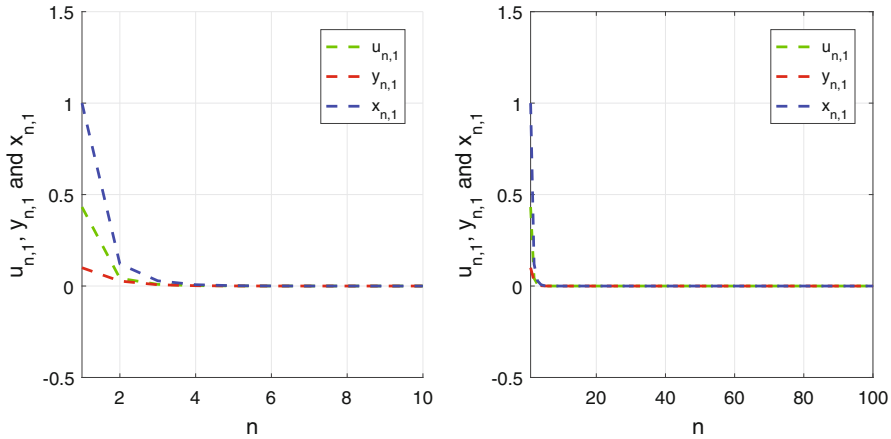


Fig. 1 Convergence of $u_{n,1}, y_{n,1}$, and x_n with initial value $x_1 = 1$

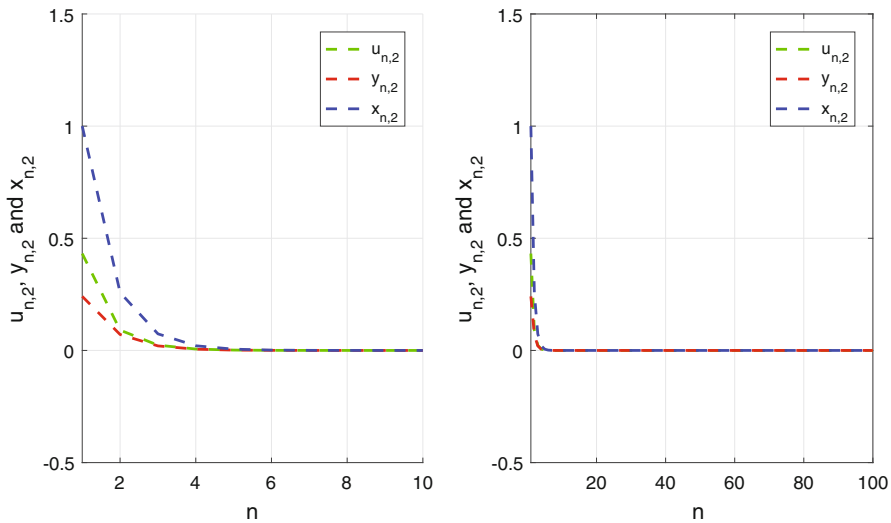


Fig. 2 Convergence of $u_{n,2}, y_{n,2}$, and x_n with initial value $x_1 = 1$

Now we assume that $\mathbf{N} = 2$, and thus mappings and sequences above will be defined as: $F_2 : C \times C \rightarrow \mathbb{R}$ such that $F_2(u, v) = 2(u+1)(v-u)$, $G_2 : Q \times Q \rightarrow \mathbb{R}$ such that $G_2(x, y) = 2(x-10)(y-x)$, and then φ_2 and ψ_2 are denoted by $\varphi_2(u, v) = 2u(v-u)$ and $\psi_2(x, y) = 2(x-y)$. Similarly, $r_{n,2} = \frac{n}{2(n+1)}$, $s_{n,2} = \frac{n}{4n+6}$, and $\alpha_{n,2} = \frac{1}{2n}$, whereas S_2 , take the new form:

$$S_2 : C \rightarrow C$$

Table 1 The values of u_n , y_n , and x_n with initial value $x_1 = 1$

n	Algorithm 3.1		
	u_n	y_n	x_n
1	0.43182	0.10000	1.00000
2	0.04224	0.02714	0.12045
3	0.00908	0.00701	0.02855
4	0.00218	0.00182	0.00725
5	0.00054	0.00047	0.00187
6	0.00013	0.00012	0.00048
7	0.00003	0.00003	0.00012
8	0.00000	0.00000	0.00003
9	0.00000	0.00000	0.00000
10	0.00000	0.00000	0.00000

$$x \rightarrow \frac{x}{20}.$$

Evidently, S_2 is nonexpansive. Since $S_2(0) = 0$, then $F(S_2) = 0$. Thus $\bigcap_{i=1}^{N=2} F(S_i) = 0$.

By the same process, we firstly estimate $z \in Q$, such that $z = T_{S_2}^{G_2, \psi_2} Ax$, for all $x \in C$, and then we get easily

$$z = \frac{22s_2 + 3x}{2s_2 + 1}.$$

Next, we determine $w \in C$, such that $w = (I - \gamma A^*(I - T_{S_2}^{G_2, \psi_2})A)x$, then

$$w = x - \gamma(9x - 3\frac{22s_2 + 3x}{2s_2 + 1}).$$

At last we estimate $u = T_{r_2}^{F_2, \varphi_2} w$, and then by similar calculus we get $u = \frac{w - 2r_2}{4r_2 + 1}$, which finally states

$$\begin{cases} u_{n,2} = \frac{1 - 9\gamma}{4r_{n,2} + 1}x_n + \frac{3\gamma(22s_{n,2} + 3x_n)}{(2s_{n,2} + 1)(4r_{n,2} + 1)} - \frac{2r_{n,2}}{4r_{n,2} + 1}, \\ y_{n,2} = \frac{\alpha_{n,2}}{20}x_n + (1 - \alpha_{n,2})u_{n,2}, \\ x_{n+1} = \gamma_n \frac{y_{n,2} + x_n}{2} + (1 - \gamma_n)y_{n,2}, \quad n \geq 1. \end{cases} \tag{25}$$

Table 2 The values of u_n , y_n , and x_n with initial value $x_1 = 1$

n	Algorithm 3.1		
	u_n	y_n	x_n
1	0.43182	0.24091	1.00000
2	0.09052	0.07112	0.25816
3	0.02353	0.02022	0.07395
4	0.00626	0.00561	0.02084
5	0.00167	0.00153	0.00575
6	0.00044	0.00041	0.00156
7	0.00011	0.00011	0.00042
8	0.00003	0.00003	0.00011
9	0.00000	0.00000	0.00003
10	0.00000	0.00000	0.00000

Table 3 The values x_n with initial value $x_1 = 1$ with three different methods

n	Algorithm 3.1	Algorithm [27]	Algorithm [15]
	x_n	x_n	x_n
1	1.00000	1.00000	1.00000
2	0.12045	0.50000	1.00000
3	0.02855	0.25000	0.75000
4	0.00725	0.12500	0.50000
5	0.00187	0.06250	0.31250
6	0.00048	0.03125	0.18750
7	0.00012	0.01562	0.10938
8	0.00003	0.00781	0.06250
9	0.00000	0.00390	0.03516
10	0.00000	0.00195	0.01953

Similarly Table 2 and Figure 2 show the behavior of the sequence x_n generated by the Algorithm 3.1. Therefore for each $i \in \{1, \dots, N\}$, $x_n \rightarrow q = 0 \in \cap_{i=1}^{N=2} F(S_i) \cap \Omega$.

In the following, we compare the proposed method with those in [27] and [15].

Remark 2 Table 3 and Figure 3 show that the sequences $\{x_n\}$ converge to 0, where $\{0\} = \cap_{i=1}^{N=2} F(S_i) \cap \Omega$.

Also Table 3 and Figure 3 show that the convergence of Algorithm 3.1 is faster than those in [27] and [15].

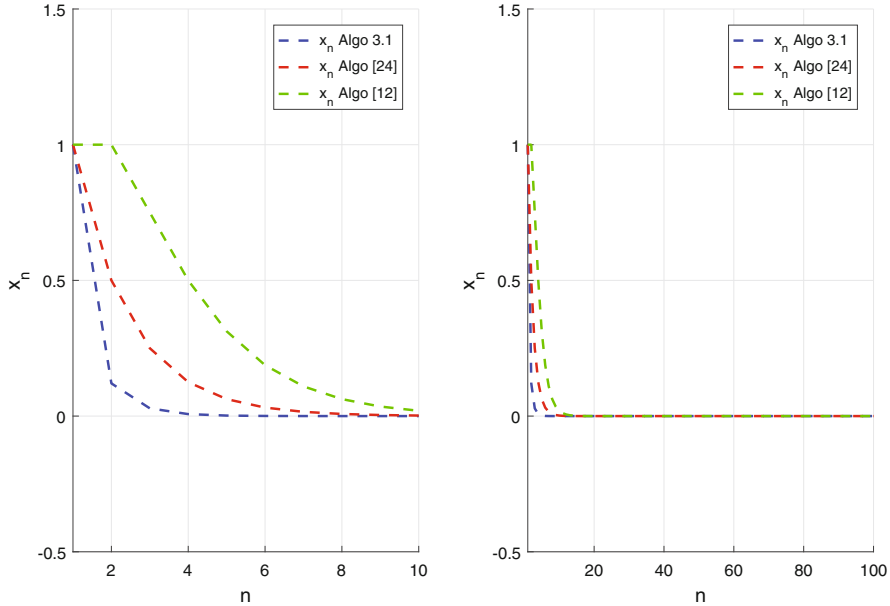


Fig. 3 Convergence of x_n with initial value $x_1 = 1$, with different number of iterations n

5 Conclusions

In this paper, we suggest and analyze an iterative method for finding the approximate element of the common set of solutions of the split generalized equilibrium problem and the set of common fixed points of a finite family of nonexpansive mappings in real Hilbert space, which can be viewed as a refinement and improvement of some existing methods for solving split generalized equilibrium problem. Some existing methods (e.g., [15, 26, 27]) can be viewed as special cases of Algorithm 3.1. Therefore, the new algorithm is expected to be widely applicable.

References

1. A. Bnouhachem, A modified projection method for a common solution of a system of variational inequalities, a split equilibrium problem and a hierarchical fixed point problem. *Fixed Point Theory Appl.* **1**(22), 1–25 (2014)
2. A. Bnouhachem, Strong convergence algorithm for approximating the common solutions of a variational inequality, a mixed equilibrium problem and a hierarchical fixed-point problem. *J. Inequal. Appl.* **2014**(154), 1–24 (2014)
3. A. Bnouhachem, An iterative method for a common solution of generalized mixed equilibrium problems, variational inequalities, and hierarchical fixed point problems. *Fixed Point Theory Appl.* **2014**(155), 1–25 (2014)

4. A. Bnouhachem, A hybrid iterative method for a combination of equilibrium problem, a combination of variational inequality problem and a hierarchical fixed point problem. *Fixed Point Theory Appl.* **2014**(163), 1–29 (2014)
5. A. Bnouhachem, An iterative algorithm for system of generalized equilibrium problems and fixed point problem. *Fixed Point Theory Appl.* **2014**(235), 1–22 (2014)
6. A. Bnouhachem, S. Al-Homidan, Q.H. Ansari, An iterative method for common solutions of equilibrium problems and hierarchical fixed point problems. *Fixed Point Theory Appl.* **2014**(194), 1–21 (2014)
7. A. Bnouhachem, Q.H. Ansari, J.C. Yao, An iterative algorithm for hierarchical fixed point problems for a finite family of nonexpansive mappings. *Fixed Point Theory Appl.* **2015**(111), 1–13 (2015)
8. A. Moudafi, Weak convergence theorems for nonexpansive mappings and equilibrium problems. *J. Nonlinear Convex Anal.* **9**, 37–43 (2008)
9. E. Blum, W. Oettli, From optimization and variational inequalities to equilibrium problems. *Math. Stud.* **63**, 123–145 (1994)
10. F. Facchinei, C. Kanzow, Generalized Nash equilibrium problems. *4OR Q. J. Belg. French Ital. Oper. Res. Soc.* **5**(3), 173–210 (2007)
11. Y. Censor, T.A. Elfving, A multiprojection algorithm using Bregman projections in a product space. *Numer. Algorithms* **8**(2), 221–239 (1994)
12. H.K. Xu, Iterative algorithms for nonlinear operators. *J. Lond. Math. Soc.* **66**(1), 240–256 (2002)
13. K.R. Kazmi, S.H. Rizvi, Iterative approximation of a common solution of a split equilibrium problem, a variational inequality problem and a fixed point problem. *J. Egypt. Math. Soc.* **21**(1), 44–51 (2013)
14. J. Deepho, J.M. Moreno, P. Kumam, A viscosity of Cesaro mean approximation method for split generalized equilibrium, variational inequality and fixed point problems. *J. Nonlinear Sci. Appl.* **9**, 1475–1496 (2016)
15. J. Deepho, W. Kumam, P. Kumam, A new hybrid projection algorithm for solving the split generalized equilibrium problems and the system of variational inequality problems. *J. Math. Model. Algor.* **13**(4), 405–423 (2014)
16. K.R. Kazmi, S.H. Rizvi, Iterative approximation of a common solution of a split generalized equilibrium problem and a fixed point problem for nonexpansive semigroup. *Math. Sci.* **7**(1), 1–10 (2013)
17. L.C. Ceng, J.C. Yao, A hybrid iterative scheme for mixed equilibrium problems and fixed point problems. *J. Comput. Appl. Math.* **214**(1), 186–201 (2008)
18. L. Grubisic, J. Tambaca, Direct solution method for the equilibrium problem for elastic stents. *Numer. Linear Algebra Appl.* **26**(3), 22–31 (2019)
19. L.C. Zeng, S.Y. Wu, J.C. Yao, Generalized KKM theorem with applications to generalized minimax inequalities and generalized equilibrium problems. *Taiwan. J. Math.* **10**(6), 1497–1514 (2006)
20. L.O. Jolaoso, K.O. Oyewole, C.C. Okeke, O.T. Mewomo, A unified algorithm for solving split generalized mixed equilibrium problem, and for finding fixed point of nonspreading mapping in Hilbert spaces. *Demonstration Math.* **51**(1), 211–232 (2018)
21. M.A. Khan, Y. Arfat, A.R. Butt, A Shrinking projection approach for equilibrium problem and fixed point problem in Hilbert spaces. *U.P.B. Sci. Bull.* **80**(1), 33–46 (2018)
22. M. Abdulaal, L.J. LeBlanc, Methods for combining modal split and equilibrium assignment models. *Transpn. Sci.* **13**(4), 292–314 (1979)
23. M. Bianchi, S. Schaible, Generalized monotone bifunctions and equilibrium problems. *J. Optim.Theory Appl.* **90**(1), 31–43 (1996)
24. P.I. Combettes, S.A. Hirstoaga, Equilibrium programming in Hilbert spaces. *J. Nonlinear Convex Anal.* **6**, 117–136 (2005)
25. S. Reich, S. Sabach, Three strong convergence theorems regarding iterative methods for solving equilibrium problems in reflexive Banach spaces. *Contemp. Math.* **56**, 225–240 (2012)

26. S. Suantai, P. Cholamjiak, Y.J. Cho, W. Cholamjiak, On solving split equilibrium problems and fixed point problems of nonspreading multi-valued mappings in Hilbert spaces. *Fixed Point Theory Appl.* **2016**(1), 1–35 (2016)
27. W. Phuengrattana, K. Lerkchaiyaphum, On solving the split generalized equilibrium problem and the fixed point problem for a countable family of nonexpansive multivalued mappings. *Fixed Point Theory Appl.* **2018**(1), 1–17 (2018)
28. X. Qin, S. Chang, Y. J. Cho, Iterative methods for generalized equilibrium problems and fixed point problems with applications. *Nonlinear Anal. Real World Appl.* **11**(4), 2963–2972 (2010)
29. X. Qin, Y.J. Cho, S.M. Kang, Viscosity approximation methods for generalized equilibrium problems and fixed point problems with applications. *Nonlinear Anal. Theory, Methods Appl.* **72**(1), 99–112 (2010)
30. Y. Censor, T. Bortfeld, B. Martin, A. Trofimov, A unified approach for inverse problem in intensity-modulated radiation therapy. *Phys. Med. Biol.* **51**(10), 2353–2365 (2006)
31. Y.J. Cho, X. Qin, J.I. Kang, Convergence theorems based on hybrid methods for generalized equilibrium problems and fixed point problems. *Nonlinear Anal. Theory Methods Appl.* **71**(9), 4203–4214 (2009)
32. K. Geobel, W.A. Kirk, *Topics in Metric Fixed Point Theory*. Cambridge Studies in Advanced Mathematics, vol. 28 (Cambridge University Press, Cambridge, 1990)
33. Y. Yao, Y.-C. Liou, S.M. Kang, Approach to common elements of variational inequality problems and fixed point problems via a relaxed extragradient method. *Comput. Math. Appl.* **59**(11), 3472–3480 (2010)

Periodic Solutions Around the Out-of-Plane Equilibrium Points in the Restricted Three-Body Problem with Radiation and Angular Velocity Variation



Vassilis S. Kalantonis, Aguda Ekele Vincent, Jessica Mrumun Gyegwe, and Efstathios A. Perdios

Abstract In the present work, we study the motion of an infinitesimal body near the out-of-plane equilibrium points of the restricted three-body problem in which the angular velocity of the two primary bodies is considered in the case where both of them are sources of radiation. Firstly, these equilibria are determined numerically, and then the influence of the system parameters on their positions is examined. Due to the symmetry of the problem, these points appear in pairs and, depending on the parameter values, their number may be zero, two, or four. The linear stability of the out-of-plane equilibrium points is also studied, and it is found that there are cases where they can be stable. In addition, periodic motion around them is investigated both analytically and numerically. Specifically, the Lindstedt–Poincaré method is used in order to obtain a second order analytical solution, while the families emanating from the out-of-plane equilibrium points are finally computed numerically either in case where the corresponding equilibrium points are stable or unstable. For the numerical computation of a three-dimensional periodic orbit, we apply known unconstrained optimization methods to an objective function that is formed by the respective periodicity conditions that have to be fulfilled.

MSC 70F07, 70F15, 70M20, 70K42

V. S. Kalantonis (✉) · E. A. Perdios

Department of Electrical & Computer Engineering, University of Patras, Patras, Greece
e-mail: kalantonis@upatras.gr; eperdios@upatras.gr

A. E. Vincent

Department of Mathematics, Nigeria Maritime University, Okerenkoko, Delta State, Nigeria
e-mail: vincentekele@yahoo.com

J. M. Gyegwe

Department of Mathematical Sciences, Federal University Lokoja, Lokoja, Kogi State, Nigeria
e-mail: jessica.gyegwe@fulokoja.edu.ng

© Springer Nature Switzerland AG 2021

T. M. Rassias, P. M. Pardalos (eds.), *Nonlinear Analysis and Global Optimization*,
Springer Optimization and Its Applications 167,
https://doi.org/10.1007/978-3-030-61732-5_11

251

1 Introduction

The restricted three-body problem concerns the motion of a massless body that moves under the gravitational attraction of two massive bodies, known as the primary and secondary bodies. The latter two bodies revolve in circular orbits around their common center of mass, and their motion is not affected by the third body of infinitesimal mass. It is well known that in the rotating frame this problem possesses five equilibrium points three of which lie on the line connecting the primaries and are called collinear points, while the other two form in the plane of motion equilateral configuration with the primaries and are called triangular points. For details on the characteristics of the restricted three-body problem, we may refer here the book by Valtonen and Karttunen [29] as well as the review article by Musielak and Quarles [10].

During the past, several variants of this classical problem have been proposed in order to make it more applicable to real systems of Dynamical Astronomy. These modifications include additional forces other than the gravitational one or take into account the shape of stars, planets, or moons (see, e.g., [4, 7, 30, 32]). So, a different version of the restricted three-body problem arises when one or both primaries are sources of radiation (the photogravitational problem) or when the angular velocity of the primaries is considered (the Chermnykh's problem). In the photogravitational problem, in addition to the five coplanar points of the classical problem, there exist equilibrium points that lie out of the orbital plane of the two primary bodies. The existence of such kind of equilibrium points was pointed out by Radzievskii [18] and several authors based their works on the Radzievskii's model in order to understand various issues related to the dynamics of a particle around radiating primaries (e.g., [2, 3, 17, 20, 21, 23–26]). The Chermnykh's problem in which the angular velocity variation of the primaries is considered has also been discussed in the context of planar or three-dimensional case with regard to its equilibrium points (see [1, 6, 8, 11, 12, 14, 15], and references therein).

Recently, Perdios et al. [16] examined equilibrium points and related periodic motions in the restricted three-body problem with angular velocity and radiation effects. That investigation was performed in the two-dimensional scenario. In this paper, we wish to study the three-dimensional case of this special modification of the restricted problem by considering the out-of-plane equilibrium points and especially the motion around them. Our aim is not to obtain any particular application of this model but to gain more insight about its dynamics. To do so, we first compute numerically the number and positions of this kind of equilibrium points and then determine their stability. We find that in contrast to the classical problem but in agreement with the photogravitational one the current problem can admit two or four such equilibria and stability may occur. As we have already mentioned, our work focuses on the periodic motion around these points; thus, an analytical as well as a numerical study have also been performed. In particular, for the analytical part of this paper, we have used the Lindstedt–Poincaré technique so as to obtain a second order semi-analytical periodic solution that has been used as a seed for the numerical

part of our work. Specifically, using initial conditions as they are determined by the analytical solution, the families of three-dimensional periodic orbits emanating from the out-of-plane equilibrium points have been determined together with their stability properties. For the numerical computation of the members of these families, we have followed the work by Kalantonis et al. [9] and have transformed the root-finding differential correction procedure (usually used for their determination) to an unconstrained optimization problem applied on an objective function that is based on the exact periodicity conditions. To accomplish it, we have adopted the corresponding unconstrained optimization algorithms developed by Broyden–Fletcher–Goldfarb–Shanno (BFGS) and Davidon–Fletcher–Powell (DFP).

Our paper is organized in five sections. In Section 2, we recall the equations of motion of the considered dynamical system. In Section 3, the positions of the out-of-plane equilibrium points are located. Specifically, we discuss the influence of the four system parameters (mass parameter μ , angular velocity ω , radiation factors q_1 and q_2 of the primaries) on the equilibrium points in a parametric way as well as we study their linear stability. The allowed regions of motion as determined by the zero velocity curves are also considered. In Section 4, we establish a second order semi-analytical periodic solution around the out-of-plane equilibrium points. The numerical determination of the families of three-dimensional periodic solutions emanating from these points is also presented in the same section. Our paper ends in Section 5 in which the obtained results and conclusions of the paper are discussed.

2 Equations of Motion

The equations of motion of the infinitesimal mass m_3 in the three-dimensional photogravitational Chermnykh’s restricted three-body problem with the origin resting at the center of mass, in a rotating system of coordinates, can be described in the dimensionless variables as (see [16], and references therein)

$$\begin{aligned} \ddot{x} - 2\omega\dot{y} &= \frac{\partial\Omega}{\partial x} = \omega^2x - \frac{q_1(1-\mu)(x+\mu)}{r_1^3} - \frac{q_2\mu(x+\mu-1)}{r_2^3}, \\ \ddot{y} + 2\omega\dot{x} &= \frac{\partial\Omega}{\partial y} = \omega^2y - \frac{q_1(1-\mu)y}{r_1^3} - \frac{q_2\mu y}{r_2^3}, \\ \ddot{z} &= \frac{\partial\Omega}{\partial z} = -\frac{q_1(1-\mu)z}{r_1^3} - \frac{q_2\mu z}{r_2^3}, \end{aligned} \tag{1}$$

where dots denote time derivatives and Ω is the potential function in synodic coordinates:

$$\Omega = \frac{1}{2}\omega^2(x^2 + y^2) + \frac{q_1(1-\mu)}{r_1} + \frac{q_2\mu}{r_2}, \tag{2}$$

while

$$r_1^2 = (x + \mu)^2 + y^2 + z^2, \quad r_2^2 = (x + \mu - 1)^2 + y^2 + z^2, \quad (3)$$

where r_1 and r_2 are the distances of the third body from the primaries. The parameter $\omega \in (0, \infty)$ is the angular velocity of the problem, μ is the mass ratio of the smaller primary to the total mass of the primaries with $0 < \mu \leq 1/2$, while the larger primary is located at the position $(-\mu, 0, 0)$ and the second primary at $(1 - \mu, 0, 0)$, correspondingly, and the unit of distance is the distance between the primaries. The radiation pressure parameters of the primaries q_1 and q_2 according to Radzievskii's theory are expressed by means of the relations $q_i = 1 - b_i$, $i = 1, 2$, where b_1 and b_2 are the ratios of the radiation force F_r to the gravitational force F_g , which results from the gravitation due to the two primary bodies m_1 and m_2 , respectively. It is interesting to note that for $q_1 = q_2 = \omega = 1$ we obtain the classical circular restricted three-body problem. It is clear that if $q_{1,2} = 1$ radiation pressure has no effect, if $0 < q_{1,2} < 1$ gravitational force exceeds radiation, if $q_{1,2} = 0$ radiation force balances the gravitational one, and if $q_{1,2} < 0$ then radiation pressure overrides the gravitational attraction. The energy (Jacobi-like) integral of this problem is given by the expression:

$$\dot{x}^2 + \dot{y}^2 + \dot{z}^2 = 2\Omega - C, \quad (4)$$

where C is the Jacobi constant, while \dot{x} , \dot{y} , and \dot{z} correspond to the components of the velocity.

3 Out-of-Plane Equilibrium Points

The equilibrium points out of the plane Oxy can be found by setting all velocity and acceleration terms equal to zero and solving the right sides of system (1). Obviously, the second equation of this system is satisfied for $y = 0$, and we solve the remaining two equations for $y = 0$ and $z \neq 0$, namely:

$$\omega^2 x_0 - \frac{q_1(1 - \mu)(x_0 + \mu)}{r_{10}^3} - \frac{q_2\mu(x_0 + \mu - 1)}{r_{20}^3} = 0, \quad (5)$$

$$\left[\frac{q_1(1 - \mu)}{r_{10}^3} + \frac{q_2\mu}{r_{20}^3} \right] z_0 = 0,$$

where the two distances can be deduced to

$$r_{10}^2 = (x_0 + \mu)^2 + z_0^2, \quad r_{20}^2 = (x_0 + \mu - 1)^2 + z_0^2, \quad (6)$$

and the subscript “0” is used to denote that these quantities have been evaluated at the equilibrium values. The latter equations for the distances give that

$$\frac{r_{20}}{r_{10}} = \left(\frac{-q_2}{q_1} \frac{\mu}{1 - \mu} \right)^{\frac{1}{3}} \equiv k, \tag{7}$$

which means that if $k = \text{constant}$ the locus of these points is an Apollonius circle. We found that the specific problem may admit up to four equilibrium points in the (x, z) plane, in particular, L_1^z, L_3^z for $z > 0$ and their symmetric points with respect to the Oxy plane, L_2^z, L_4^z , respectively (i.e., for $z < 0$). Their positions are hard to be obtained with analytical expressions; however, they can be approximated by using any numerical method for solving nonlinear algebraic systems. Note that the existence, number, and location of these equilibria depend on the parameters ω, μ, q_1 , and q_2 . From Equation (7), it can be easily seen that for the existence of any real solution the following condition is necessary to hold:

$$q_1 q_2 < 0, \tag{8}$$

which means that the radiation pressure force of just one of the primaries exceeds its gravitational attraction. We also mention here that in previous studies about the photogravitational version of the restricted problem a necessary but not sufficient condition for the existence of the out-of-plane equilibrium points was also to consider negative values for the radiation factors (see, e.g., [19, 25, 27]).

Next, we shall discuss the positions of the out-of-plane equilibrium points of the test body under the above condition for angular velocity variation $\omega \in (0, 5]$, whereas the radiation factors q_1 and q_2 vary in the interval $q_i \in [-1, 1], i = 1, 2$, while they always satisfy the condition $q_1 q_2 < 0$. Our target is not to provide a systematic search of their location and existence but to detect them for several combinations of the parameters of the problem so as to generate the corresponding three-dimensional periodic orbits that originate out of them.

To investigate the influence of the mass parameter and angular velocity on the positions of the equilibria under consideration, the radiation factor of the first primary is arbitrary set to be $q_1 = -0.01$, while that of the second primary is set to be $q_2 = 0.5$, thus satisfying the requested condition (8). The coordinates of the numerically determined out-of-plane equilibrium points are shown in Table 1 for various values of the angular velocity ω and the mass parameter μ . For mass parameter $\mu = 0.1$, the locations of the equilibrium points with respect to different values of ω are presented in the second column of Table 1. We observe that with the increase of the angular velocity parameter ω , the z -coordinates of the equilibria decrease while at the same time the x -coordinates approach the origin. The third and fourth columns of Table 1 show the variational trend of the positions of the equilibria with the variation of ω for the values of the mass parameter $\mu = 0.2$ and $\mu = 0.25$, respectively. Evidently, the variational trend of the corresponding positions is similar to the scenario with $\mu = 0.1$ as described previously. It is

Table 1 The positions $(x_0, \pm z_0)$ of the out-of-plane equilibrium points as a function of ω for $q_1 = -0.01$ and $q_2 = 0.5$

ω	$\mu = 0.1$	$\mu = 0.2$	$\mu = 0.25$
0.5	$(-0.1093610, \pm 0.690410)$	$(-0.2528320, \pm 0.499293)$	$(-0.320035, \pm 0.448833)$
1.0	$(-0.0346804, \pm 0.634498)$	$(-0.1019530, \pm 0.416740)$	$(-0.140896, \pm 0.360004)$
1.5	$(-0.0164374, \pm 0.618705)$	$(-0.0546692, \pm 0.374972)$	$(-0.080534, \pm 0.301557)$
2.0	$(-0.0094826, \pm 0.612435)$	$(-0.0336957, \pm 0.352851)$	$(-0.051897, \pm 0.264657)$
2.5	$(-0.0061438, \pm 0.609373)$	$(-0.0226790, \pm 0.340139)$	$(-0.036018, \pm 0.240316)$
3.0	$(-0.0042958, \pm 0.607664)$	$(-0.0162333, \pm 0.332307)$	$(-0.026338, \pm 0.223632)$
3.5	$(-0.0031694, \pm 0.606617)$	$(-0.0121607, \pm 0.327197)$	$(-0.020032, \pm 0.211821)$
4.0	$(-0.0024332, \pm 0.605931)$	$(-0.0094339, \pm 0.323701)$	$(-0.015713, \pm 0.203221)$
4.5	$(-0.0019262, \pm 0.605457)$	$(-0.0075235, \pm 0.321216)$	$(-0.012634, \pm 0.196805)$
5.0	$(-0.0015623, \pm 0.605117)$	$(-0.0061355, \pm 0.319391)$	$(-0.010367, \pm 0.191913)$

Table 2 The positions $(x_0, \pm z_0)$ of the out-of-plane equilibrium points as a function of q_2 for $\mu = 0.3$ and $\omega = 0.75$

q_2	$q_1 = -0.05$	$q_1 = -0.03$	$q_1 = -0.01$
1.0	$(-0.327184, \pm 0.574423)$	$(-0.349825, \pm 0.471707)$	$(-0.377815, \pm 0.310964)$
0.9	$(-0.304154, \pm 0.589214)$	$(-0.327545, \pm 0.484068)$	$(-0.356382, \pm 0.321954)$
0.8	$(-0.279374, \pm 0.605808)$	$(-0.303641, \pm 0.497241)$	$(-0.333459, \pm 0.332499)$
0.7	$(-0.252452, \pm 0.625177)$	$(-0.277761, \pm 0.511759)$	$(-0.308735, \pm 0.342759)$
0.6	$(-0.222829, \pm 0.649047)$	$(-0.249413, \pm 0.528542)$	$(-0.281778, \pm 0.353003)$
0.5	$(-0.189672, \pm 0.680798)$	$(-0.217867, \pm 0.549332)$	$(-0.251962, \pm 0.363741)$
0.4	$(-0.151654, \pm 0.728114)$	$(-0.181978, \pm 0.577854)$	$(-0.218312, \pm 0.376088)$
0.3	$(-0.106477, \pm 0.813339)$	$(-0.139772, \pm 0.623759)$	$(-0.179175, \pm 0.392896)$
0.2	$(-0.050027, \pm 1.045970)$	$(-0.087408, \pm 0.722465)$	$(-0.131332, \pm 0.423550)$

also observed from the same table that the positions of these two equilibria vary in a relatively small range with $\omega \geq 1$ compared to that with $\omega < 1$ as the mass parameter μ varies. This indicates that low angular velocity values have greater impact on the out-of-plane equilibria than the high angular velocity values.

Similarly, for the investigation of the influence of the radiation parameters q_1 and q_2 on the positions of the out-of-plane equilibrium points, we set for the mass parameter and angular velocity the values $\mu = 0.3$ and $\omega = 0.75$, respectively. The coordinates of the corresponding equilibrium points are shown in Table 2 for increasing values of radiation factors q_1 and q_2 . Recall here that radiation pressure increases when $q_{1,2}$ decreases. In particular, the second column of Table 2 shows the variational trend of the position of the equilibrium points when the values of $\mu = 0.3, \omega = 0.75, q_1 = -0.05$ are fixed along with the variation of q_2 . With the decrease of the radiation factor q_2 from 1 to 0.2, the x -coordinates approach the origin, while at the same time the values of the z -coordinates increase. The third and fourth columns of this table show the variational trend of the positions of the equilibria with the variation of q_2 for the same fixed values $\mu = 0.3, \omega = 0.75$ but for $q_1 = -0.03$ and $q_1 = -0.01$, respectively. It is obvious that the variational

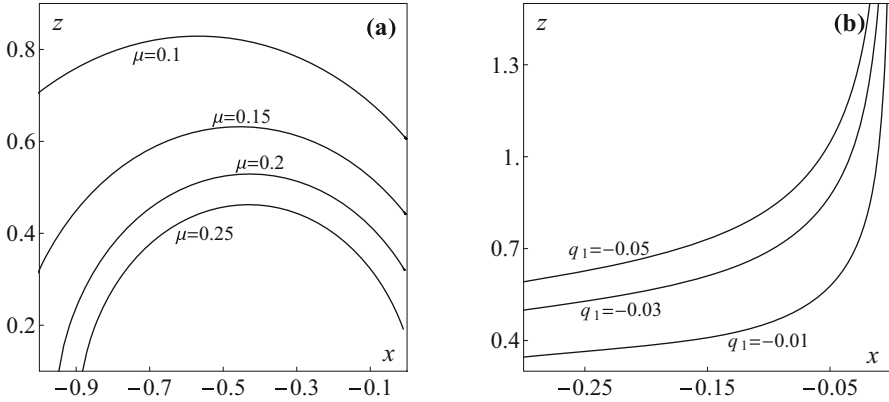


Fig. 1 Positions of the positive out-of-plane equilibrium points for (a) fixed values of the radiation factors $q_1 = -0.01$ and $q_2 = 0.5$ and various values of the mass parameter when the angular velocity ω varies (b) fixed values of the angular velocity and mass parameters $\omega = 0.75$ and $\mu = 0.3$, respectively, and various values of the radiation factor q_1 when the radiation factor q_2 varies

trend of the corresponding equilibria position is similar to the previously discussed scenario with $q_1 = -0.05$. So, by decreasing the values of the radiation factor q_2 , the out-of-plane equilibria move away from the primaries and approach the Oz -axis.

The aforementioned discussion can be summarized in Figure 1, where we plot the positions of the positive out-of-plane equilibrium points L_1^z . In particular, in frame (a) of this figure, we show the respective positions when the values of the radiation factors are kept fixed to $q_1 = -0.01$ and $q_2 = 0.5$, respectively, when the angular velocity varies and the mass parameter takes the values $\mu = 0.1$, $\mu = 0.15$, $\mu = 0.2$, and $\mu = 0.25$. In frame (b), we present the positions of the out-of-plane equilibria in the (x, z) -plane as q_2 varies, for the fixed values of $q_1 = -0.05$, $q_1 = -0.03$, and $q_1 = -0.01$, while the remaining parameters are $\omega = 0.75$ and $\mu = 0.3$. As we see, the variational trend of the equilibria positions in Figure 1a, b is similar to the corresponding scenarios presented in Tables 1 and 2, respectively. In Figure 2, we show a representative case where two positive out-of-plane equilibrium points exist simultaneously, namely the points L_1^z and L_3^z . The corresponding figure is illustrated for fixed values of the mass parameter $\mu = 0.4$ and radiation factor $q_2 = -1.5$ for three specific values of the radiation factor q_1 , while the angular velocity ω varies. In the first frame, we present the position of the out-of-plane equilibrium point L_1^z , while in the second frame the position of the equilibrium point L_3^z .

In the following, we present, as a matter of interest, the contours of the surface (4) on the (x, z) plane for zero velocity, which provide the zero velocity curves. These curves define the areas on this plane where the motion of the test particle is allowed or forbidden. In particular, in Figure 3, we present the zero velocity curves for fixed values of $\mu = 0.1$, $q_1 = -0.01$, and $q_2 = 0.5$ and for the three values of

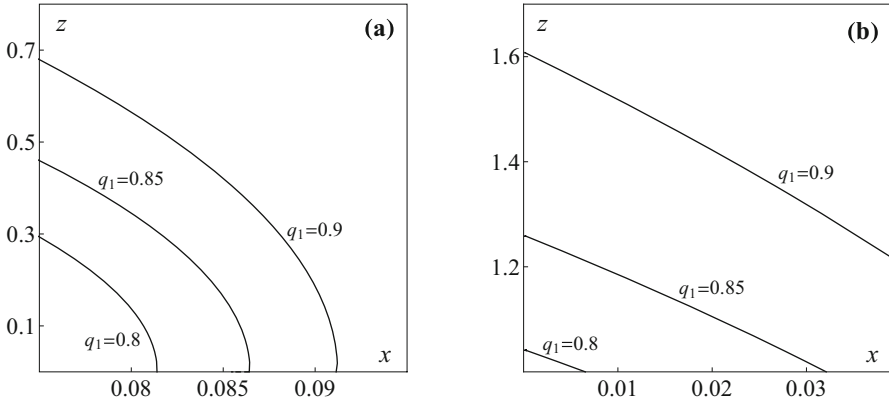


Fig. 2 Positions of the two positive out-of-plane equilibrium points for fixed values of the mass parameter $\mu = 0.4$ and radiation factor $q_2 = -1.5$ and various values of the radiation factor q_1 when the angular velocity ω varies (a) out-of-plane equilibrium point L_1^z , and (b) out-of-plane equilibrium point L_2^z

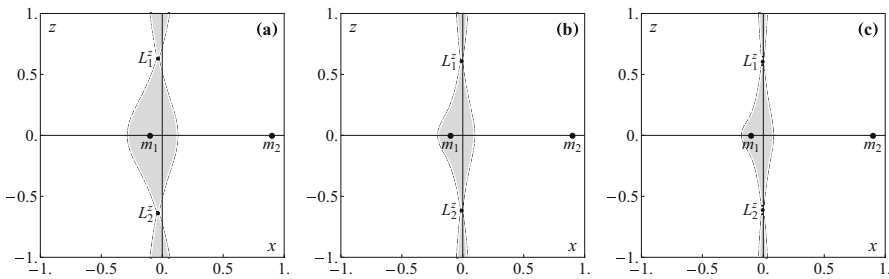


Fig. 3 Zero velocity curves in the (x, z) plane and locations of the out-of-plane equilibria for (a) $\omega = 1$, (b) $\omega = 2$, and (c) $\omega = 2.75$, correspondingly. The locations of the primary bodies are presented too. Note: The values of $\mu = 0.1$, $q_1 = -0.01$, and $q_2 = 0.5$ are fixed for all cases

angular velocity $\omega = 1$, $\omega = 2$, and $\omega = 5$, correspondingly. It can be seen that the zero velocity curves between the out-of-plane equilibrium points form regions not allowed to possible motion, which shrink as the angular velocity increases. So, it results from the equipotential curves of (4) that the value of ω has significant effects on the topological structure of the forbidden regions to motion of the massless body. Figure 4 illustrates the zero velocity curves for fixed values of $\omega = 2$, $q_1 = -0.01$, and $q_2 = 0.5$ and for the three values of mass parameter $\mu = 0.15$, $\mu = 0.2$, and $\mu = 0.25$, correspondingly. It can be observed that with the increase of the mass parameter, the zero velocity curves up to the out-of-plane equilibria go approaching the bigger primary body m_1 . It means that the regions of forbidden motion shrink with its increase. In Figure 5, the zero velocity curves for fixed values of $\mu = 0.3$, $q_1 = -0.05$, and $\omega = 0.75$ and for three values of radiation factor ($q_2 = 1$,

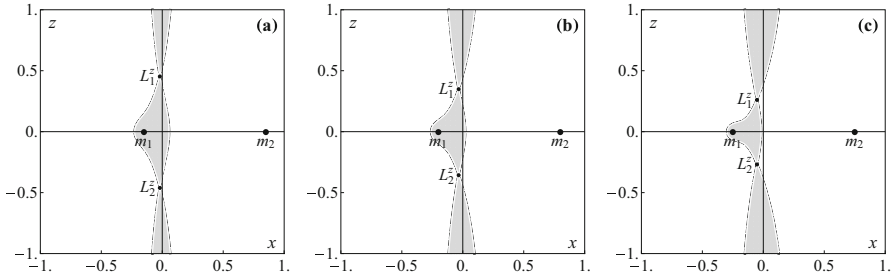


Fig. 4 Zero velocity curves in the (x, z) plane and locations of the out-of-plane equilibria for (a) $\mu = 0.15$, (b) $\mu = 0.2$, and (c) $\mu = 0.25$, correspondingly. The locations of the primary bodies are presented too. Note: The values of $\omega = 2$, $q_1 = -0.01$ and $q_2 = 0.5$ are fixed for all cases

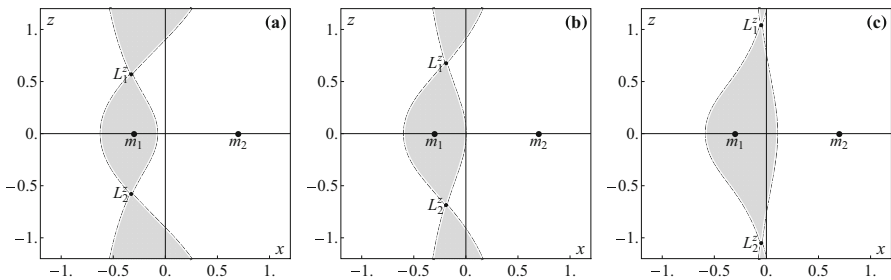


Fig. 5 Zero velocity curves in the (x, z) plane and locations of the out-of-plane equilibria for (a) $q_2 = 1$, (b) $q_2 = 0.5$, and (c) $q_2 = 0.2$, correspondingly. The locations of the primary bodies are presented too. Note: The values of $\mu = 0.3$, $\omega = 0.75$, and $q_1 = -0.05$ are fixed for all cases

$q_2 = 0.5$, and $q_2 = 0.2$) are illustrated. It is clear that the forbidden region of motion expands as the radiation pressure q_2 increases. In Figure 6, we present the zero velocity curves with $\mu = 0.3$, $q_2 = 0.9$, and $\omega = 0.75$ when the radiation parameter q_1 takes the values $q_1 = -0.05$, $q_1 = -0.03$, and $q_1 = -0.01$, correspondingly. And in this case, around the dominant primary body m_1 and up to the out-of-plane equilibrium points, the zero velocity curves form small ovals of regions not allowed to motion, which shrink as the radiation pressure q_1 increases.

To determine the linear stability of an out-of-plane equilibrium point L_i^z , $i = 1, 2, 3, 4$, we transfer the origin to its position $(x_0, 0, z_0)$ by introducing the new variables (ξ, η, ζ) and linearize the equations of motion obtaining:

$$\begin{aligned} \ddot{\xi} - 2\omega\dot{\eta} &= \Omega_{\xi\xi}^{(0)}\xi + \Omega_{\xi\eta}^{(0)}\eta + \Omega_{\xi\zeta}^{(0)}\zeta, \\ \ddot{\eta} + 2\omega\dot{\xi} &= \Omega_{\eta\xi}^{(0)}\xi + \Omega_{\eta\eta}^{(0)}\eta + \Omega_{\eta\zeta}^{(0)}\zeta, \\ \ddot{\zeta} &= \Omega_{\zeta\xi}^{(0)}\xi + \Omega_{\zeta\eta}^{(0)}\eta + \Omega_{\zeta\zeta}^{(0)}\zeta, \end{aligned} \tag{9}$$

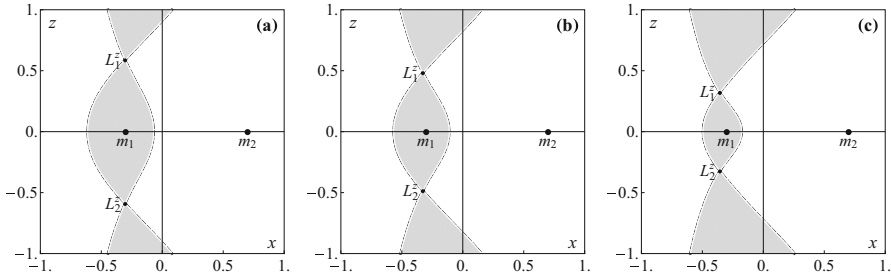


Fig. 6 Zero velocity curves in the (x, z) plane and locations of the out-of-plane equilibria for (a) $q_1 = -0.05$, (b) $q_1 = -0.03$, and (c) $q_1 = -0.01$, correspondingly. The locations of the primary bodies are presented too. Note: The values of $\mu = 0.3$, $\omega = 0.75$, and $q_2 = 0.9$ are fixed for all cases

where the superscript “(0)” indicates that the partial derivatives have been evaluated at the equilibrium point. Explicitly, the partial derivatives of system (9) are $\Omega_{\xi\eta}^{(0)} = \Omega_{\eta\xi}^{(0)} = \Omega_{\zeta\eta}^{(0)} = 0$ and:

$$\begin{aligned} \Omega_{\xi\xi}^{(0)} &= \omega^2 - \frac{(1-\mu)q_1}{r_{10}^3} - \frac{\mu q_2}{r_{20}^3} + \frac{3(1-\mu)(x_0 + \mu)^2 q_1}{r_{10}^5} + \frac{3\mu(x_0 + \mu - 1)^2 q_2}{r_{20}^5}, \\ \Omega_{\xi\zeta}^{(0)} &= \Omega_{\zeta\xi}^{(0)} = 3z_0 \left[\frac{(1-\mu)(x_0 + \mu)q_1}{r_{10}^5} + \frac{\mu(x_0 + \mu - 1)q_2}{r_{20}^5} \right], \\ \Omega_{\eta\eta}^{(0)} &= \omega^2 - \frac{(1-\mu)q_1}{r_{10}^3} - \frac{\mu q_2}{r_{20}^3}, \\ \Omega_{\zeta\zeta}^{(0)} &= -\frac{(1-\mu)q_1}{r_{10}^3} - \frac{\mu q_2}{r_{20}^3} + \frac{3(1-\mu)q_1 z_0^2}{r_{10}^5} + \frac{3\mu q_2 z_0^2}{r_{20}^5}, \end{aligned} \tag{10}$$

with

$$r_{10}^2 = (x_0 + \mu)^2 + z_0^2, \quad r_{20}^2 = (x_0 + \mu - 1)^2 + z_0^2. \tag{11}$$

The characteristic equation corresponding to system (9) is given by

$$\lambda^6 + a\lambda^4 + b\lambda^2 + c = 0, \tag{12}$$

where

Table 3 The eigenvalues $\lambda_{1,2}, \lambda_{3,4,5,6}$ of Equation (12) for the out-of-plane equilibrium points $L_{1,2}^z$

q_2	q_1	ω	$\mu = 0.1$	$\mu = 0.3$
0.9	-0.05	0.75	$\pm 0.17272i, \pm(0.10994 \pm 0.74811i)$	$\pm 0.71282i, \pm(0.42692 \pm 0.70050i)$
0.7	-0.04	1.50	$\pm 0.14891i, \pm(0.10278 \pm 1.49983i)$	$\pm 0.90444i, \pm(0.41697 \pm 1.41945i)$
0.5	-0.03	2.00	$\pm 0.11696i, \pm(0.08193 \pm 1.99997i)$	$\pm 0.71473i, \pm(0.37629 \pm 1.97134i)$
0.3	-0.02	2.75	$\pm 0.07481i, \pm(0.05277 \pm 2.75000i)$	$\pm 0.50428i, \pm(0.31129 \pm 2.74449i)$
0.1	-0.01	3.00	$\pm 0.00709i, \pm(0.00501 \pm 3.00000i)$	$\pm 0.26538i, \pm(0.18066 \pm 2.99957i)$

Table 4 The eigenvalues $\lambda_{1,2}, \lambda_{3,4}, \lambda_{5,6}$, of Equation (12) and the corresponding positions of the out-of-plane equilibrium points $L_{1,2}^z$ and $L_{3,4}^z$ for the value of the mass parameter $\mu = 0.4$

q_2	q_1	ω	$L_{1,2}^z$	$L_{3,4}^z$
-1.5	0.9	4.0	$\pm 0.95627i, \pm 3.09459i, \pm 4.63779i$	$\pm 0.14898, \pm 3.87865i, \pm 4.12047i$
			$(x, z) = (0.07943819, \pm 0.57904060)$	$(x, z) = (0.00845626, \pm 1.53273515)$
-1.4	0.8	4.5	$\pm 0.98123i, \pm 3.53042i, \pm 5.20321i$	$\pm 0.22101, \pm 4.30972i, \pm 4.68777i$
			$(x, z) = (0.07485926, \pm 0.48895229)$	$(x, z) = (0.01129513, \pm 1.21242661)$
-1.3	0.7	5.0	$\pm 0.95519i, \pm 4.01241i, \pm 5.74354i$	$\pm 0.30080, \pm 4.72488i, \pm 5.26934i$
			$(x, z) = (0.07062798, \pm 0.40300427)$	$(x, z) = (0.01412884, \pm 0.97512312)$
-1.2	0.6	5.5	$\pm 0.85759i, \pm 4.56523i, \pm 6.23885i$	$\pm 0.38428, \pm 5.12248i, \pm 5.86583i$
			$(x, z) = (0.06639573, \pm 0.31682111)$	$(x, z) = (0.01715166, \pm 0.78085851)$

$$\begin{aligned}
 a &= 4\omega^2 - \Omega_{\xi\xi}^{(0)} - \Omega_{\eta\eta}^{(0)} - \Omega_{\zeta\zeta}^{(0)}, \\
 b &= \Omega_{\xi\xi}^{(0)}\Omega_{\eta\eta}^{(0)} + \Omega_{\eta\eta}^{(0)}\Omega_{\zeta\zeta}^{(0)} + \Omega_{\zeta\zeta}^{(0)}\Omega_{\xi\xi}^{(0)} - 4\omega^2\Omega_{\zeta\zeta}^{(0)} - [\Omega_{\xi\zeta}^{(0)}]^2, \\
 c &= [\Omega_{\xi\zeta}^{(0)}]^2\Omega_{\eta\eta}^{(0)} - \Omega_{\xi\xi}^{(0)}\Omega_{\eta\eta}^{(0)}\Omega_{\zeta\zeta}^{(0)},
 \end{aligned}
 \tag{13}$$

which is a polynomial of sixth degree in λ . The eigenvalues of the characteristic equation (12) determine the stability or instability of the respective equilibrium points. An equilibrium point is linearly stable only when all roots of the characteristic equation for λ are pure imaginary. Otherwise, the equilibrium point is unstable.

As a particular example, we compute the characteristic roots $\lambda_i, i = 1, 2, \dots, 6$, which are shown in Table 3 for different values of μ and for a wide range of the remaining parameters ω, q_1 , and q_2 . In addition, in Table 4, we provide sample cases at which the values of the corresponding roots are all purely imaginary (equilibrium points $L_{1,2}^z$), thus leading to stability, while for the equilibria $L_{3,4}^z$ we get two opposite real roots and four imaginary, which means that due to the real roots these points are unstable. So, our analysis reveals that there are cases in which the eigenvalues are all imaginary, and this leads to linear stability of the out-of-plane equilibrium points.

4 Spatial Periodic Orbits Around the Out-of-Plane Equilibria

In this section, we describe the evolution of the families of three-dimensional periodic orbits emanating from the out-of-plane equilibrium points. Our study is divided into two parts. Firstly, based on the Poincaré-Lindstedt method, a second order semi-analytical solution of a three-dimensional periodic orbit around these equilibria will be determined. Then, this semi-analytical solution will be used for the determination of appropriate initial conditions around these points for the numerical integration of the full equations of motion (1).

4.1 The Analytical Approximation

In order to obtain the analytical solution, we initially expand the equations of motion (1) around an out-of-plane equilibrium point up to second order terms obtaining:

$$\begin{aligned} \ddot{\xi} - 2\omega\dot{\eta} &= A_{100}\xi + A_{001}\zeta + A_{101}\xi\zeta + A_{200}\xi^2 + A_{020}\eta^2 + A_{002}\zeta^2 = f(\xi, \eta, \zeta), \\ \ddot{\eta} + 2\omega\dot{\xi} &= B_{010}\eta + B_{110}\xi\eta + B_{011}\eta\zeta = g(\xi, \eta, \zeta), \\ \ddot{\zeta} &= C_{100}\xi + C_{001}\zeta + C_{200}\xi^2 + C_{020}\eta^2 + C_{002}\zeta^2 + C_{101}\xi\zeta = h(\xi, \eta, \zeta), \end{aligned} \tag{14}$$

where the coefficients of the first equation are given by the following formulae:

$$\begin{aligned} A_{100} &= \Omega_{\xi\xi}^{(0)}, \quad A_{001} = \Omega_{\xi\zeta}^{(0)}, \quad A_{020} = \frac{A_{001}}{2z_0}, \\ A_{101} &= 3z_0 \left[Q_1 \left(\frac{1}{r_{10}^5} - \frac{5\alpha^2}{r_{10}^7} \right) + Q_2 \left(\frac{1}{r_{20}^5} - \frac{5\beta^2}{r_{20}^7} \right) \right], \\ A_{200} &= 3\alpha Q_1 \left(\frac{1}{r_{10}^5} + \frac{z_0^2 - 4\alpha^2}{2r_{10}^7} \right) + 3\beta Q_2 \left(\frac{1}{r_{20}^5} + \frac{z_0^2 - 4\beta^2}{2r_{20}^7} \right), \\ A_{002} &= \frac{3\alpha Q_1(\alpha^2 - 4z_0^2)}{2r_{10}^7} + \frac{3\beta Q_2(\beta^2 - 4z_0^2)}{2r_{20}^7}, \end{aligned}$$

and the coefficients of the second one by

$$B_{010} = \Omega_{\eta\eta}^{(0)}, \quad B_{011} = 3z_0 \left(\frac{Q_1}{r_{10}^5} + \frac{Q_2}{r_{20}^5} \right), \quad B_{110} = \frac{A_{001}}{2},$$

while the coefficients of the third equation have been abbreviated as

$$\begin{aligned}
 C_{001} &= \Omega_{\zeta\zeta}^{(0)}, \quad C_{100} = \Omega_{\zeta\xi}^{(0)} = A_{001}, \quad C_{020} = \frac{B_{011}}{2}, \quad C_{101} = 2A_{002}, \\
 C_{200} &= \frac{3Q_1z_0(z_0^2 - 4\alpha^2)}{2r_{10}^7} + \frac{3Q_2z_0(z_0^2 - 4\beta^2)}{2r_{20}^7}, \\
 C_{002} &= 3z_0 \left[\frac{Q_1(3\alpha^2 - 2z_0^2)}{2r_{10}^7} + \frac{Q_2(3\beta^2 - 2z_0^2)}{2r_{20}^7} \right],
 \end{aligned}$$

where we have also abbreviated $Q_1 = q_1(1 - \mu)$, $Q_2 = q_2\mu$, and $\alpha = x_0 + \mu$, $\beta = x_0 + \mu - 1$. We look for solutions of system (14) in powers of a small orbital parameter ε of the following form:

$$\begin{aligned}
 \xi(\tau) &= \xi_1\varepsilon + \xi_2\varepsilon^2 + \mathcal{O}(\varepsilon^3), \\
 \eta(\tau) &= \eta_1\varepsilon + \eta_2\varepsilon^2 + \mathcal{O}(\varepsilon^3), \\
 \zeta(\tau) &= \zeta_1\varepsilon + \zeta_2\varepsilon^2 + \mathcal{O}(\varepsilon^3),
 \end{aligned} \tag{15}$$

where time t has been strained through the transformation:

$$t = (1 + \kappa)\tau, \quad \kappa = b_1\varepsilon + b_2\varepsilon^2 + \mathcal{O}(\varepsilon^3), \tag{16}$$

and b_1 and b_2 have to be determined in order to avoid any secular term. So, by introducing τ as the new variable, system (14) takes the form:

$$\begin{aligned}
 (1 + \kappa)^{-2}\xi'' - 2\omega(1 + \kappa)^{-1}\eta' &= f(\xi, \eta, \zeta), \\
 (1 + \kappa)^{-2}\eta'' + 2\omega(1 + \kappa)^{-1}\xi' &= g(\xi, \eta, \zeta), \\
 (1 + \kappa)^{-2}\zeta'' &= h(\xi, \eta, \zeta).
 \end{aligned} \tag{17}$$

By substituting (15) into the above system (17) and by equating the coefficients of the same powers of ε , we obtain the following two systems. The first system corresponds to the linear terms with respect to the orbital parameter ε :

$$\begin{aligned}
 \xi_1'' - 2\omega\eta_1' - A_{100}\xi_1 - A_{001}\zeta_1 &= 0, \\
 \eta_1'' + 2\omega\xi_1' - B_{010}\eta_1 &= 0, \\
 \zeta_1'' - C_{001}\zeta_1 - A_{001}\xi_1 &= 0,
 \end{aligned} \tag{18}$$

while the second one has been arisen by the second order terms with respect to the orbital parameter ε :

$$\begin{aligned}
 \xi_2'' - 2\omega\eta_2' - A_{100}\xi_2 - A_{001}\zeta_2 &= A_{200}\xi_1^2 + A_{101}\xi_1\zeta_1 + A_{002}\zeta_1^2 + A_{020}\eta_1^2, \\
 \eta_2'' + 2\omega\xi_2' - B_{010}\eta_2 &= B_{110}\xi_1\eta_1 + B_{011}\eta_1\zeta_1, \\
 \zeta_2'' - A_{001}\xi_2 - C_{001}\zeta_2 &= C_{200}\xi_1^2 + C_{002}\zeta_1^2 + C_{020}\eta_1^2 + C_{101}\xi_1\zeta_1.
 \end{aligned} \tag{19}$$

In order to obtain the second order semi-analytical solution around the out-of-plane equilibria, the above two systems have to be solved successively. System (18) admits the following general solution:

$$\xi_1(\tau) = \sum_{i=1}^6 \alpha_i e^{\lambda_i \tau}, \quad \eta_1(\tau) = \sum_{i=1}^6 \beta_i e^{\lambda_i \tau}, \quad \zeta_1(\tau) = \sum_{i=1}^6 \gamma_i e^{\lambda_i \tau}, \quad (20)$$

where λ_i , $i = 1, 2, \dots, 6$ are the roots of the characteristic equation (12) and α_i , β_i , γ_i , $i = 1, 2, \dots, 6$, are arbitrary constants. If a pair of roots of the characteristic equation are purely imaginary, i.e., $\lambda_{1,2} = \pm w_0 i$, then the special solution of the general solution (20) corresponding to $\alpha_i = \beta_i = \gamma_i = 0$, $i = 3, 4, 5, 6$, is periodic with period $T = 2\pi/w_0$, with w_0 being the respective frequency, and has the following form:

$$\begin{aligned} \xi_1(\tau) &= x_{11} + c_{111} \cos(w_0 \tau) + s_{111} \sin(w_0 \tau), \\ \eta_1(\tau) &= y_{11} + c_{121} \cos(w_0 \tau) + s_{121} \sin(w_0 \tau), \\ \zeta_1(\tau) &= z_{11} + c_{131} \cos(w_0 \tau) + s_{131} \sin(w_0 \tau), \end{aligned} \quad (21)$$

where the involved coefficients have to be determined. Substitution of the last special solution into (18) leads to the requested first order solution:

$$\xi_1(\tau) = s_{111} \sin(w_0 \tau), \quad \eta_1(\tau) = c_{121} \cos(w_0 \tau), \quad \zeta_1(\tau) = \sin(w_0 \tau), \quad (22)$$

with

$$s_{111} = -\frac{A_{001}(B_{010} + w_0^2)}{\Lambda}, \quad c_{121} = -\frac{2\omega A_{001} w_0}{\Lambda},$$

and $\Lambda = w_0^4 + (A_{100} + B_{010} - 4\omega^2)w_0^2 + A_{100}B_{010}$, while the remaining coefficients of (21) have been eliminated in the process. We substitute now the determined solution (22) into the RHSs of system (19) and look for a respective second order solution of the form:

$$\begin{aligned} \xi_2(\tau) &= x_{21} + c_{211} \cos(w_0 \tau) + s_{211} \sin(w_0 \tau) + c_{212} \cos(2w_0 \tau) + s_{212} \sin(2w_0 \tau), \\ \eta_2(\tau) &= y_{21} + c_{221} \cos(w_0 \tau) + s_{221} \sin(w_0 \tau) + c_{222} \cos(2w_0 \tau) + s_{222} \sin(2w_0 \tau), \\ \zeta_2(\tau) &= z_{21} + c_{231} \cos(w_0 \tau) + s_{231} \sin(w_0 \tau) + c_{232} \cos(w_0 \tau) + s_{232} \sin(w_0 \tau), \end{aligned} \quad (23)$$

where the unknown coefficients involved in this solution have also to be determined. The above substitutions result to the following linear system:

$$\begin{bmatrix} -A_{100} & 0 & 0 & -A_{001} & 0 \\ 0 & -A_{100} - 4w_0^2 & -4\omega w_0 & 0 & -A_{001} \\ 0 & -4\omega w_0 & -B_{010} - 4w_0^2 & 0 & 0 \\ -A_{001} & 0 & 0 & -C_{001} & 0 \\ 0 & -A_{001} & 0 & 0 & -C_{001} - 4w_0^2 \end{bmatrix} \begin{bmatrix} x_{21} \\ c_{212} \\ s_{222} \\ z_{23} \\ c_{232} \end{bmatrix} = \begin{bmatrix} \Theta_1 \\ \Theta_2 \\ \Theta_3 \\ \Theta_4 \\ \Theta_5 \end{bmatrix},$$

where $\Theta_i, i = 1, 2, \dots, 5$ are given by

$$\begin{aligned} \Theta_1 &= \frac{1}{4z_0} A_{001} c_{121}^2 + \frac{1}{2} [A_{002} + s_{111} (A_{101} + A_{200} s_{111})], \\ \Theta_2 &= \frac{1}{4z_0} A_{001} c_{121}^2 - \frac{1}{2} [A_{002} + s_{111} (A_{101} + A_{200} s_{111})], \\ \Theta_3 &= \frac{1}{2z_0} [c_{121} (A_{001} s_{111} + B_{011} z_0)], \\ \Theta_4 &= \frac{1}{4} [B_{011} c_{121}^2 + 2C_{002} + 2s_{111} (2A_{002} + s_{111} C_{200})], \\ \Theta_5 &= \frac{1}{4} [B_{011} c_{121}^2 - 2C_{002} - 2s_{111} (2A_{002} + s_{111} C_{200})], \end{aligned}$$

which has to be solved for the determination of the involved unknown coefficients of (23), while the remaining coefficients of this solution have been found to be equal to zero. The requested second order semi-analytical solution is finally obtained in the form:

$$\begin{aligned} \xi_2(\tau) &= x_{21} + c_{212} \cos(2w_0\tau), \quad \eta_2(\tau) = s_{222} \sin(2w_0\tau), \\ \zeta_2(\tau) &= z_{23} + c_{232} \cos(2w_0\tau), \end{aligned} \tag{24}$$

where the respective coefficients have been determined by solving the aforementioned linear system and are

$$x_{21} = \frac{\Psi_1}{\Psi}, \quad z_{23} = \frac{\Psi_2}{\Psi}, \quad c_{212} = \frac{\Phi_1}{\Phi}, \quad s_{222} = \frac{\Phi_2}{2\Phi}, \quad c_{232} = \frac{\Phi_3}{\Phi},$$

with

$$\begin{aligned} \Psi_1 &= 2z_0 [s_{111}^2 (A_{200} C_{001} - A_{001} C_{200}) + s_{111} (A_{101} C_{001} - 2A_{001} A_{002}) \\ &\quad + A_{002} C_{001} - A_{001} C_{002}] + A_{001} c_{121}^2 (C_{001} - B_{011} z_0), \\ \Psi_2 &= -2z_0 [s_{111}^2 (A_{001} A_{200} - A_{100} C_{200}) + s_{111} (A_{001} A_{101} - 2A_{002} A_{100}) \\ &\quad + A_{001} A_{002} - A_{100} C_{002}] + c_{121}^2 (A_{100} B_{011} z_0 - A_{001}^2), \\ \Phi_1 &= 2z_0 \varphi_2 s_{111}^2 (A_{001} C_{200} - A_{200} \varphi_3) - 2z_0 \varphi_2 s_{111} (A_{101} C_{001} - 2A_{001} A_{002}) \end{aligned}$$

$$\begin{aligned}
& +4A_{101}w_0^2) + A_{001}\varphi_2c_{121}^2(\varphi_3 - z_0B_{011}) + 2z_0\varphi_2(A_{001}C_{002} - A_{002}\varphi_3) \\
& -8\omega w_0\varphi_3c_{121}(A_{001}s_{111} + z_0B_{011}), \\
\Phi_2 = & -c_{121}(A_{001}s_{111} + B_{011}z_0) \left[A_{001}^2 - A_{100}C_{001} - 4w_0^2(A_{100} + C_{001}) - 16w_0^4 \right] \\
& -4\omega w_0z_0 \left[s_{111}(A_{101}C_{001} - 2A_{001}A_{002} + 4A_{101}w_0^2) - A_{001}C_{002} + A_{002}C_{001} \right. \\
& \quad \left. + 4A_{002}w_0^2 + s_{111}^2(A_{200}C_{001} - A_{001}C_{200} + 4A_{200}w_0^2) \right] \\
& + 2\omega w_0A_{001}c_{121}^2(\varphi_3 - B_{011}z_0), \\
\Phi_3 = & c_{121}^2(B_{011}z_0\varphi_6 - A_{001}^2\varphi_2) + 2z_0s_{111}^2(A_{001}A_{200}\varphi_2 - C_{200}\varphi_6) \\
& + 8\omega w_0A_{001}^2c_{121}s_{111} + 2z_0s_{111}(A_{001}A_{101}\varphi_2 - 2A_{002}\varphi_6) \\
& + 8\omega w_0z_0A_{001}B_{011}c_{121} + 2A_{001}A_{002}z_0\varphi_2 - 2z_0C_{002}\varphi_6,
\end{aligned}$$

and

$$\Psi = 4z_0(A_{001}^2 - A_{100}C_{001}), \quad \Phi = 4z_0\varphi_2A_{001}^2 - 4z_0\varphi_3\varphi_6,$$

while we have also set for abbreviation:

$$\begin{aligned}
\varphi_1 &= 4w_0^2 + A_{001}, \quad \varphi_2 = 4w_0^2 + B_{010}, \quad \varphi_3 = 4w_0^2 + C_{001}, \\
\varphi_4 &= A_{100} + B_{010} + 4\omega^2, \quad \varphi_5 = A_{100}B_{010} + 16w_0^4, \quad \varphi_6 = \varphi_5 + 4w_0^2\varphi_4.
\end{aligned}$$

Eventually, the resulting second order approximate periodic solution around the out-of-plane equilibrium points is obtained in the form:

$$x = x_0 + \xi(\tau), \quad y = \eta(\tau), \quad z = z_0 + \zeta(\tau), \quad (25)$$

where

$$\begin{aligned}
\xi(\tau) &= s_{111} \sin(w_0\tau)\varepsilon + [x_{21} + c_{212} \cos(2w_0\tau)] \varepsilon^2, \\
\eta(\tau) &= c_{121} \cos(w_0\tau)\varepsilon + s_{222} \sin(2w_0\tau)\varepsilon^2, \\
\zeta(\tau) &= \sin(w_0\tau)\varepsilon + [z_{23} + c_{232} \cos(2w_0\tau)] \varepsilon^2,
\end{aligned} \quad (26)$$

while the particular coefficients of this approximate solution have been determined as it was discussed in our previous analysis. Note that the coefficients b_1 and b_2 of time transformation (16) have been arbitrarily chosen to be equal to zero in order to avoid secular solutions, which means that finally we get the relation for time transformation $t = \tau$.

In Figure 7, we show approximate initial conditions of three-dimensional periodic orbits as they result from our semi-analytical solution (25) for the values

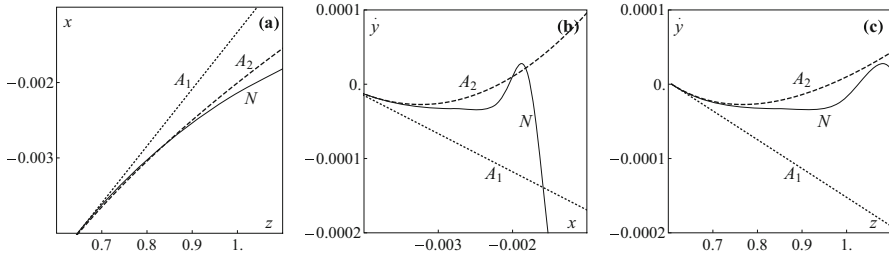


Fig. 7 The validity of the first (A_1) and second (A_2) order semi-analytical solutions w.r.t. the corresponding numerical solution (N) in several projections. The involved solutions have been computed for the parameter values $\mu = 0.1$, $\omega = 3$, $q_1 = -0.01$, and $q_2 = 0.5$

of the parameters of the problem $\mu = 0.1$, $\omega = 3$, $q_1 = -0.01$, and $q_2 = 0.5$, in the range of the orbital parameter $\varepsilon \in [0, 0.45]$. To establish the validity of our analysis, the family of three-dimensional periodic orbits, which has been computed numerically by integrating the full equations of motion (1), is also shown in this figure. The presented solutions are marked by A_1 , for the linear terms of (25), A_2 for the second order analytical solution, and N for the numerical one, respectively. As we observe, in all projections, the second order analytical solution A_2 is a better approximation of the numerical one N than the first order analytical solution A_1 . Note here that, due to the symmetry of the problem, we have chosen to show in this figure initial conditions of the form (x_0, z_0, \dot{y}_0) , as it will be explained afterwards

4.2 The Numerical Approximation

We consider a dynamical system expressed by the following system of first order differential equations:

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}; t), \tag{27}$$

with $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{F} = (F_1, F_2, \dots, F_n) : \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$, where t is the independent variable. A solution \mathbf{x} of system (27) is periodic of period T if it satisfies the following conditions:

$$\mathbf{x}(\mathbf{x}_0; t = 0) = \mathbf{x}(\mathbf{x}_0; t = T), \tag{28}$$

where \mathbf{x}_0 is the initial point of this orbit at $t = 0$. In general, the above conditions do not hold; therefore, given of an initial guess \mathbf{x}_0^* of the initial point of a periodic orbit, we seek the necessary corrections of this point such that:

$$\mathbf{x}(\mathbf{x}_0^* + \delta\mathbf{x}_0^*; t = 0) = \mathbf{x}(\mathbf{x}_0^* + \delta\mathbf{x}_0^*; t = T^* + \delta T^*). \tag{29}$$

In order to solve the above nonlinear system of algebraic equations, we may minimize the following objective function:

$$f(\mathbf{x}_0^* + \delta\mathbf{x}_0^*) = \sum_{i=1}^n [x_i(\mathbf{x}_0^* + \delta\mathbf{x}_0^*; t = 0) - x_i(\mathbf{x}_0^* + \delta\mathbf{x}_0^*; t = T^* + \delta T^*)]^2, \quad (30)$$

arising from the corresponding periodicity conditions (29) by using any optimization method. For the prediction of a next member orbit that belongs to a family of periodic orbits, we work as follows. Suppose that the initial conditions \mathbf{x}_0 and the period T of the member orbit are already known with a predetermined accuracy. Then, to find a new periodic orbit of this family, we have to compute appropriate modifications $\delta\mathbf{x}_0$ and δT of the initial conditions and period, so that the solution with initial point $\mathbf{x}_0 + \delta\mathbf{x}_0$ to be periodic with period $T + \delta T$ thus satisfying the conditions:

$$x_i(\mathbf{x}_0 + \delta\mathbf{x}_0; t = 0) - x_i(\mathbf{x}_0 + \delta\mathbf{x}_0; t = T + \delta T) = \mathbf{0}. \quad (31)$$

By expanding the LHSs of the above equations up to first order terms, we obtain

$$\delta x_{0i} + \sum_{j=1}^n \frac{\partial x_i}{\partial x_{0j}} \delta x_{0j} + \frac{\partial x_i}{\partial t} \delta T = 0, \quad i = 1, \dots, n, \quad (32)$$

and by fixing

$$\sum_{j=1}^n \delta x_{0j}^2 = d^2 = \text{const}, \quad (33)$$

we are able to approximate the δ -modifications by solving (32) and (33), so that the distance between the initial points of the two periodic solutions remains equal to d . Obviously, this prediction does not satisfy the periodicity conditions with the predetermined accuracy; therefore, it can be corrected to give a periodic solution with initial point \mathbf{x}_0^* and period T^* by minimizing the objective function f given in (30) together with the function:

$$h(\mathbf{x}_0) = \left| \sum_{i=1}^n (x_{0i}^* - x_{0i})^2 - d^2 \right|. \quad (34)$$

The minimization of h ensures that the distance between the initial points of the two periodic solutions will remain equal to d .

As an optimization technique for the numerical determination of periodic solutions around the out-of-plane equilibrium points, we have adopted here the algorithms developed by Broyden–Fletcher–Goldfarb–Shanno (BFGS) and Davidon–Fletcher–Powell (DFP), which are briefly described in the following (details for

these two quasi-Newton methods can be found in, e.g., [5]). Consider the following set of objective functions:

$$\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), g_2(\mathbf{x}), \dots, g_n(\mathbf{x})).$$

Minimizing simultaneously all these functions is equivalent to the minimization of function:

$$f(\mathbf{x}) = \sum_{i=1}^n g_i^2(\mathbf{x}),$$

and for the computation of the minima of function f , we use the numerical optimization technique:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \lambda^k \mathbf{A}_k \nabla f(\mathbf{x}^k), \quad k = 0, 1, 2, \dots, \tag{35}$$

where

$$\begin{aligned} \mathbf{A}_{k+1} &= \mathbf{A}_k + \frac{\mathbf{r}^k (\mathbf{r}^k)^\top}{(\mathbf{r}^k)^\top \mathbf{q}^k} - \frac{\mathbf{A}_k \mathbf{q}^k (\mathbf{q}^k)^\top \mathbf{A}_k}{(\mathbf{q}^k)^\top \mathbf{A}_k \mathbf{q}^k} + \gamma (\mathbf{q}^k)^\top \mathbf{A}_k \mathbf{q}^k \mathbf{u}^k (\mathbf{u}^k)^\top, \\ \mathbf{u}^k &= \frac{\mathbf{r}^k}{(\mathbf{r}^k)^\top \mathbf{q}^k} - \frac{\mathbf{A}_k \mathbf{q}^k}{(\mathbf{q}^k)^\top \mathbf{A}_k \mathbf{q}^k}, \quad \mathbf{r}^k = \mathbf{x}^{k+1} - \mathbf{x}^k, \quad \mathbf{q}^k = \nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^k), \end{aligned}$$

with \mathbf{A}_0 being an arbitrary symmetric and positive definite matrix, usually taken to be the identity matrix, and λ^k is the optimal length in the direction $\mathbf{p}^k = -\mathbf{A}_k \nabla f(\mathbf{x}^k)$. For $\gamma = 1$, we obtain the BFGS method, while for $\gamma = 0$ the DFP one. The aforementioned technique has been successfully applied by Kalantonis et al. [9] for the determination of periodic orbits as fixed points on the Poincaré surface of section in a different model problem of Celestial Mechanics.

In the considered problem, there is the symmetry property with respect to the Oxz plane, which means that all the determined periodic orbits are symmetric with it. By exploiting this particular symmetry, we start the integration of the equations of motion with initial position components on this plane and initial velocity perpendicular to it and look for a perpendicular intersection of the orbit with the Oxz plane at the half period, so the mirror theorem will be satisfied (see [22]). Thus, the conditions that must be fulfilled for an orbit emanating from an out-of-plane equilibrium point in order to be periodic are

$$\mathbf{y}(\mathbf{x}_0; t = T/2) = 0, \quad \dot{\mathbf{x}}(\mathbf{x}_0; t = T/2) = 0, \quad \dot{\mathbf{z}}(\mathbf{x}_0; t = T/2) = 0, \tag{36}$$

where $\mathbf{x}_0 = (x_0, y_0 = 0, z_0, \dot{x}_0 = 0, \dot{y}_0, \dot{z}_0 = 0)$ is the initial state vector, i.e., at $t = 0$, and T is the orbit's period. If these conditions are not satisfied, we seek for proper corrections δx_0 , δz_0 , $\delta \dot{y}_0$, and δT of the initial state vector and period, respectively, such that

$$\begin{aligned}
g_1(\delta x_0, \delta z_0, \delta \dot{y}_0; \delta T) &= y(x_0 + \delta x_0, 0, z_0 + \delta z_0, 0, \dot{y}_0 + \delta \dot{y}_0, 0; T/2 + \delta T) = 0, \\
g_2(\delta x_0, \delta z_0, \delta \dot{y}_0; \delta T) &= \dot{x}(x_0 + \delta x_0, 0, z_0 + \delta z_0, 0, \dot{y}_0 + \delta \dot{y}_0, 0; T/2 + \delta T) = 0, \\
g_3(\delta x_0, \delta z_0, \delta \dot{y}_0; \delta T) &= \dot{z}(x_0 + \delta x_0, 0, z_0 + \delta z_0, 0, \dot{y}_0 + \delta \dot{y}_0, 0; T/2 + \delta T) = 0,
\end{aligned} \tag{37}$$

at half of the period. We then form the objective function:

$$G(\delta x_0, \delta z_0, \delta \dot{y}_0; \delta T) = \sum_{i=1}^3 g_i^2(\delta x_0, \delta z_0, \delta \dot{y}_0; \delta T) \tag{38}$$

and apply either the BFGS or DFP algorithms in order to compute the corresponding corrections as it was discussed previously.

By applying the above mentioned technique, several families of symmetric (w.r.t. the Oxz plane) periodic orbits emanating from the out-of-plane equilibrium points have been computed. In order to determine a member of each one of the computed families in the vicinity of an out-of-plane equilibrium point, we have used the analytical solution (25) obtained in the previous subsection, for a relatively small value of the orbital parameter ε , e.g., $\varepsilon = 0.01$, which works as the appropriate seed for our numerical computations. Then, this approximate orbit is corrected numerically with the accuracy of eight significant figures and the family is continued up to its end. Note that the time t in our analytical solution (25) is always selected such that to obtain $y_0 = 0$, $\dot{x}_0 = 0$, and $\dot{z}_0 = 0$, i.e., to start perpendicularly from the Oxz plane. Also, for the numerical determination of the stability of a three-dimensional periodic orbit, we recall here that it can be determined by integrating numerically the respective variational equations (see, e.g., [28]). Such an orbit will be stable if the following conditions hold simultaneously [31]:

$$|P| < 2 \quad \text{and} \quad |Q| < 2, \tag{39}$$

where $P, Q = (2 - \text{Tr } V \pm \sqrt{\Delta})/2$, and $\Delta = (2 - \text{Tr } V)^2 - 2[(2 - \text{Tr } V)^2 + 2 - \text{Tr } V^2] + 8$, while V is the variational matrix. For stability of a three-dimensional periodic orbit in the restricted problem, we may also refer to Perdios [13].

So, we have computed the two families emanating from the unstable out-of-plane equilibrium points L_1^z for the parameter values $\mu = 0.1$, $\omega = 3.0$, $q_1 = -0.01$, $q_2 = 0.5$, and $\mu = 0.3$, $\omega = 0.75$, $q_1 = -0.05$, $q_2 = 0.5$, corresponding to the positions (x_0, z_0) of the out-of-plane equilibrium points given to the second column of the sixth row of Tables 1 and 2, respectively. Both families consist of unstable periodic orbits that continue to exist for large values of z and are presented in Figure 8 where we plot their respective member orbits.

In Figure 9, we present in the space of initial conditions the characteristic curves of the families emanating from the out-of-plane equilibrium points L_1^z and L_3^z , together with their respective projections (light grey), corresponding to the case of the parameter values $\mu = 0.4$, $\omega = 4.0$, $q_1 = 0.9$, and $q_2 = -1.5$, which

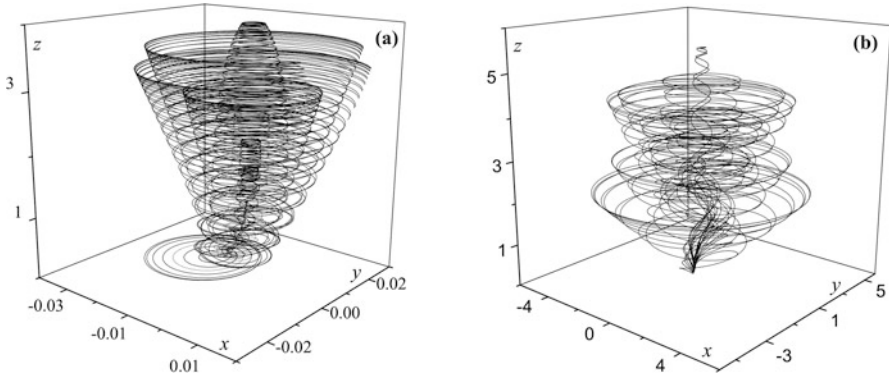


Fig. 8 Periodic orbits of the families emanating from the two unstable out-of-plane equilibrium points L_1^z for the parameter values (a) $\mu = 0.1$, $\omega = 3.0$, $q_1 = -0.01$, $q_2 = 0.5$ and (b) $\mu = 0.3$, $\omega = 0.75$, $q_1 = -0.05$, $q_2 = 0.5$

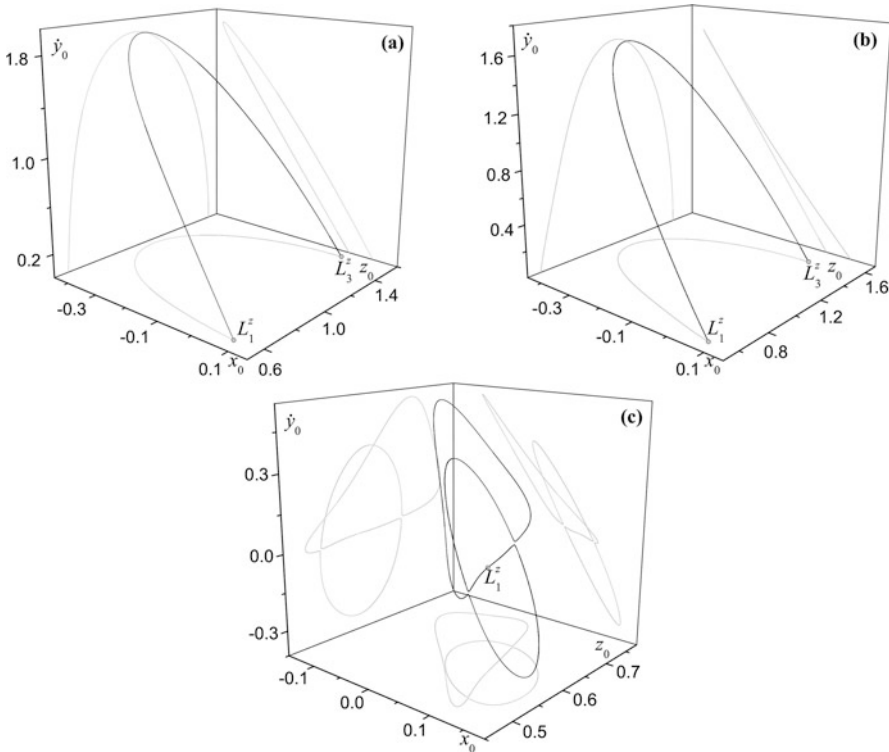


Fig. 9 Characteristic curves of the three families emanating from the stable out-of-plane equilibrium points L_1^z and L_3^z for $\mu = 0.4$, $\omega = 4.0$, $q_1 = 0.9$, and $q_2 = -1.5$

Table 5 Initial conditions for the families emanating from the out-of-plane equilibrium points (0.07943819, 0.57904060) and (0.00845626, 1.53373515) for the parameter values $\mu = 0.4$, $\omega = 4.0$, $q_1 = 0.9$ and $q_2 = -1.5$. For each orbit, we also give the corresponding stability indices P and Q

Family	$T/2$	x_0	z_0	\dot{y}_0	P	Q
f_s	0.67739031	0.07925352	0.57907274	0.00084713	0.99359	-0.54361
	0.70518031	-0.30853936	0.76395514	1.55934870	-0.42186	-1.84500
	0.86262166	0.00845621	1.53273515	0.00000019	-1.33685	-1.92126
f_i	1.01518902	0.07905774	0.57918829	0.00113941	1.99993	0.72469
	0.92247511	-0.19670076	0.74723476	0.85158941	0.87442	-1.00011
	0.84418676	0.00845617	1.53273515	0.00000033	-1.55292	-2.00368
f_l	3.28528233	0.07924189	0.57973515	0.00008423	-0.17409	-1.17465
	3.63898863	0.11070274	0.48690961	-0.01352081	0.93780	0.60193
	3.28528661	0.07967298	0.57821451	-0.00010256	-0.17403	-1.17470

is presented in the first row of Table 4. For this set of values of the parameters, there are two out-of-plane equilibrium points (and their corresponding symmetric w.r.t. the Ox -axis equilibria L_2^z and L_4^z , as well) one of which (L_1^z) is stable and the other one (L_3^z) is unstable as it can be established by the respective eigenvalues presented in that table. From the stable equilibrium point L_1^z , three families emanate from it, one of which has the short period as it can be obtained from the largest frequency (denoted by f_s in this figure), the other that has the long period arising from the smallest frequency (denoted by f_l in the figure), and the last one of the intermediate period (denoted by f_i , respectively) that corresponds to the remaining frequency where its value is between the other two. Accordingly, from the unstable equilibrium point L_3^z , two families emanate from it since its stability analysis has shown that it possesses two pairs of pure imaginary roots. Therefore, one is the family of the short period orbits and the other family consists of the long period orbits. A remarkable result is that two families emanating from the stable out-of-plane equilibrium point L_1^z terminate at the unstable out-of-plane equilibrium point L_3^z or vice versa interconnecting these equilibria. The third family emanating from L_1^z terminates at itself.

In Table 5, we give three initial conditions for sample members of each one of the three previously mentioned computed families. For each family, the first set of initial conditions corresponds to a three-dimensional periodic orbit near the beginning of the family, i.e., in the immediate neighborhood of the equilibrium point, while the third set to the family’s termination means that it also corresponds to a similar periodic orbit around the equilibrium. In particular, in this table, we provide the half of the period $T/2$, the components of the positions x_0 , z_0 , and velocity \dot{y}_0 on the Oxz plane as well as the stability properties by presenting the values of the stability indices P and Q as they were defined by (39).

5 Discussion and Conclusion

We studied the photogravitational Chermnykh's restricted three-body problem in terms of its three-dimensional dynamical properties and found that the equations of motion given in the literature allow the existence of out of the orbital plane equilibrium points. Our aim was not to obtain any particular application of this problem but to gain some new dynamical features that have not been observed for the classical restricted three-body problem. In particular, it was found that due to the symmetry of the problem the out-of-plane equilibrium points always appear in pairs so two or four such points may lie on the Oxz plane in symmetrical positions with respect to the Oxy plane, i.e., in the form $(x_0, \pm z_0)$. It was observed that the involved parameters of the problem not only affect the number and positions of the corresponding equilibria but they play significant role on their stability since it was identified that there are values of these parameters for which the respective points may be linearly stable.

Our main results concentrated on periodic motion around the out-of-plane equilibrium points. To this purpose, by using the Lindstedt–Poincaré method, we determined an approximate analytical solution for periodic orbits around them. This solution was then used for obtaining appropriate initial conditions for the computation of the whole families of three-dimensional periodic orbits emanating from these points. In the case of stable equilibrium points, three such families may originate, one of which has the short period, the other is with the long period, and the last one that has the period corresponding to the frequency value between the largest and smallest frequencies. For the numerical determination of a three-dimensional periodic orbit, the periodicity conditions that must be fulfilled compose a system of nonlinear algebraic equations, and its solution is usually obtained by applying any numerical technique for solving such nonlinear systems. However, in our approach, we used the periodicity conditions in order to derive a proper objective function and look for optimizing it. For the optimization treatment of the obtained objective function, we applied the Broyden–Fletcher–Goldfarb–Shanno and Davidon–Fletcher–Powell algorithms.

The existence of out-of-plane equilibrium points gives rise to new dynamical features that deserve to be also explored. So, except from periodic orbits around these points, a case that was studied here, we may look for three-dimensional homoclinic or heteroclinic connections between these orbits by computing the relevant invariant manifolds, a case that is of special importance in space mission design. Another interesting case of motion, which can also be investigated, is asymptotic motion to the out-of-plane equilibrium points themselves, i.e., to seek three-dimensional asymptotic solutions that depart asymptotically from an out-of-plane equilibrium point and arrive asymptotically at the same or another such point. This case of motion traps for infinite time the particle of infinitesimal mass in the vicinity of an unstable equilibrium point and may be useful for future space colonization.

References

1. S.V. Chermnykh, Stability of libration points in a gravitational field. *Leningradskii Universitet Vestnik Matematika Mekhanika Astronomiia* **2**, 73–77 (1987)
2. Y.A. Chernikov, The photogravitational restricted three-body problem. *Sov. Astron.* **14**, 176–181 (1970)
3. M.K. Das, P. Narang, S. Mahajan, M. Yussa, On out of plane equilibrium points in photo-gravitational restricted three-body problem. *J. Astrophys. Astron.* **30**, 177–185 (2009)
4. M.T. de Bustos, M.A. Lopez, R. Martinez, J.A. Vera, On the periodic solutions emerging from the equilibria of the Hill Lunar problem with oblateness. *Qual. Theory Dyn. Syst.* **17**, 331–344 (2018)
5. J. Dennis, R.B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations* (SIAM, Philadelphia, 1996)
6. K. Gozdziowski, A.J. Maciejewski, Nonlinear stability of the Lagrangian libration points in the Chermnykh problem. *Celest. Mech. Dyn. Astron.* **70**, 41–58 (1988)
7. A.A. Hussain, A. Umar, Generalized out-of-plane equilibrium points in the frame of elliptic restricted three-body problem: Impact of oblate primary and Luminous-triaxial secondary. *Adv. Astron.* **2019**, 3278946 (2019)
8. I.G. Jiang, L.C. Yeh, On the Chermnykh-like problems: I. The mass parameter $\mu = 0.5$. *Astrophys. Space Sci.* **305**, 341–348 (2006)
9. V.S. Kalantonis, E.A. Perdios, A.E. Perdiou, O. Ragos, M.N. Vrahatis, On the application of optimization methods to the determination of members of families of periodic solutions. *Astrophys. Space Sci.* **288**, 581–590 (2003)
10. Z.E. Musielak, B. Quarles, The three-body problem. *Rep. Progr. Phys.* **77**, 065901 (2014)
11. K.E. Papadakis, The 3D restricted three-body problem under angular velocity variation. *Astron. Astrophys.* **425**, 1133–1142 (2004)
12. K.E. Papadakis, Numerical exploration of Chermnykh’s problem. *Astrophys. Space Sci.* **299**, 67–81 (2005)
13. E.A. Perdios, The manifolds of families of 3D periodic orbits associated to Sitnikov motions in the restricted three-body problem. *Celest. Mech. Dyn. Astron.* **99**, 85–104 (2007)
14. E.A. Perdios, O. Ragos, Asymptotic and periodic motion around collinear equilibria in Chermnykh’s problem. *Astron. Astrophys.* **414**, 361–371 (2004)
15. A.E. Perdiou, A.A. Nikaki, E.A. Perdios, Periodic motions in the spatial Chermnykh restricted three-body problem. *Astrophys. Space Sci.* **345**, 57–66 (2013)
16. E.A. Perdios, V.S. Kalantonis, A.E. Perdiou, A.A. Nikaki, Equilibrium points and related periodic motions in the restricted three-body problem with angular velocity and radiation effects. *Adv. Astron.* **2015**, 473–483 (2015)
17. A.A. Perezhogin, Stability of the sixth and seventh libration points in the photogravitational restricted circular three-body problem. *Sov. Astron. Lett.* **2**, 174–175 (1976)
18. V. Radzievskii, The restricted problem of three bodies taking account of light pressure. *Astron. Zh.* **27**, 250–256 (1950)
19. V. Radzievskii, The space photogravitational restricted three-body problem. *Astron. Zh.* **30**, 265–269 (1953)
20. O. Ragos, C. Zagouras, The zero velocity surfaces in the photogravitational restricted three-body problem. *Earth Moon Planets* **41**, 257–278 (1988)
21. O. Ragos, C. Zagouras, Periodic solutions about the ‘out of plane’ equilibrium points in the photogravitational restricted three-body problem. *Celest. Mech. Dyn. Astron.* **44**, 135–154 (1988)
22. A.E. Roy, M.W. Ovenden, On the occurrence of commensurable mean motions in the solar system II. The mirror theorem. *Mon. Not. R. Astron. Soc.* **115**, 296–309 (1955)
23. D.W. Schuerman, The restricted three-body problem including radiation pressure. *Astrophys. J.* **238**, 337–342 (1980)

24. Shankaran, J.P. Sharma, B. Ishwar, Out-of-plane equilibrium points and stability in the generalised photogravitational restricted three body problem. *Astrophys. Space Sci.* **332**, 115–119 (2011)
25. 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)
26. J. Singh, Motion around the Out of plane equilibrium points in the generalized perturbed restricted three-body problem. *Astrophys. Space Sci.* **342**, 303–310 (2012)
27. J. Singh, A.E. Vincent, Out-of-plane equilibrium points in the photogravitational restricted four-body problem. *Astrophys. Space Sci.* **359**, 38 (2015)
28. J. Singh, A.E. Perdiou, J.M. Gyegwe, V.S. Kalantonis, Periodic orbits around the collinear equilibrium points for binary Sirius, Procyon, Luhman 16, α -Centuari and Luyten 726-8 systems: the spatial case. *J. Phys. Commun.* **1**, 025008 (2017)
29. M. Valtonen, H. Karttunen, *The Three-Body Problem* (Cambridge University Press, Cambridge, 2006)
30. 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)
31. C. Zagouras, V.V. Markellos, Axisymmetric periodic orbits of the restricted problem in three dimensions. *Astron. Astrophys.* **59**, 79–89 (1977)
32. E.E. Zotos, M.S. Suraj, R. Aggrawal, S.K. Satya, Investigating the basins of convergence in the circular Sitnikov three-body problem with non-spherical primaries. *Few-Body Syst.* **59**, 69 (2018)

Optimal Lot Size with Partial Backlogging Under the Occurrence of Imperfect Quality Items



G. Karakatsoulis and K. Skouri

Abstract In this study, a continuous review inventory system with deterministic demand, partial backlogging, and imperfect quality items is considered. More precisely, the fraction of imperfect quality items is assumed as a random variable with a known distribution function. The order quantity is subjected to a 100%, error-free, screening process, with finite screening rate. After inspection, the imperfect quality items can be classified into two categories: low quality items and defective items. The demand rate is constant and manifests even during screening period. The demand during the stockout period is satisfied partially as soon as stock is available and before the new demand is met. Perfect and imperfect quality items are charged with different holding cost, giving the chance of different treatment for the two categories of products. The objective is to find the order quantity that maximizes the total profit of the system per unit time. Beyond, the analytical properties are established, the impact of imperfect quality and holding costs differentiation are examined and the behavior of the relative error using the EOQ with partial backlogging solution is displayed graphically. Finally, it is shown that this model can be reduced to other models existing in the literature.

1 Introduction

Supply uncertainty in the raw material and production stages of a supply chain impacts on the performance and effectiveness of the whole supply chain. In response to this phenomenon, production-inventory models, which incorporate related issues, have attracted the interest of researchers in recent years. In the literature, several forms of supply uncertainty have appeared: Disruption, yield uncertainty, capacity uncertainty, lead time uncertainty, although the boundaries among them are often indistinguishable. In this study, supplier's delivery capacity is considered as a

G. Karakatsoulis · K. Skouri (✉)

Department of Mathematics, University of Ioannina, Ioannina, Greece

e-mail: kskouri@uoi.gr

© Springer Nature Switzerland AG 2021

T. M. Rassias, P. M. Pardalos (eds.), *Nonlinear Analysis and Global Optimization*,

Springer Optimization and Its Applications 167,

https://doi.org/10.1007/978-3-030-61732-5_12

random variable, so that each lot may contain a random percentage of imperfect quality items. The production of imperfect quality items is a common problem in high-tech and electronic industries. Comprehensive quality control inspections help to prevent defective products from reaching the customer, causing loss of brand loyalty and brand trust and then loss in market share. To this end, a piece by piece inspection could be used to improve quality and minimize or eliminate defects.

In EOQ type models framework, Salameh and Jaber [16] proposed an inventory model with a random proportion of imperfect quality items in each batch, assuming that no non-planned stockouts occur. Cárdenas-Barrón et al. [2] amended a flaw in the optimal order quantity and Goyal and Cárdenas-Barrón [4] proposed an approximation of the objective function. Papachristos and Konstantaras [12] showed that the constraint used by Salameh and Jaber [16] does not necessarily prevent the non-planned shortages. It should be noticed that although this event has small probability, however, this could be vital for some products like medical equipment.

Another issue concerns the storage of imperfect and perfect quality items. Since, usually, the imperfect and perfect quality items are stored in different warehouses with lower running costs, Wahab and Jaber [19] modified the model of Salameh and Jaber [16] assuming that the holding cost of the imperfect quality items is lower from the holding cost of the perfect quality items. Rezaei [14] extended the work of Salameh and Jaber [16] allowing shortages with complete backlogging, but assumed that the backorders are satisfied immediately as soon as a new batch arrives. This is not feasible, since upon arrival, it is not known which items are of perfect quality and which are not. This was corrected by Eroglu and Ozdemir [3], assuming also that the imperfect quality items can be classified as low quality and defective. Wang et al. [20] extended the model of Salameh and Jaber [16] assuming partial backlogging. Jaber et al. [6] and Alamri et al. [1] investigated models with a reduction in the percentage of imperfect quality items according to a learning curve while Khan et al. [8] and Konstantaras et al. [10] assumed learning in the inspection. Hauck and Vörös [5] treated the inspection rate as a decision variable, assuming that the decision maker could increase the speed of inspection through the corresponding investment. Jaber et al. [7] assumed zero reorder point and repair of the imperfect quality items while Taleizadeh et al. [18] assumed partial backlogging with repair option. A comprehensive review, regarding the extensions/modifications of the EOQ model for imperfect items, suggested by Salameh and Jaber [16], was provided by Khan et al. [9].

In this study, the work of Wang et al. [20] is revisited assuming different holding costs between perfect and imperfect quality items and two classes of imperfect quality products: low quality and defective. A different solution procedure is provided (compared to that of Wang et al. [20]) leading to analytical closed-form solutions. Then, the impact of the imperfect quality items on optimal solution is examined, through numerical examples, in relation to other model parameters, and several managerial insights are provided. Also, the relative error of using the EOQ with partial backlogging solution instead of the optimal one is displayed graphically, highlighting the profit reduction that could cause an inaccurate model.

2 Preliminaries

In this section, the main notation and the assumptions, under which the model is developed, are introduced:

2.1 Assumptions

1. The planning horizon is infinite.
2. The demand rate is known and constant.
3. The lead time is zero.
4. An order, of size Q , is placed once the inventory position drops to B . The time between two consecutive orders is defined as a cycle and it is of length T .
5. Each batch contains a random percentage of imperfect quality items with a known probability density function, $f(p)$. The percentage of imperfect quality items, in a cycle, is independent from the percentage of imperfect quality items in other cycles. The imperfect quality items contain a known percentage, θ , of low quality items while the rest are defective.
6. Each lot is subjected to 100% screening process at a finite rate, x , with $x > D$. The screening process is assumed to be error-free.
7. During the screening process, the perfect quality items which are found per unit time, are more than the demand.
8. The imperfect quality items are sold, as a single batch, at a lower price at the end of the screening process and the defective items are rejected at a cost per unit.
9. During the shortage period, the demand is partially backlogged at a known rate.
10. The backorders are satisfied before the new demand is met.
11. The holding cost of the imperfect quality items is lower from the holding cost of the perfect ones, due to different stock keeping requirements.

2.2 Notation

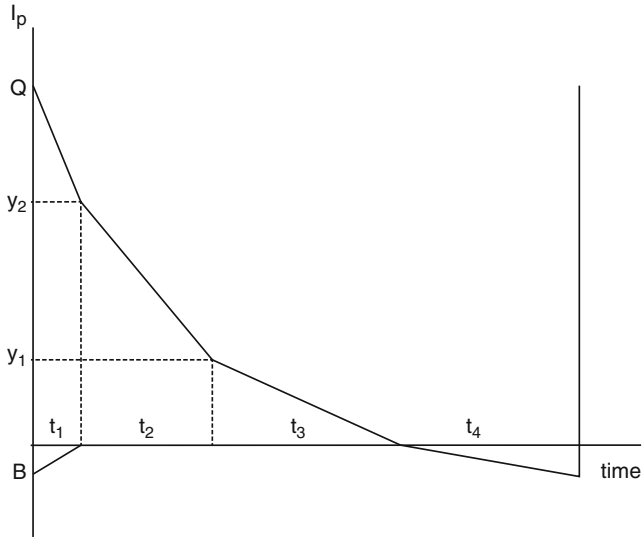
Q	Order Quantity (in units) [decision variable]
B	Maximum level of backorders [decision variable]
D	Demand rate (in units, per unit time)
x	Screening rate (in units, per unit time)
b	Backordering rate of the demand
s	Selling price of the perfect quality items
v	Selling price of the low quality items
K	Fixed ordering cost (per order)
h_g	Holding cost of perfect quality items (per unit, per unit time)

h_d	Holding cost of imperfect quality items (per unit, per unit time)
c_b	Backlogging cost (per unit, per unit time)
c_{ls}	Loss of goodwill cost (per unit)
c_r	Rejection cost
p	Percentage of defective items
$f(p)$	Probability density function of p
θ	Percentage of the low quality items in the imperfect items
T	Length of cycle
z	D/x
A_1	$1 - p - bz$
A	$1 - p - z$

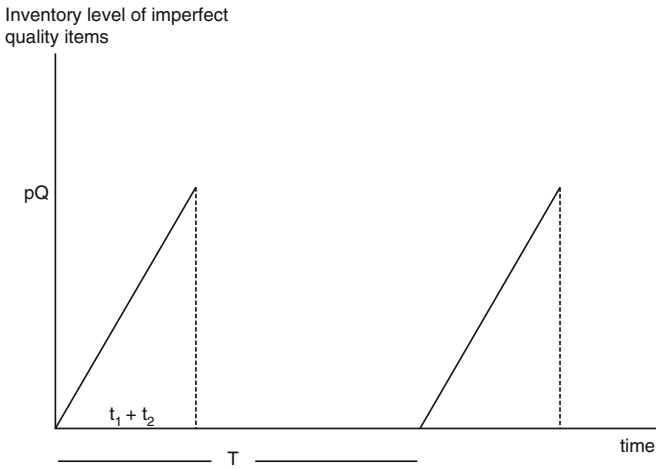
3 Model Formulation and Optimal Policy Determination

Figure 1a represents the inventory level of perfect quality items and the level of backorders during a cycle, while Figure 1b represents the inventory level of imperfect quality items. Specifically, the inventory system operates as follows. At time zero, a new order of size Q is received and the inventory level is equal to Q , since the batch received cannot be used immediately in order to satisfy the backorders, due to quality uncertainty. Then, a screening process starts, which lasts for $t_1 + t_2$ time units. During the time period t_1 the unsatisfied demand from the previous cycle is satisfied in priority to the new demand which is also partially backlogged. During time period t_2 , the screening process proceeds, but there are no backorders any more, so the detected perfect quality items are used in order to satisfy the new demand. Also, during the screening process, the imperfect quality items are stored in a separate warehouse and are withdrawn at the end of it (Figure 1b). During the time period of length t_3 the inventory depletes due to constant demand, while during the time period of length t_4 the system allows for shortages, that are partially backlogged. The reorder interval is comprised of t_1, t_2, t_3, t_4 i.e. it is of length T with $T = t_1 + t_2 + t_3 + t_4$.

The objective is the determination of order quantity and maximum backlogging level that maximize the system profit per unit time. The profit is comprised of the revenue generated from the sales of the perfect and low quality items, the fixed cost per order, the holding costs of the perfect and imperfect items, the backlogging cost, the loss of goodwill cost, and the rejection cost of the defective items. For the calculation of system profit per unit time the Renewal Reward Theorem will be used. Thus, it is necessary to calculate the expected total profit per cycle and the expected cycle length.



(a) Inventory level of perfect quality items over time



(b) Inventory level of imperfect quality items over time

Fig. 1 The inventory level for perfect (a) and imperfect (b) quality items

3.1 Cycle Length

In this section, the expected cycle length is derived. As it has already been noticed, the cycle length is

$$T = t_1 + t_2 + t_3 + t_4 \tag{1}$$

During the shortage period of length t_4 , a fraction b of demand is backlogged, while the rest $1 - b$ is lost. So,

$$t_4 = \frac{B}{bD} \tag{2}$$

Because of the assumption (7), the relation

$$P((1 - p)x > D) = 1 \Leftrightarrow P(p < 1 - z) = 1$$

must hold.

During the time interval $[0, t_1]$, x items are screened per unit time and $(1 - p)x$ are of perfect quality. Thus, the backlogged quantity of the previous cycle is satisfied at a rate of $(1 - p)x$. At the same time, the new demand cannot be satisfied and due to partial backlogging, new backorders are created at a rate b . Hence, at time t_1 the accumulated unsatisfied demand of previous cycle, B , is satisfied at a rate of $(1 - p)x - bD$, and consequently

$$t_1 = \frac{B}{(1 - p)x - bD} = \frac{B}{xA_1} \tag{3}$$

So, taking into account that the inventory level of perfect quality items decreases at a rate of x , the inventory level at time instant t_1 is

$$y_2 = Q - xt_1 = Q - \frac{B}{A_1} \tag{4}$$

Furthermore, during the screening process, Q items are screened at a rate x and the screening process lasts for $t_1 + t_2$ time units so

$$t_1 + t_2 = \frac{Q}{x}$$

Hence,

$$t_2 = \frac{Q}{x} - t_1 = \frac{Q}{x} - \frac{B}{xA_1} \tag{5}$$

During $[t_1, t_1 + t_2]$, the demand is D units per unit time, and px imperfect quality items are identified per unit time. Hence, the inventory level of perfect quality items decreases at a rate of $D + px$. At time instant t_1 the inventory level of perfect quality items is y_2 (see Figure 1a). Hence, at time instant $t_1 + t_2$ the inventory level of perfect quality items is

$$y_1 = y_2 - (D + px)t_2 = A Q - \frac{A}{A_1} B \tag{6}$$

During the time period of length t_3 the inventory level depletes at a rate D with initial inventory level y_1 . Therefore,

$$t_3 = \frac{y_1}{D} = \frac{A}{D} Q - \frac{A}{DA_1} B \tag{7}$$

Finally, from (1), (2), (3), (5), and (7), the cycle length is

$$T = \frac{1}{D} \left[(1 - p)Q + \frac{(1 - b)(1 - p)}{bA_1} B \right] \tag{8}$$

with expected value

$$ET = \frac{1}{D} \left[E(1 - p)Q + \frac{(1 - b)}{b} E \left(\frac{1 - p}{A_1} \right) B \right] \tag{9}$$

3.2 Total Profit Formulation

In order to derive the profit, the revenue is firstly calculated. The revenue is comprised of the sales of perfect and low quality items, so at a cycle the total revenue is

$$TR_c(Q, B) = [s(1 - p) + vp\theta]Q \tag{10}$$

Then, in order to derive the total cost of the system per cycle, the holding cost and the shortage cost are calculated.

The inventory level of perfect quality items at time t is

$$I_p(t) = \begin{cases} Q - xt, & t \in [0, t_1] \\ y_2 - (D + px)t, & t \in [t_1, t_1 + t_2] \\ y_1 - Dt, & t \in [t_1 + t_2, t_1 + t_2 + t_3] \\ 0, & t \in [t_1 + t_2 + t_3, T] \end{cases} \tag{11}$$

Therefore, the holding cost of the perfect quality items is

$$\frac{h_g}{2D} \left[zQ^2 + A(1 - p) \left(Q - \frac{B}{A_1} \right)^2 \right] \tag{12}$$

The inventory level for imperfect quality items is

$$I_{im}(t) = \begin{cases} pxt, & t \in [0, t_1 + t_2] \\ 0, & \text{elsewhere.} \end{cases} \tag{13}$$

So, the holding cost of the imperfect quality items is

$$\frac{h_d}{2D} pzQ^2 \tag{14}$$

The inventory level during shortages period at time t is described as

$$I(t) = \begin{cases} B - xA_1t, & t \in [0, t_1] \\ bDt, & t \in [t_1 + t_2 + t_3, T] \\ 0, & \text{elsewhere} \end{cases} \tag{15}$$

Consequently, the backordering cost per cycle is

$$\frac{c_b}{2D} \left(\frac{z}{A_1} + \frac{1}{b} \right) B^2 \tag{16}$$

Moreover, the loss of goodwill cost is

$$c_{ls} \frac{(1-b)}{b} B + c_{ls} \frac{(1-b)z}{A_1} B \tag{17}$$

Hence, from (16) and (17) the shortage cost per cycle is

$$\frac{c_b}{2bD} \left(\frac{zb}{A_1} + 1 \right) B^2 + \frac{c_{ls}(1-b)}{b} \left(1 + \frac{zb}{A_1} \right) B \tag{18}$$

Finally, taking into account the fixed ordering cost per cycle (i.e. K), the purchase cost (i.e. cQ), and the rejection cost of the defective items (i.e. $c_r(1-p)(1-\theta)Q$), the total cost of the system per cycle is

$$TC_c(Q, B) = K + [c + c_r p(1-\theta)]Q + \frac{h_g}{2D} \left[zQ^2 + A(1-p) \left(Q - \frac{B}{A_1} \right)^2 \right] \tag{19}$$

$$+ \frac{h_d}{2D} pzQ^2 + \frac{c_b}{2bD} \left(\frac{zb}{A_1} + 1 \right) B^2 + \frac{c_{ls}(1-b)}{b} \left(1 + \frac{zb}{A_1} \right) B \tag{20}$$

with expected value

$$ETC_c(Q, B) = K + [c + c_r(1-\theta)E_1]Q + \frac{h_g z}{2D} Q^2 + \frac{h_g}{2D} E_3 Q^2$$

$$\begin{aligned}
 & -\frac{h_g}{D} E_4 Q B + \frac{h_g}{2D} E_5 B^2 + \frac{h_d z}{2D} E(p) Q^2 \\
 & + \frac{c_b}{2bD} E_2 B^2 + \frac{c_{ls}(1-b)}{b} E_2 B
 \end{aligned}$$

where

$$\begin{aligned}
 E_1 &= E(1-p), \quad E_2 = E\left(\frac{1-p}{A_1}\right), \quad E_3 = E[(1-p)A] \\
 E_4 &= E\left[\frac{(1-p)A}{A_1}\right], \quad E_5 = E\left[\frac{(1-p)A}{A_1^2}\right]
 \end{aligned}$$

Using Renewal Reward Theorem, the total profit of the system per unit time is given by

$$TP_{ut}(Q, B) = \frac{ETR_c(Q, B) - ETC_c(Q, B)}{ET} \tag{21}$$

To facilitate the analysis, the transformation $X = ET$ is used, so

$$Q = \frac{D}{E_1} X - \frac{(1-b)E_2}{bE_1} B \tag{22}$$

Also, the following quantities are defined:

$$\alpha = (h_g z + h_g E_3 + h_d z E(p)) \frac{(1-b)^2 E_2^2}{b^2 E_1^2} + h_g E_5 + \frac{c_b E_2}{b} + \frac{2h_g E_4 (1-b) E_2}{b E_1}$$

$$\gamma = \frac{D(1-b)E_2 c_2}{b}$$

$$\delta = (h_g z + h_g E_3 + h_d z E(p)) \frac{D^2}{E_1^2}$$

$$\epsilon = (h_g z + h_g E_3 + h_d z E(p)) \frac{D(1-b)E_2}{bE_1^2} + \frac{Dh_g E_4}{E_1}$$

Then, the total profit per unit time as a function of (X, B) is

$$TP_{ut}(X, B) = \frac{c_1 D}{E_1} - \frac{\gamma B}{DX} - \frac{K}{X} - \frac{\delta X}{2D} + \frac{\epsilon B}{D} - \frac{\alpha B^2}{2DX} \tag{23}$$

where $c_1 = sE_1 + vE(p)\theta - c - c_r(1-\theta)E(p)$ and $c_2 = \frac{c_1}{E_1} + c_{ls}$.

Obviously, the condition $c_1 \geq 0$ must hold. Otherwise, the expected income from the sales would be less than the purchase cost plus the cost of rejection, hence there should not be any profit from system operation.

3.3 Optimal Policy

Aiming to derive the optimal policy for the system under consideration (i.e. optimal values for B and X and so for Q), required analytical results are derived. The next Proposition states properties for the function TP_{ut} , useful for the determination of the optimal policy.

Proposition 1 *The function TP_{ut} is concave in X for given B and concave in B for given X . The function TP_{ut} is concave in (X, B) if*

$$2KD\alpha > \gamma^2 \tag{24}$$

It should be noticed for the EOQ model with partial backlogging, Rosenberg [15] proved that it is optimal to allow for shortages when a relation connecting the cycle length of EOQ model without shortages and the lost sales cost holds. To this end, the model without shortages that is the model analyzed by Wahab and Jaber [19] is used in order for an analogous relation to be developed. So, let T_w :

$$T_w = \sqrt{\frac{2KD}{\delta}} \tag{25}$$

which, is the cycle length derived by Wahab and Jaber [19]. Using the above relation (25) the next Proposition provides the optimal policy for the system under consideration.

Proposition 2

1. *If the inequalities*

$$\alpha\delta > \epsilon^2 \tag{26}$$

and

$$T_w > \frac{\gamma}{\epsilon} \tag{27}$$

hold, then

$$B^* = -\frac{\gamma}{\alpha} + \frac{\epsilon}{\alpha} \sqrt{\frac{2KD\alpha - \gamma^2}{\alpha\delta - \epsilon^2}}$$

$$Q^* = \frac{D}{E_1} \sqrt{\frac{2KD\alpha - \gamma^2}{\alpha\delta - \epsilon^2}} - \frac{(1-b)E_2}{bE_1} B^* \tag{28}$$

If inequality (26) does not hold, then: $B^* = 0, Q^* = 0$

If inequality (27) does not hold, then: $B^* = 0, Q^* = \frac{D}{E_1} T_W$

2. If the inequality (24) does not hold, then either there should not be any inventory system (i.e. $Q^* = 0$ and $B^* = 0$) or

$$B^* = 0$$

$$Q^* = \sqrt{\frac{2KD}{(h_g z + h_g E_3 + h_d z E(p))}}$$

3.4 Special Cases

It is worthwhile to note an additional contribution of the model under consideration; from this model, the already existing models, presented in Table 1, arise.

4 Numerical Comparisons

In this section, numerical comparisons are presented in order to highlight the effects of supply uncertainty, holding costs differentiation and backlogging rate to the optimal policy and profit. To this end, the following set of values for model parameters are used: $D = 1000, h_g = 10, K = 500, c_b = 5, c_{ls} = 1, x = 4000, b = 0.7, \theta = 0.8, s = 50, v = 12, c = 41,$ and $cr = 10$. These values are based on the parametric values used by Wang et al. [20], with required modification. The modification is necessary since the model investigated by Wang et al. [20] aims to minimize the total cost, while in the present model the objective is to maximize the total profit.

Table 1 Special cases arising from the proposed model

If:	Resulting model
$h_d = h_g$ and $\theta = 1$	Wang et al. [20]
$h_d = h_g$ and $b = 1$	Eroglu and Ozdemir [3]
$c_b \rightarrow \infty$ and $\theta = 1$	Wahab and Jaber [19]
$c_b \rightarrow \infty, h_d = h_g$ and $\theta = 1$	Maddah and Jaber [11]
$h_d = h_g, \theta = 1, x \rightarrow \infty$ and $c_b \rightarrow \infty$	Silver [17]
$x \rightarrow \infty$ and $P(p = 0) = 1$	Park [13]

In order to examine the impact, of different holding cost for perfect and imperfect quality items (i.e. $h_d \neq h_g$), on the optimal policy and profit, let us set $\phi = \frac{h_d}{h_g}$, $\phi \in [0, 1]$. This means that as ϕ increases the holding cost of imperfect quality items increases and obviously when $\phi = 1$, perfect and imperfect quality items have the same holding cost. Figure 2 presents the combined impact of imperfect quality items holding cost for various values of defective rate, as expressed through β . For ϕ a step size of 0.1 is used ($\phi \in [0, 1]$), while $p \sim U(0, \beta)$, where $\beta \in (0.1, 0.3)$. Specifically, Figure 2a, c, d present the impact of ϕ on optimal profit, Q^* and B^* respectively, for various values of β , while Figure 2b presents $DP = TP(Q^*, B^*; \phi = i) - TP(Q^*, B^*; \phi = 0.1)$, $i = 0.2, \dots, 1$ (i.e. the decrease in profit).

It is observed that an increase in h_d (for the same level of β) causes a small decrease in profit, which becomes higher as the imperfect quality rate increases. The changes in optimal order quantity and maximum backlogging level are also minor, for the same β . However, the impact of β is considerably high. Notice that the decrease in profit for the case that corresponds to the parameters pair ($\phi = 0.9, \beta = 0.3$) related to the pair ($\phi = 0.1, \beta = 0.1$) is more than 95%. For the same pairs, the increase in the optimal order quantity is more than 25% and the increase in maximum backlogging is greater than 80%, a fact that could justify the decrease in profit, as cost increases due to shortages.

In order to examine the combined effect of the backorder rate, b and the percentage of the imperfect quality items, on the optimal policy and the optimal profit, the following values are used: $\phi = 0.2$, $\theta = 0.8$, $p \sim U(0, \beta)$, where β takes values from 0.1 to 0.3 with a step of size 0.05 and $b \in [0.2, 1]$. Figure 3a, c, d, present the impact of b on optimal profit, Q and B respectively for various values of β , while Figure 3b presents $Pinc = TP(Q^*, B^*; b = i) - TP(Q^*, B^*; b = 0.2)$, $i = 0.25, \dots, 0.95$ (i.e. the increase in profit). The backlogging rate (b) impacts significantly on the system operation and performance. Precisely, the increase of backlogging rate leading to increase in profit from 16% to almost 400% depends on β (the higher impact is caused by higher values of β). Also, the increase in b causes increase in order quantity (approximately 30% increase depends on β). Low values of backlogging rate imply low level of maximum backlogging level (B) for low values of β and higher for higher values of β .

5 Cost Penalty of Employing the EOQ with Partial Backorders

Because of the simplicity of the EOQ model, it is interesting to explore the impact in the total profit of the system, if the EOQ with backorders is used (Q_{EOQ}, B_{EOQ}) instead of the optimal decision variable (Q^*, B^*).

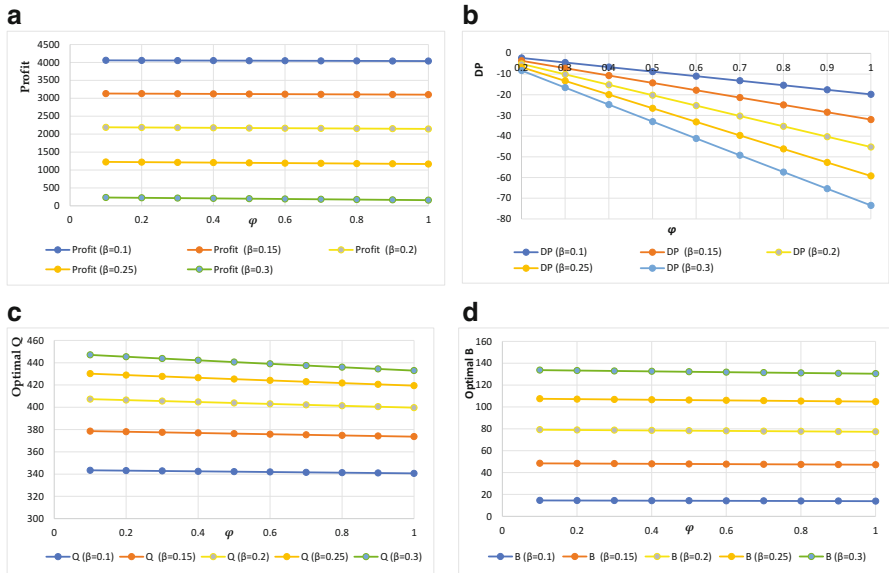


Fig. 2 Impact of the imperfect quality items holding cost for different values of β . **(a)** Impact of h_d increase ($\phi = h_d/h_g, h_g$ constant) on system profit, for various values of β . **(b)** Impact of h_d ($\phi = h_d/h_g, h_g$ constant) increase on system profit decrease, for various values of β . **(c)** Impact of h_d increase ($\phi = h_d/h_g, h_g$ constant) on optimal order quantity, for various values of β . **(d)** Impact of h_d increase ($\phi = h_d/h_g, h_g$ constant) on optimal backlogging quantity, for various values of β

Notice that for $x \rightarrow \infty$, and $P(p = 0) = 1$ the inequality (26) holds and the inequality (27) becomes

$$\sqrt{\frac{2K}{Dh}} > \frac{(1 - b)c_2}{h} \tag{29}$$

where $c_2 = s - c + c_{Is}$.

Suppose that the inequality (29) holds. Then

$$B_{EOQ} = \frac{-b(1 - b)Dc_2 + b\sqrt{h[2KD(h + bc_b) - (1 - b)^2c_2^2D^2]}}{bc_b}$$

$$Q_{EOQ} = \sqrt{\frac{2KD(h + bc_b) - (1 - b)^2D^2c_2^2}{hbc_b}} - \frac{(1 - b)}{b}B_{EOQ} \tag{30}$$

Then computations were performed that depicted in Figure 4 in order to obtain insight into the effect of ignoring the existence of imperfect quality items on system

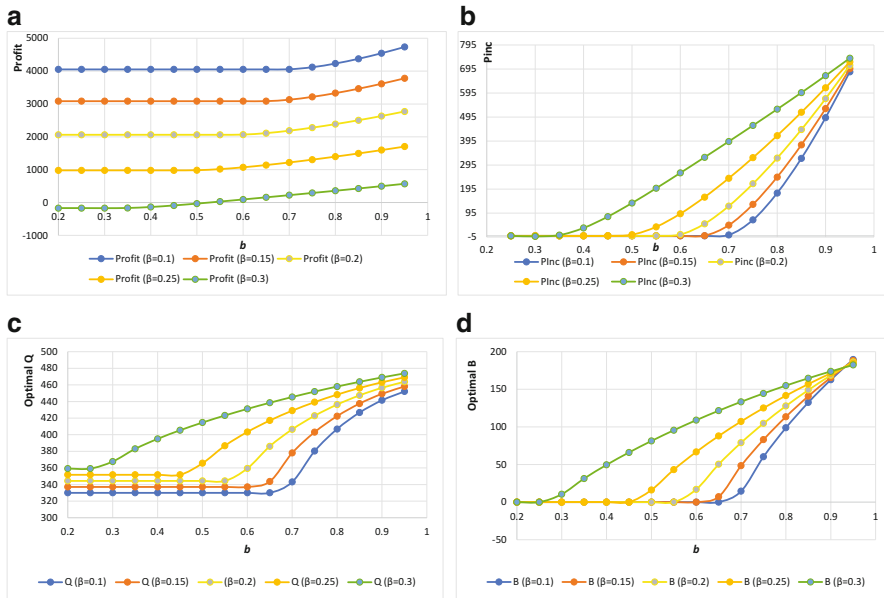


Fig. 3 Impact of backlogging rate, b , for different values of β . **(a)** Impact of b on system profit, for various values of β . **(b)** Impact of b on system profit increase, for various values of β . **(c)** Impact of b on optimal order quantity, for various values of β . **(d)** Impact of b on optimal backlogging quantity, for various values of β

performance. Thus, the EOQ with partial backlogging is used and the following performance indicator is calculated:

$$\Delta = \frac{TP_{ut}(Q_{EOQ}, B_{EOQ}) - TP_{ut}(Q^*, B^*)}{TP_{ut}(Q^*, B^*)} \tag{31}$$

Figure 4 gives the graphical representation of Δ with respect to β using the same parametric values as in previous section. From this figure it seems that ignoring the existence of imperfect quality items may cause significant reduction in system profit, mainly when the percentage of imperfect quality items increases.

6 Conclusions

Supply uncertainty has become an active research area having considerable effects on supply chain performance. In this paper, a single-echelon inventory model under deterministic demand, partial backlogging, and defective items are considered. After modeling long-run average profit, taking into account that the holding cost of perfect and imperfect quality items may be different, the optimal policy is determined in

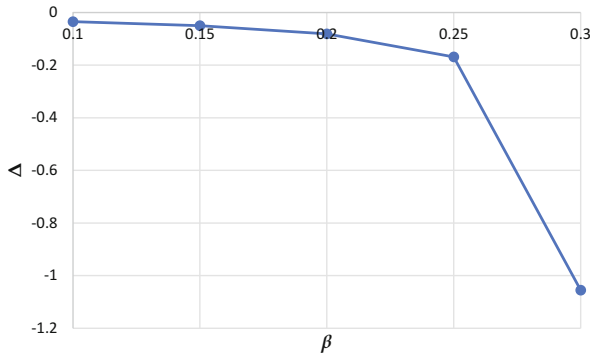


Fig. 4 Δ as a function of β

closed-form. It is worth noting that the form and the solution of this model constitute a direct extension of several existing models in the literature, among others, the classical EOQ (with partial backlogging) model if $P(p = 0) = 1$. In addition, the numerical results demonstrated the clear outperformance of exact representation of system (taking into account the supply uncertainty) in relation to classical EOQ with partial backlogging.

There are several directions for further research. One direction could be to use this model as an approximation for more complex systems with stochastic demand. Another research direction could be the consideration of a non-zero stochastic lead time. Lastly, the consideration of alternative accounting schemes for inventory cost evaluation (like end-of-cycle costing) could be also of interest.

Appendix

Proof of Proposition 1

$$\frac{\partial T P_{ut}(X, B)}{\partial X} = \frac{\gamma B}{DX^2} + \frac{K}{X^2} - \frac{\delta}{2D} + \frac{\alpha B^2}{2DX^2} \tag{32}$$

$$\frac{\partial T P_{ut}(X, B)}{\partial B} = -\frac{\gamma}{DX} + \frac{\epsilon}{D} - \frac{\alpha B}{DX} \tag{33}$$

$$\frac{\partial^2 T P_{ut}(X, B)}{\partial X^2} = -\frac{2\gamma B}{DX^3} - \frac{2K}{X^3} - \frac{\alpha B^2}{DX^3} \tag{34}$$

$$\frac{\partial^2 T P_{ut}(X, B)}{\partial B \partial X} = \frac{\gamma}{DX^2} + \frac{\alpha B}{DX^2} \tag{35}$$

$$\frac{\partial^2 T P_{ut}(X, B)}{\partial B^2} = -\frac{\alpha}{DX} \tag{36}$$

Hence the function TP_{ut} is concave in X for B constant, concave in B for X constant, and the Hessian matrix is negative definite if

$$2KD\alpha > \gamma^2 \tag{37}$$

□

Proof of Proposition 2 If the inequality (24) holds, then in order to maximize $TP_{ut}(X, B)$ it is sufficient to solve the following system of equations:

$$\alpha B^2 + 2\gamma B - \delta X^2 + 2KD = 0 \tag{38}$$

$$\alpha B - \epsilon X + \gamma = 0 \tag{39}$$

Equation (39) always has the solution:

$$X^* = \frac{\alpha}{\epsilon}B + \frac{\gamma}{\epsilon}$$

Replacing X by X^* in (38)

$$\frac{\partial TP_{ut}(X, B)}{\partial X} \Big|_{X=X^*} = \alpha(\epsilon^2 - \alpha\delta)B^2 + 2\gamma(\epsilon^2 - \alpha\delta)B + 2KD\epsilon^2 - \gamma^2\delta \tag{40}$$

is obtained. This quantity has two roots if

$$(\epsilon^2 - \alpha\delta)(\gamma^2 - 2KD\alpha) > 0$$

Because of the inequality (37), the above inequality holds if

$$\epsilon^2 - \alpha\delta < 0$$

or, equivalently

$$h_g^2 E_4^2 < \left(h_g E_5 + \frac{c_b E_2}{b} \right) [h_g z + h_g E_3 + h_d z E(p)]$$

One root is always negative. The other is positive if

$$T_w > \frac{\gamma}{\epsilon}$$

Taking into account the inequality for the concavity of the objective function and making simplifications, it follows that it is optimal to allow shortages if

$$(\epsilon^2 - \alpha\delta)(\gamma^2 - 2KD\alpha) > 0$$

and

$$T_w > \frac{\gamma}{\epsilon}$$

where T_w in (25) is the cycle length in [19] model.

Hence, if the inequalities (26) and (27) hold

$$B^* = -\frac{\gamma}{\alpha} + \frac{\epsilon}{\alpha} \sqrt{\frac{2KD\alpha - \gamma^2}{\alpha\delta - \epsilon^2}}$$

$$X^* = \sqrt{\frac{2KD\alpha - \gamma^2}{\alpha\delta - \epsilon^2}}$$

Using the relation (22) the result is obtained.

Otherwise, $B^* = 0$ and either there should be an inventory system (i.e. $Q^* > 0$) or not. If inequality (26) does not hold, then $\frac{\partial T P_{ut}(X, B)}{\partial X} |_{X=X^*} > 0$, thus the function $T P_{ut}(X, B)$ is increasing in X and the maximum value is obtained for $X \rightarrow \infty$ which gives

$$B^* = 0$$

$$Q^* = 0$$

If inequality (27) does not hold, then

$$B^* = 0$$

$$Q^* = \sqrt{\frac{2KD}{(h_g z + h_g E_3 + h_d z E(p))}}$$

If the inequality (24) does not hold, then either there should not be any inventory

system (i.e. $Q^* = 0 = B^*$) or $Q^* = \sqrt{\frac{2KD}{(h_g z + h_g E_3 + h_d z E(p))}}$ and $B^* = 0$. □

References

1. A.A. Alamri, I. Harris, A.A. Syntetos, Efficient inventory control for imperfect quality items. *Eur. J. Oper. Res.* **254**(1), 92–104 (2016)
2. L.E. Cárdenas-Barrón et al., Observation on: “economic production quantity model for items with imperfect quality” [*Int. J. Prod. Econ.* **64** 59–64 (2000)]. *Int. J. Prod. Econ.* **67**(2), 201–201 (2000)

3. A. Eroglu, G. Ozdemir, An economic order quantity model with defective items and shortages. *Int. J. Prod. Econ.* **106**(2), 544–549 (2007)
4. S.K. Goyal, L.E. Cárdenas-Barrón, Note on: economic production quantity model for items with imperfect quality—a practical approach. *Int. J. Prod. Econ.* **77**(1), 85–87 (2002)
5. Z. Hauck, J. Vörös, Lot sizing in case of defective items with investments to increase the speed of quality control. *Omega* **52**, 180–189 (2015)
6. M. Jaber, S. Goyal, M. Imran, Economic production quantity model for items with imperfect quality subject to learning effects. *Int. J. Prod. Econ.* **115**(1), 143–150 (2008)
7. M.Y. Jaber, S. Zaroni, L.E. Zavanella, Economic order quantity models for imperfect items with buy and repair options. *Int. J. Prod. Econ.* **155**, 126–131 (2014)
8. M. Khan, M. Jaber, M. Wahab, Economic order quantity model for items with imperfect quality with learning in inspection. *Int. J. Prod. Econ.* **124**(1), 87–96 (2010)
9. M. Khan, M. Jaber, A. Guiffrida, S. Zolfaghari, A review of the extensions of a modified EOQ model for imperfect quality items. *Int. J. Prod. Econ.* **132**(1), 1–12 (2011)
10. I. Konstantaras, K. Skouri, M. Jaber, Inventory models for imperfect quality items with shortages and learning in inspection. *Appl. Math. Model.* **36**(11), 5334–5343 (2012)
11. B. Maddah, M.Y. Jaber, Economic order quantity for items with imperfect quality: revisited. *Int. J. Prod. Econ.* **112**(2), 808–815 (2008)
12. S. Papachristos, I. Konstantaras, Economic ordering quantity models for items with imperfect quality. *Int. J. Prod. Econ.* **100**(1), 148–154 (2006)
13. K.S. Park, Inventory model with partial backorders. *Int. J. Syst. Sci.* **13**(12), 1313–1317 (1982)
14. J. Rezaei, Economic order quantity model with backorder for imperfect quality items, in *Proceedings. 2005 IEEE International Engineering Management Conference*, vol 2 (IEEE, Piscataway, 2005), pp. 466–470
15. D. Rosenberg, A new analysis of a lot-size model with partial backlogging. *Naval Res. Logist. Q.* **26**(2), 349–353 (1979)
16. M. Salameh, M. Jaber, Economic production quantity model for items with imperfect quality. *Int. J. Prod. Econ.* **64**(1–3), 59–64 (2000)
17. E. Silver, Establishing the order quantity when the amount received is uncertain. *Inf. Syst. Oper. Res.* **14**(1), 32–39 (1976)
18. A.A. Taleizadeh, M.P.S. Khanbaglo, L.E. Cárdenas-Barrón, An EOQ inventory model with partial backordering and reparation of imperfect products. *Int. J. Prod. Econ.* **182**, 418–434 (2016)
19. M. Wahab, M.Y. Jaber, Economic order quantity model for items with imperfect quality, different holding costs, and learning effects: a note. *Comput. Ind. Eng.* **58**(1), 186–190 (2010)
20. W.T. Wang, H.M. Wee, Y.L. Cheng, C.L. Wen, L.E. Cárdenas-Barrón, EOQ model for imperfect quality items with partial backorders and screening constraint. *Eur. J. Ind. Eng.* **9**(6), 744–773 (2015)

Error Analysis Through Energy Minimization and Stability Properties of Exponential Integrators



Odysseas Kosmas and Dimitrios Vlachos

Abstract In this article, the stability property and the error analysis of higher-order exponential variational integration are examined and discussed. Toward this purpose, at first we recall the derivation of these integrators and then address the eigenvalue problem of the amplification matrix for advantageous choices of the number of intermediate points employed. Obviously, the latter determines the order of the numerical accuracy of the method. Following a linear stability analysis process we show that the methods with at least one intermediate point are unconditionally stable. Finally, we explore the behavior of the energy errors of the presented schemes in prominent numerical examples and point out their excellent efficiency in long term integration.

1 Introduction and Motivation

During the last decades, there have been developed several numerical integration methods for Lagrangian systems, where the integrator is derived by discretizing the Hamilton's principle. This class of integration methods is known as discrete variational integrators and have specific advantages that make them attractive for many applications in mechanical systems. They are appropriate for both conservative and nearly dissipative (forced) systems. The conservative nature of variational integrators can allow substantially more accurate simulations at lower cost [1–4].

In solving numerical ordinary differential equations, one of the most difficult problems is related to the development of integrators for highly oscillatory systems [1]. As is well known, standard numerical schemes may require a huge number of

O. Kosmas

Modelling and Simulation Centre, MACE, The University of Manchester, Manchester, UK
e-mail: odysseas.kosmas@manchester.ac.uk

D. Vlachos (✉)

Department of Economics, University of Peloponnese, Tripoli, Greece
e-mail: dvlachos@uop.gr

© Springer Nature Switzerland AG 2021

T. M. Rassias, P. M. Pardalos (eds.), *Nonlinear Analysis and Global Optimization*,
Springer Optimization and Its Applications 167,
https://doi.org/10.1007/978-3-030-61732-5_13

295

time steps to track the oscillations. But, even with small size steps they may alter the dynamics, unless the chosen method has specific advantages. A useful category of them is the group of geometric integrators, which are numerical schemes that preserve some geometric features of the dynamical system. These integrators endow highly qualified simulations with longer time running without spurious effects (like bad energy behavior of conservative systems) than the traditional ones [5–10].

Historically, the exponential integrators were initially proposed long ago by Hersch [11] who constructed the first exponential integrator for linear ordinary differential equations (ODEs) with constant coefficients. Then Certainé [12] introduced the first multi-step integrator of this type by using a variation of parameters approach. Roughly speaking, the development of exponential integrators followed two general directions: (1) those derived to solve first-order ODE systems, and (2) those derived to solve “directly” second-order differential equations (direct integrators for second-order DE, where the term direct means without reducing it to first-order), see [13–16]. Most of these integrators rely on the variation of parameters approach.

The existence of the exponential function in this kind of integrators justifies the term *exponential integrator* firstly used by Hochbruck et al. [15]. Furthermore, a great number of exponential integrators derived for solving second-order differential equations with respect to time (e.g. the Newton’s equations of motion that govern the dynamic equilibrium of elastodynamic systems), are known as Exponential-Time Integrators (ETI) [17]. Regarding applications for solving the category of Hamiltonian systems on which we focus in the present work, the exponential integrators were introduced by Hairer, Lubich, and Wanner [5] and by Leimkuhler and Reich [6]. In our present paper, special emphasis will be given on the derivation of advantageous exponential integrators with respect to geometric characteristics and on the investigation of the stability property and the error analysis of higher-order exponential variational integration.

In the present, we start by recalling the exponential integrators for Hamiltonian systems in Section 2. After a short review of the variational integrators we define a special case of them, namely the exponential variational integrators in Section 3. For those high-order schemes we then investigate their stability region in Section 4 and, finally, their numerical convergence via an error analysis in numerical examples.

2 Exponential Integrators for Hamiltonian Systems

Exponential integrators have been introduced in order to solve numerically Hamiltonian systems of the form [5, 6]

$$\ddot{q} + \Omega q = g(q), \quad g(q) = -\nabla U(q), \quad (1)$$

where Ω is a diagonal matrix which may contain diagonal entries ω with relatively large modulus. $g(q)$ represent the force field created through the negative gradient

from the potential function $U(q)$. We further consider that the potential function $U(q)$ is generally smooth across its domain. Introducing time step h , when focusing on the long time behavior of the numerical solutions of the above systems, many authors present solutions for large values of ωh [5].

For such systems it has been shown that an exact discretization of (1) satisfies the equation [5]

$$q_{n+1} - 2 \cos(h\omega)q_n + q_{n-1} = 0. \tag{2}$$

Thus, by combining those we may write

$$q_{n+1} - 2 \cos(h\omega)q_n + q_{n-1} = h^2 \psi(\omega h) g(\phi(\omega h)q_n), \tag{3}$$

where the functions $\psi(\omega h)$ and $\phi(\omega h)$ are even, real-valued functions satisfying the conditions $\psi(0) = \phi(0) = 1$. For appropriately defined functions ψ and ϕ , the latter equations constitute exponential integrators [14, 16].

3 High-Order Exponential Variational Integrators

In giving a brief overview over the variational integrators, we first consider the simple case, where the discrete configurations and velocities are defined at the nodes of a time grid only [4]. Using the notation Q for the configuration manifold, the derivation of variational integrators uses the discrete Lagrangian map $L_d : Q \times Q \rightarrow \mathbb{R}$, which may be considered as an approximation of a continuous action with Lagrangian $L : TQ \rightarrow \mathbb{R}$, i.e.

$$L_d(q_k, q_{k+1}) \approx \int_{t_k}^{t_{k+1}} L(q, \dot{q}) dt, \tag{4}$$

in the time interval $[t_k, t_{k+1}] \subset \mathbb{R}$. Recalling that, the action sum $S_d : Q^{N+1} \rightarrow \mathbb{R}$ comes out of the Lagrangian L_d , S_d is defined by

$$S_d(\gamma_d) = \sum_{k=0}^{N-1} L_d(q_k, q_{k+1}), \tag{5}$$

with $\gamma_d = (q_0, \dots, q_N)$ representing the discrete trajectory. Next, the discrete Hamilton principle states that a motion γ_d for the discrete mechanical system extremizes the action sum, i.e. $\delta S_d = 0$. Afterwards, making the differentiation and rearrangement of the terms in the latter equation and having in mind that both q_0 and q_N are fixed, the discrete Euler–Lagrange equations are obtained [4] as

$$D_2 L_d(q_{k-1}, q_k) + D_1 L_d(q_k, q_{k+1}) = 0, \quad k = 1, \dots, N - 1. \tag{6}$$

The notation $D_i L_d$ indicates the slot derivative with respect to the argument of L_d .

Those derivatives are used for the definition of the discrete conjugate momentum at time steps k and $k + 1$ via the Legendre transforms, i.e.

$$p_k = -D_1 L_d(q_k, q_{k+1}), \quad p_{k+1} = D_2 L_d(q_k, q_{k+1}), \quad k = 0, \dots, N - 1. \quad (7)$$

The latter equations, also known as position-momentum form of a variational integrator, can be used when an initial condition (q_0, p_0) is known, to obtain (q_1, p_1) [4, 8–10].

We can now extend the definitions of the previous section by introducing an arbitrary number of intermediate points. Those high-order variational integrator schemes will then be introduced for the Hamiltonian systems of (1) within the context of the exponential integrators of (3).

To construct high-order methods, we approximate the action integral along the curve segment between q_k and q_{k+1} using a discrete Lagrangian that depends only on the end points. This way, we obtain expressions for configurations q_k^j and velocities \dot{q}_k^j for $j = 0, \dots, S - 1$, $S \in \mathbb{N}$ at time $t_k^j \in [t_k, t_{k+1}]$ by expressing $t_k^j = t_k + C_k^j h$ for $C_k^j \in [0, 1]$ such that $C_k^0 = 0$, $C_k^{S-1} = 1$ using

$$q_k^j = g_1(t_k^j)q_k + g_2(t_k^j)q_{k+1}, \quad \dot{q}_k^j = \dot{g}_1(t_k^j)q_k + \dot{g}_2(t_k^j)q_{k+1}, \quad (8)$$

where $h \in \mathbb{R}$ is the time step. For the purposes of our present work, we choose functions of the form

$$g_1(t_k^j) = \sin\left(u - \frac{t_k^j - t_k}{h}u\right) (\sin u)^{-1}, \quad g_2(t_k^j) = \sin\left(\frac{t_k^j - t_k}{h}u\right) (\sin u)^{-1} \quad (9)$$

to represent the oscillatory behavior of the solution [8–10, 18–20]. Then, for the sake of continuity, the conditions

$$g_1(t_{k+1}) = g_2(t_k) = 0 \quad (10)$$

and

$$g_1(t_k) = g_2(t_{k+1}) = 1 \quad (11)$$

must be fulfilled. It should be noted that other interpolations (e.g. linear, cubic splines, etc.) are possible as alternatives to (9), see [9, 10].

Moreover, for any choice of interpolation we define the discrete Lagrangian in the form of the weighted sum

$$L_d(q_k, q_{k+1}) = h \sum_{j=0}^{S-1} w^j L\left(q(t_k^j), \dot{q}(t_k^j)\right), \quad (12)$$

where it can be easily proved that

$$\sum_{j=0}^{S-1} w^j \left(C_k^j\right)^m = \frac{1}{m+1}, \tag{13}$$

where $m = 0, 1, \dots, S - 1$ and $k = 0, 1, \dots, N - 1$ must hold, see e.g. [8, 18].

In applying the above interpolation technique with the trigonometric expressions (9), the parameter u can be chosen as $u = \omega h$. Furthermore, for problems that involve a constant and known domain frequency ω , the parameter u can be easily computed, while for the solution of orbital problems of the general N -body problem, a new parameter u may be computed by estimating the frequency of the motion of any moving point mass during the course of motion [9, 10, 18].

Moreover, for the derivation of exponential variational integrators, one may apply the steps of deducing high-order variational integrators to Hamiltonian system (1). Then, the discrete Euler–Lagrange equations (6) lead to the expressions

$$q_{n+1} + \Lambda(u, \omega, h, S)q_n + q_{n-1} = h^2\Psi(\omega h)g(\Phi(\omega h)q_n), \tag{14}$$

where

$$\Lambda(u, \omega, h, S) = \frac{\sum_{j=0}^{S-1} w^j \left[\dot{g}_1(t_k^j)^2 + \dot{g}_2(t_k^j)^2 - \omega^2(g_1(t_k^j)^2 + g_2(t_k^j)^2) \right]}{\sum_{j=0}^{S-1} w^j \left[\dot{g}_1(t_k^j)\dot{g}_2(t_k^j) - \omega^2 g_1(t_k^j)g_2(t_k^j) \right]}. \tag{15}$$

In addition, based on the latter two expressions, one may derive exponential variational integrators that use the configurations q_k^j and velocities \dot{q}_k^j of (8). In this case we get

$$\Lambda(u, \omega, h, S) = -2 \cos(\omega h). \tag{16}$$

We note that, whenever the latter equation holds, the defined above high-order variational integrators are of exponential type [8–10, 19–21].

4 Stability Analysis of Exponential Variational Integrators

The investigation of the stability properties of exponential variational integrators is important. Here, following [6, 7, 22], we restrict ourselves to a linear stability analysis, and toward this aim, we start by considering the simple case of a Hamiltonian system (1), i.e. the harmonic oscillator

$$\ddot{q} + \omega^2 q = 0. \tag{17}$$

This system is described by the smooth Lagrangian function

$$L(q, \dot{q}) = \frac{1}{2} \dot{q}^2 - \frac{1}{2} \omega^2 q^2, \quad \omega \in \mathbb{R}. \tag{18}$$

If we write (17) using the equivalent Hamilton’s equations, we obtain

$$\dot{q} = p, \quad \dot{p} = -\omega^2 q, \tag{19}$$

where the exact solution can be written as [7]

$$\begin{pmatrix} p(t) \\ q(t) \end{pmatrix} = \begin{pmatrix} \cos(\omega t) & -\omega \sin(\omega t) \\ \frac{\sin(\omega t)}{\omega} & \cos(\omega t) \end{pmatrix} \begin{pmatrix} p(0) \\ q(0) \end{pmatrix} = M_\omega \begin{pmatrix} p(0) \\ q(0) \end{pmatrix}. \tag{20}$$

But because $\det(M_\omega) = 1$, the eigenvalues of M_ω are $\lambda_{1,2} = e^{\pm i\omega t}$ and thus $|\lambda_{1,2}| = 1$.

Next, recalling that a numerical solution is asymptotically stable when the growth of the solution is asymptotically bounded, a sufficient condition for asymptotic stability implies that the eigenvalues of $M_{h\omega}$ must be on the unit disk of the complex plane. In addition, they are simple if they lie on the unit circle. In the next subsections, we investigate the latter property for selected variational integrators.

At this point it is worth noting that, for the Lagrangian of the harmonic oscillator with frequency ω , the discrete Lagrangian (12) with interpolation functions defined via (9) reads

$$L_d(q_k, q_{k+1}) = \frac{h}{2} \left[\sum_{j=0}^{S-1} w^j \left(\dot{g}_1(t_k^j) q_k + \dot{g}_2(t_k^j) q_{k+1} \right)^2 - \omega^2 \sum_{j=0}^{S-1} w^j \left(g_1(t_k^j) q_k + g_2(t_k^j) q_{k+1} \right)^2 \right]. \tag{21}$$

For the latter discrete Lagrangian, the discrete Euler–Lagrange equations (6) yield

$$q_{k+1} + \frac{\sum_{j=0}^{S-1} w^j \left[\dot{g}_1(t_k^j)^2 + \dot{g}_2(t_k^j)^2 - \omega^2 (g_1(t_k^j)^2 + g_2(t_k^j)^2) \right]}{\sum_{j=0}^{S-1} w^j \left[\dot{g}_1(t_k^j) \dot{g}_2(t_k^j) - \omega^2 g_1(t_k^j) g_2(t_k^j) \right]} q_k + q_{k-1} = 0. \tag{22}$$

This expression is explicit for any choice of interpolation functions [23, 24].

4.1 Exponential Variational Integrator for $S = 2$

In the special case when no intermediate points are used, i.e. $S = 2$, $t_k^0 = t_k$, and $t_k^1 = t_{k+1}$, the coefficients C_k^j and w^j described in Section 3 take the values

$$\begin{aligned} C_k^0 &= 0, & C_k^1 &= 1 \\ w^0 &= \frac{1}{2}, & w^1 &= \frac{1}{2}. \end{aligned} \tag{23}$$

Relying on the concrete expressions (6) and (7) and using (8)–(22) and (23), the eigenvalues $\lambda_{1,2}$ of the matrix $M_{h,\omega}$ can be written as

$$\lambda_{1,2} = \frac{2 \cos(2\omega h) + 2 \pm \sqrt{2 \cos(4\omega h) - 2}}{4 \cos(\omega h)}. \tag{24}$$

Since

$$2 \cos(4\omega h) - 2 \leq 0, \quad \forall \omega h \in \mathbb{Z}, \tag{25}$$

in order to show the stability of the resulting numerical scheme, we prove the following theorem.

Theorem 4.1 *The phase fitted variational integrator using trigonometric interpolation for $S = 2$ is stable for $\omega h \neq \nu\pi + \frac{\pi}{2}, \nu \in \mathbb{Z}$.*

Proof

First Case $\omega h = \nu\pi + \frac{\pi}{2}, \nu \in \mathbb{Z}$

For these values of ωh , the denominator of the eigenvalues in (24) vanishes, creating an unstable method, see Figure 1.

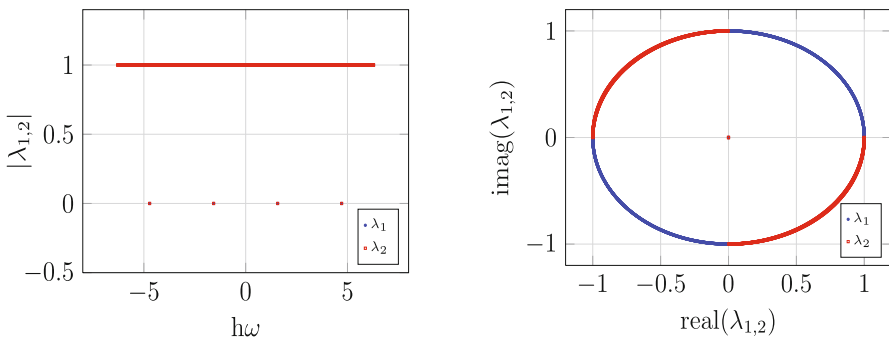


Fig. 1 Modulus of the eigenvalues (24) of $M_{h,\omega}$ for the phase fitted variational integrators for $S = 2$ for $\omega h \in [-2\pi, 2\pi]$ and real and imaginary part

Second Case $\omega h \neq \nu\pi + \frac{\pi}{2}, \nu \in \mathbb{Z}$

For the case when

$$2 \cos(4\omega h) - 2 = 0, \quad (26)$$

which is equivalent to

$$\omega h = \nu\pi \quad (\text{since } \omega h \neq \nu\pi + \frac{\pi}{2}, \nu \in \mathbb{Z}), \quad (27)$$

the eigenvalues (24) are

$$\lambda_{1,2} = \frac{2 \cos(2\nu\pi) + 2}{4 \cos(\nu\pi)} = 1. \quad (28)$$

This means that these choices of ωh create a stable integrator. Furthermore, in the case when

$$2 \cos(4\omega h) - 2 < 0, \quad (29)$$

which is equivalent to

$$\omega h \neq \nu\frac{\pi}{2}, \nu \in \mathbb{Z}, \quad (30)$$

both eigenvalues in (24) are complex numbers. The modulus of the eigenvalues is $|\lambda_{1,2}|^2 = 1$ since

$$\begin{aligned} |\lambda_{1,2}|^2 &= \left(\frac{2 \cos(2\omega h) + 2}{4 \cos(\omega h)} \right)^2 + \left(\frac{\sqrt{2 \cos(4\omega h) - 2}}{4 \cos(\omega h)} \right)^2 \\ &= \left(\frac{2 \cos(2\omega h) + 2}{4 \cos(\omega h)} \right)^2 + \frac{|2 \cos(4\omega h) - 2|}{(4 \cos(\omega h))^2} \\ &= \frac{4 \cos^2(2\omega h) + 4 + 8 \cos(2\omega h) + 2 - 2 \cos(4\omega h)}{(4 \cos(\omega h))^2} \end{aligned}$$

$$\text{using } \cos(4\omega h) = 2 \cos^2(2\omega h) - 1$$

$$= \frac{8 \cos(2\omega h) + 8}{(4 \cos(\omega h))^2}$$

$$\text{using } \cos(2\omega h) = 2 \cos^2(\omega h) - 1$$

$$= 1. \quad (31)$$

□

For an illustration, the stability region of the exponential variational integrator (22) for $S = 2$ is shown in Figure 1. It is clear from this figure that for $\omega h \neq \nu\pi + \frac{\pi}{2}, \nu \in \mathbb{Z}$ the integrator is stable. The unstable choices of ωh are illustrated in Figure 1.

4.2 Exponential Variational Integrator for $S = 3$

For the specific case when one intermediate point in each time interval is used, i.e. $S = 3$ and $t_k^0 = t_k, t_k^1 = t_k + \frac{h}{2},$ and $t_k^2 = t_{k+1},$ the coefficients C_k^j and w^j described in Section 3 are

$$\begin{aligned} C_k^0 &= 0, & C_k^1 &= \frac{1}{2}, & C_k^2 &= 1, \\ w^0 &= \frac{1}{6}, & w^1 &= \frac{1}{6}, & w^2 &= \frac{1}{6}. \end{aligned} \tag{32}$$

Again, by using the discrete Euler–Lagrange equations (6) for the discrete Lagrangian (21), the eigenvalues of the matrix $M_{h,\omega}$ can be cast in the form

$$\lambda_{1,2} = 2 + \frac{\cos^2(\omega h) - 4}{2 \cos^2\left(\frac{\omega h}{2}\right) + 1} \pm \frac{1}{2} \sqrt{\frac{1}{32 \cos^2\left(\frac{\omega h}{2}\right) + 4 \cos^2(\omega h)} \Delta_3(\omega, h)}. \tag{33}$$

Then, the function $\Delta_3(\omega, h)$ is given by the expression

$$\Delta_3(\omega, h) = 4 \cos^2(2\omega h) - 32 \cos^2\left(\frac{\omega h}{2}\right) + 32 \cos^2\left(\frac{3\omega h}{2}\right) + 64 \cos^2(\omega h) - 68. \tag{34}$$

It is worth noting that both eigenvalues $\lambda_{1,2}$ of (33) are complex numbers for $\omega h \neq 6\nu\pi, \nu \in \mathbb{Z}$ since $\Delta_3(\omega, h) < 0$ (the period of $\Delta_3(\omega, h)$ is 6π). Furthermore, for the validity of the stability for trigonometric interpolation with $S = 3,$ we prove the following theorem.

Theorem 4.2 *The phase fitted variational integrator, that uses trigonometric interpolation for $S = 3,$ is stable for any $\omega h \in \mathbb{R}.$*

Proof

First Case $\omega h = 6\nu\pi, \nu \in \mathbb{Z}$

For the special case when $\omega h = 6\nu\pi$ we have $\Delta_3(\omega, h) = 0$ and, the eigenvalues in (33) lie on the unit circle, because $\lambda_{1,2} = 1.$

Second Case $\omega h \neq 6\nu\pi, \nu \in \mathbb{Z}$

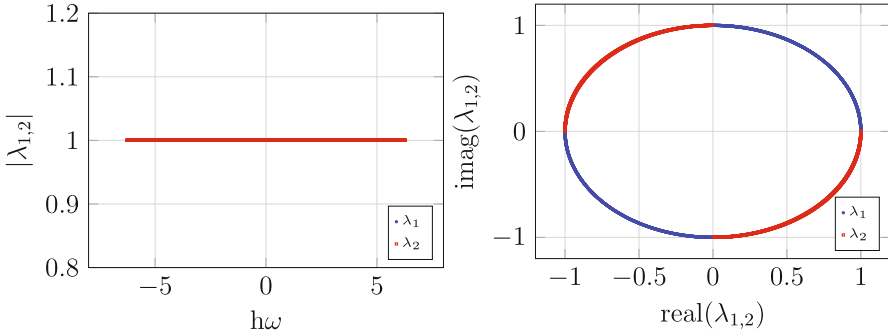


Fig. 2 Modulus of the eigenvalues (33) of $M_{h,\omega}$ for the phase fitted variational integrators for $S = 3$ for $\omega h \in [-2\pi, 2\pi]$ and real and imaginary part

This choice of ωh leads to complex numbers for both eigenvalues of (33) which now have magnitude equal to unity as $|\lambda_{1,2}|^2 = 1$, (the proof follows the ONE of Theorem 4.1). □

The degree of stability (stability region) of the phase fitted variational integrator (22) with trigonometric interpolation, for $S = 3$, is shown in Figure 2. One can conclude that, for $\omega h \in \mathbb{R}$, the integrator is unconditionally stable.

5 Error Analysis from Testing the Variational Integrators in Hamiltonian Systems

The numerical convergence of the proposed variational integrators, may be illustrated by considering the harmonic oscillator with frequency $\omega = 1$ which is described by the Lagrangian (18). For concrete numerical tests, following [9, 10, 25], we choose as initial conditions $(q_0, p_0) = (2, 1)$ and as time interval $[0, 25]$. In Figure 3a, b the evolution of the errors in the position q and the momentum p are plotted for the Störmer–Verlet [5] and the trigonometric interpolation method with $S = 3$ ($u = \omega h$). From the comparison, one can imply that the errors using the trigonometric interpolation are much smaller (also bounded for all the integration time) than those obtained by using Störmer–Verlet method of [5, 18, 26–28].

As an additional test, the global errors for the position and momentum components at $t = 3$ and time steps $h \in \{0.05, 0.1, 0.5\}$ are compared in Figure 4 with the Störmer–Verlet method [5, 18, 27, 28] for the system of harmonic oscillator. Evidently, while both methods are of the same order, for all the step sizes tested, smaller errors in position and momentum result when trigonometric interpolation for $u = \omega h$ is employed.

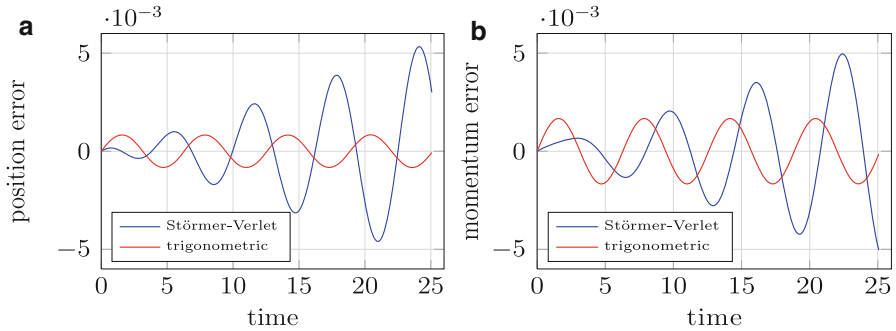


Fig. 3 Harmonic oscillator with $\omega = 1$, using $h = 0.05$ and $S = 3$. (a) Errors in position and (b) errors in momentum for the Störmer-Verlet [5, 18, 27, 28] and the trigonometric interpolation method with $u = \omega h$

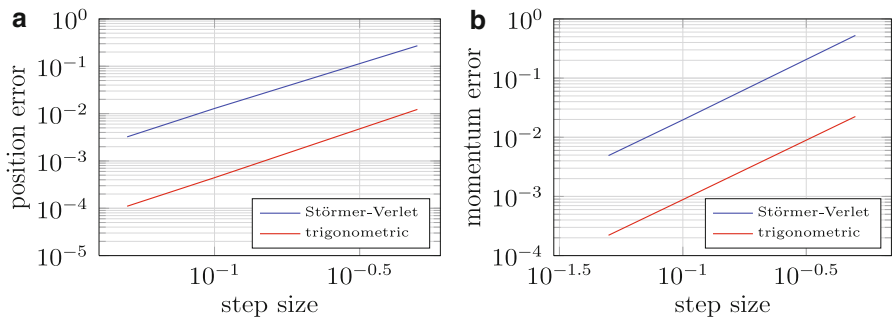


Fig. 4 Harmonic oscillator with $\omega = 1$ and $S = 3$. Global errors of (a): the position and (b): the momentum using three step sizes h for the Störmer-Verlet [5] and the trigonometric interpolation method with $u = \omega h$

6 Conclusions

In this article, we deal with a special kind of high-order numerical method which relies on the exponential variational integrators. It is mostly appropriate for Hamiltonian systems and for this reason its efficiency is tested in oscillatory problems. We then explore their linear stability properties of this exponential integrator via evaluating the eigenvalues of the amplification matrix. Such a consideration provided us with high confidence level regarding the applicability of these methods and pointed out that they are stable for a wide range of parameters. Finally, we investigated the numerical convergence of the exponential integrators via an error analysis in prominent numerical examples. The conclusion extracted from the latter application in studying the evolution of the errors in position q and momentum p through the proposed simulation technique shows a very good behavior.

Acknowledgement Dr. Odysseas Kosmas wishes to acknowledge the support of EPSRC via grand EP/N026136/1 “Geometric Mechanics of Solids.”

References

1. B. Engquist, A. Fokas, E. Hairer, A. Iserles, *Highly Oscillatory Problems* (Cambridge University Press, Cambridge, 2009)
2. J. Wendlandt, J.E. Marsden, Mechanical integrators derived from a discrete variational principle. *Phys. D*, **106**, 223–246 (1997)
3. C. Kane, J.E. Marsden, M. Ortiz, Symplectic-energy-momentum preserving variational integrators. *J. Math. Phys.* **40**, 3353–3371 (2001)
4. J.E. Marsden, M. West, Discrete mechanics and variational integrators. *Acta Numer.* **10**, 357–514 (2001)
5. E. Hairer, C. Lubich, G. Wanner, Geometric numerical integration illustrated by the Störmer-Verlet method. *Acta Numer.* **12**, 399–450 (2003)
6. B. Leimkuhler, S. Reich, *Simulating Hamiltonian Dynamics* (Cambridge Monographs on Applied and Computational Mathematics, Cambridge, 2004)
7. S. Ober-Blöbaum, Galerkin variational integrators and modified symplectic Runge–Kutta methods. *IMA J. Numer. Anal.* **37**, 375–406 (2017)
8. O.T. Kosmas, D.S. Vlachos, Phase-fitted discrete Lagrangian integrators. *Comput. Phys. Commun.* **181**, 562–568 (2010)
9. O.T. Kosmas, D.S. Leyendecker, Analysis of higher order phase fitted variational integrators. *Adv. Comput. Math.* **42**, 605–619 (2016)
10. O.T. Kosmas, S. Leyendecker, Variational integrators for orbital problems using frequency estimation. *Adv. Comput. Math.* **45**, 1–21 (2019). <https://doi.org/10.1007/s10444-018-9603-y>
11. J. Hersch, Contribution a la methode des equations aux differences. *Z. Angew. Math. Phys.* **9**, 129–180 (1958)
12. J. Certaine, *The Solution of Ordinary Differential Equations with Large Time Constants*. Mathematical Methods for Digital Computers (Wiley, New York, 1960), pp. 128–132
13. D.A. Pope, An exponential method of numerical integration of ordinary differential equations. *Commun. ACM* **6**, 491–493 (1963)
14. P. Deuffhard, A study of extrapolation methods based on multistep schemes without parasitic solutions. *Z. Angew. Math. Phys.* **30**, 177–189 (1979)
15. M. Hochbruck, C. Lubich, H. Selfhofer, Exponential integrators for large systems of differential equations. *SIAM J. Sci. Comput.* **19**, 1552–1574 (1998)
16. B. García-Archilla, M.J. Sanz-Serna, R.D. Skeel, Long-time-step methods for oscillatory differential equations. *SIAM J. Sci. Comput.* **20**, 930–963 (1999)
17. A. Nealen, M. Mueller, R. Keiser, E. Boxerman, M. Carlson, Physically based deformable models in computer graphics. *Comput. Graph. Forum* **25**, 809–836 (2006)
18. O.T. Kosmas, S. Leyendecker, Phase lag analysis of variational integrators using interpolation techniques. *PAMM Proc. Appl. Math. Mech.* **12**, 677–678 (2012)
19. O.T. Kosmas, D. Papadopoulos, Multisymplectic structure of numerical methods derived using nonstandard finite difference schemes. *J. Phys. Conf. Ser.* **490**, 012205 (2014)
20. O.T. Kosmas, D. Papadopoulos, D. Vlachos, Geometric derivation and analysis of multi-symplectic numerical schemes for differential equations, in *Computational Mathematics and Variational Analysis. Springer Optimization and Its Applications*, vol. 159, ed. by N. Daras, T. Rassias (2020), pp. 207–226
21. O.T. Kosmas, Exponential variational integrators for the dynamics of multibody systems with holonomic constraints. *J. Phys. Conf. Ser.* **1391**, 012170 (2019)
22. O.T. Kosmas, S. Leyendecker, Stability analysis of high order phase fitted variational integrators, in *Proceedings of WCCM XI: ECCM V—ECFD VI*, vol. 1389 (2014), pp. 865–866

23. A. Stern, E. Grinspun, Implicit-explicit integration of highly oscillatory problems. *SIAM Multiscale Model. Simul.* **7**, 1779–1794 (2009)
24. S. Reich, Backward error analysis for numerical integrators. *SIAM J. Numer. Anal.* **36**, 1549–1570 (1999)
25. M. Leok, J. Zhang, Discrete Hamiltonian variational integrators. *IMA J. Numer. Anal.* **31**, 1497–1532 (2011)
26. O.T. Kosmas, D.S. Vlachos, A space-time geodesic approach for phase fitted variational integrators. *J. Phys. Conf. Ser.* **738**, 012133 (2016)
27. O.T. Kosmas, Charged particle in an electromagnetic field using variational integrators. *ICNAAM Numer. Anal. Appl. Math.* **1389**, 1927 (2011)
28. O.T. Kosmas, D.S. Vlachos, Local path fitting: a new approach to variational integrators. *J. Comput. Appl. Math.* **236**, 2632–2642 (2012)

A Degenerate Kirchhoff-Type Inclusion Problem with Nonlocal Operator



Dumitru Motreanu

Abstract The chapter focuses on a Kirchhoff-type elliptic inclusion problem driven by a generalized nonlocal fractional p -Laplacian whose nonlocal term vanishes at finitely many points and for which the multivalued term is in the form of the generalized gradient of a locally Lipschitz function. The corresponding elliptic equation has been treated in (Liu et al., Existence of solutions to Kirchhoff-type problem with vanishing nonlocal term and fractional p -Laplacian). Multiple nontrivial solutions are obtained by applying the nonsmooth critical point theory combined with truncation techniques.

1 Introduction

Let $\Omega \subset \mathbb{R}^N$, with $N \geq 1$, be a bounded domain with Lipschitz boundary $\partial\Omega$ and fix constants $s \in (0, 1)$ and $p \in (1, +\infty)$ with $ps < N$. We denote by $|\cdot|$ the Euclidean norm in \mathbb{R}^N and by $B_\rho(x)$ the open ball in \mathbb{R}^N centered at $x \in \mathbb{R}^N$ and of radius $\rho > 0$.

In the present chapter we study the nonlocal differential inclusion

$$\begin{cases} m \left(\int_{\mathbb{R}^{2N}} |u(x) - u(y)|^p K(x - y) dx dy \right) \mathfrak{L}_K^p u \in [\underline{f}(u), \overline{f}(u)] & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \quad (1)$$

Here we have a continuous function $m : [0, +\infty) \rightarrow \mathbb{R}$ that can vanish, a nonlocal fractional p -Laplace-type operator \mathfrak{L}_K^p given by

D. Motreanu (✉)

Département de Mathématiques, Université de Perpignan, Perpignan, France
e-mail: motreanu@univ-perp.fr

$$\mathfrak{L}_K^p u(x) := 2 \lim_{\varepsilon \downarrow 0} \int_{\mathbb{R}^N \setminus B_\varepsilon(x)} |u(x) - u(y)|^{p-2} (u(x) - u(y)) K(x - y) dy$$

for a.e. $x \in \mathbb{R}^N$ and corresponding to a function $f \in L^\infty_{\text{loc}}(\mathbb{R})$ we have set

$$\underline{f}(s) = \lim_{\delta \rightarrow 0} \text{essinf}_{|\tau-s| < \delta} f(\tau), \quad \forall s \in \mathbb{R} \tag{2}$$

and

$$\overline{f}(s) = \lim_{\delta \rightarrow 0} \text{esssup}_{|\tau-s| < \delta} f(\tau), \quad \forall s \in \mathbb{R}. \tag{3}$$

We emphasize the degenerate character of (1) which comes from the possibility of the function m to vanish. It is also worth mentioning that it is essential to take $f \in L^\infty_{\text{loc}}(\mathbb{R})$ in order that (2) and (3) be well defined.

We assume that K verifies the conditions:

(K) $K : \mathbb{R}^N \setminus \{0\} \rightarrow (0, +\infty)$ is a measurable function satisfying

- (i) the function $x \mapsto \min \{1, |x|^p\} K(x)$ belongs to $L^1(\mathbb{R}^N)$;
- (ii) there exists a constant $\alpha > 0$ such that

$$K(x) \geq \alpha |x|^{-(N+ps)}, \quad \forall x \in \mathbb{R}^N \setminus \{0\}.$$

An important particular case of the singular kernel K is $K(x) = |x|^{-(N+ps)}$, for which problem (1) reduces to

$$\begin{cases} m \left(\int_{\mathbb{R}^{2N}} \frac{|u(x) - u(y)|^p}{|x - y|^{N+ps}} dx dy \right) (-\Delta)_s^p u \in [\underline{f}(u), \overline{f}(u)] & \text{in } \Omega \\ u = 0 & \text{in } \mathbb{R}^N \setminus \Omega, \end{cases} \tag{4}$$

where $(-\Delta)_s^p$ stands for the fractional p -Laplacian

$$(-\Delta)_s^p u(x) := 2 \lim_{\varepsilon \downarrow 0} \int_{\mathbb{R}^N \setminus B_\varepsilon(x)} \frac{|u(x) - u(y)|^{p-2} (u(x) - u(y))}{|x - y|^{N+ps}} dy$$

for a.e. $x \in \mathbb{R}^N$. In the limit as $s \rightarrow 1^-$ we recover from $(-\Delta)_s^p$ the (negative) p -Laplacian operator $-\Delta_p : W_0^{1,p}(\Omega) \rightarrow W^{-1,p'}(\Omega) = W_0^{1,p}(\Omega)^* (\frac{1}{p} + \frac{1}{p'} = 1)$, which is given by

$$\langle -\Delta_p u, v \rangle = \int_\Omega |\nabla u|^{p-2} \nabla u \cdot \nabla v dx \quad \text{for all } u, v \in W_0^{1,p}(\Omega).$$

If the function f is continuous, then the interval $[f(u(x)), \overline{f}(u(x))]$ collapses to the singleton $f(u(x))$. Consequently, in this case (4) reduces to the quasilinear Dirichlet equation

$$\begin{cases} m \left(\int_{\mathbb{R}^{2N}} \frac{|u(x) - u(y)|^p}{|x - y|^{N+ps}} dx dy \right) (-\Delta)_s^p u = f(u) & \text{in } \Omega \\ u = 0 & \text{in } \mathbb{R}^N \setminus \Omega. \end{cases} \tag{5}$$

Equation (5) has been examined in [11]. When $p = 2$ and $s \rightarrow 1^-$, problem (5) turns to be the degenerate elliptic equation of Kirchhoff-type

$$\begin{cases} -m(\|u\|_{H_0^1(\Omega)}) \Delta u = f(u) & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{6}$$

which has been studied by Santos Júnior and Siciliano [22] (see also [8]). The multivalued problem (1) driven by a nonlocal fractional operator was not considered before in the literature, not even for the case $K(x) = |x|^{-(N+ps)}$ in (4). It represents a generalization of all the previous works.

The origin of this type of problems lies in the work of Kirchhoff [10]. Many real life phenomena are governed by what we nowadays call Kirchhoff-type equations. Stationary and non-stationary Kirchhoff-type problems are extensively studied for their rich mathematical insight, see, e.g., [1, 7–9, 13, 20–22]. On the other hand, much interest is paid in recent years to problems involving fractional and nonlocal operators, see, e.g., [6, 12, 14–17, 25]. A further development is represented by dealing with Kirchhoff-type problems driven by nonlocal fractional operators. It is motivated by accurate models in population dynamics and anomalous diffusion processes. We briefly review a few works in this direction. Xiang et al. [24] examined an equation with fractional p -Laplacian and a nonvanishing Kirchhoff function. Autuori et al. [2] studied a stationary Kirchhoff problem with critical nonlinearity and Kirchhoff function vanishing at zero. Pan et al. [19] established the existence of a global solution to a degenerate diffusion problem of Kirchhoff-type. Xiang et al. [23] pointed out the blow-up for a nonlocal diffusion equation with Kirchhoff function that vanishes at zero. Liu et al. [11] investigated the nonlocal equation (5) with multiple points at which the Kirchhoff function m vanishes.

Here we show the existence of multiple solutions to the nonlocal integro-differential inclusion (1) that are located by using the zeros of the Kirchhoff function m . It is for the first time when such a property is established for an inclusion problem. In the particular case of problem (5) the result becomes the theorem given in [11], so the present work is a generalization of [11].

Our approach consists in building a nonsmooth version of [11]. Specifically, we rely on the nonsmooth version of mountain-pass theorem stated in [4] (see also [18, Chapter 3]) instead of the classical mountain-pass theorem. In this respect, we introduce intermediate problems through suitable truncations and the zeros of the

Kirchhoff function m . For each intermediate problem we set forth a variational framework of mountain-pass type and prove the existence of solutions to the intermediate problems. Then we observe that they are solutions to the original problem (1), thus achieving the desired conclusion. Essential differences occur with respect to the treatment for the smooth case in [11]. Basically this is caused by the fact that here we have to work with the pair of functions (f, \bar{f}) introduced in (2)–(3) in place of the function f . Furthermore, the nonsmooth frame has to match the functional setting of fractional p -Laplacian-type operators. We illustrate the applicability of our main result with an example.

The source of inspiration for us was Santos Júnior and Siciliano [22] that dealt with problem (6). Nevertheless, we have substantially modified the arguments in [22]. These modifications were necessary because some arguments in [22] were not accurate. For instance, the fact used in the proof of Proposition 3.1 of [22] consisting in deriving from the equality $f_*(u_*) = 0$ that $u_* = 0$ is not true (note that if $u_* = s_*$ on a set of positive measure, then therein $f_*(s_*) = 0$ with $s_* > 0$ by assumption (f)). Our reasoning is considerably changed with respect to [22]. Even the hypotheses for the nonlinearity $f(u)$ are strongly changed. By the way, assumption (f) employed in [22] is simply dropped.

The rest of the paper is organized as follows. Section 2 contains basic elements related to the mathematical background. Section 3 presents our hypotheses and main result. Section 4 studies the associated truncated problems. Section 5 describes the mountain-pass geometry in connection with the truncated problems. Section 6 discusses the Palais–Smale condition for the Euler functionals corresponding to the intermediate problems. Section 7 provides the proof of our main result and an example.

2 Mathematical Background

First we provide some needed prerequisites related to the nonlocal functional setting. More details can be found in [6, 16, 17].

Let $\Omega \subset \mathbb{R}^N$ be a bounded domain with Lipschitz boundary $\partial\Omega$. The notation $|\Omega|$ will designate the Lebesgue measure of Ω . Denote $\mathcal{O} = (\mathbb{R}^N \setminus \Omega) \times (\mathbb{R}^N \setminus \Omega) \subset \mathbb{R}^{2N}$ and $Q = \mathbb{R}^{2N} \setminus \mathcal{O}$ and fix constants $s \in (0, 1)$ and $p \in (1, +\infty)$. The fractional critical exponent p_s^* is defined by

$$p_s^* = \begin{cases} \frac{Np}{N-sp} & \text{if } sp < N \\ +\infty & \text{if } sp \geq N. \end{cases}$$

Let $K : \mathbb{R}^N \setminus \{0\} \rightarrow (0, +\infty)$ be a singular kernel functional satisfying hypotheses (K). We note that

$$X = \left\{ u : \mathbb{R}^N \rightarrow \mathbb{R} \text{ measurable} : u|_{\Omega} \in L^p(\Omega), \int_Q |u(x) - u(y)|^p K(x - y) \, dx dy < \infty \right\}$$

is a Banach space equipped with the norm

$$\|u\|_X = \|u\|_{L^p(\Omega)} + \left(\int_Q |u(x) - u(y)|^p K(x - y) \, dx dy \right)^{\frac{1}{p}}.$$

The natural space for problem (1) is the closed linear subspace of X given by

$$X_0 := \left\{ u \in X : u(x) = 0 \text{ for a.e. } x \in \mathbb{R}^N \setminus \Omega \right\},$$

which is continuously embedded in $L^r(\Omega)$ for all $r \in [1, p_s^*]$. Actually, taking also into account assumption (K) , there exists a constant $c_0(r) > 0$ such that

$$\|u\|_{L^r(\Omega)}^p \leq c_0(r) \int_{\Omega} \int_{\Omega} \frac{|u(x) - u(y)|^p}{|x - y|^{N+sp}} \, dx dy \leq \frac{c_0(r)}{\alpha} \int_Q |u(x) - u(y)|^p K(x - y) \, dx dy$$

whenever $u \in X_0$. From the preceding inequality with $r = p$, it holds

$$\begin{aligned} & \left(\int_Q |u(x) - u(y)|^p K(x - y) \, dx dy \right)^{\frac{1}{p}} \leq \|u\|_X \\ & = \|u\|_{L^p(\Omega)} + \left(\int_Q |u(x) - u(y)|^p K(x - y) \, dx dy \right)^{\frac{1}{p}} \\ & \leq \left(\left(\frac{c_0(p)}{\alpha} \right)^{\frac{1}{p}} + 1 \right) \left(\int_Q |u(x) - u(y)|^p K(x - y) \, dx dy \right)^{\frac{1}{p}} \end{aligned}$$

for all $u \in X_0$. It follows that

$$\|u\|_{X_0} := \left(\int_Q |u(x) - u(y)|^p K(x - y) \, dx dy \right)^{\frac{1}{p}}$$

is an equivalent norm on X_0 . In the sequel, X_0 will be endowed with the norm $\|u\|_{X_0}$ becoming a reflexive Banach space. Hereafter the notation $\langle \cdot, \cdot \rangle$ stands for the duality brackets for the dual pair (X_0^*, X_0) . For any $q \in [1, p_s^*]$ we denote by $S(q) > 0$ the smallest positive constant such that

$$\|u\|_q := \|u\|_{L^q(\Omega)} \leq S(q) \|u\|_{X_0} \text{ for all } u \in X_0.$$

The embedding of X_0 into $L^q(\Omega)$ is compact when $q \in [1, p_s^*)$. We mention that for any $u \in X_0$ we have

$$u^\pm := \max\{\pm u, 0\} \in X_0, \quad u = u^+ - u^-, \quad |u| = u^+ + u^-.$$

For a later use, we denote by r' the Hölder conjugate of any $r \in (1, +\infty)$, i.e., $r' = \frac{r}{r-1}$.

We also recall a few things about the nonsmooth critical point theory for locally Lipschitz functions for which we refer to [4, 18]. This theory is based on the notion of generalized gradient (see [5]). The generalized directional derivative of a locally Lipschitz function $\Phi : X \rightarrow \mathbb{R}$ on a Banach space X at $u \in X$ in the direction $v \in X$ is defined as

$$\Phi^0(u; v) := \limsup_{w \rightarrow u, t \rightarrow 0^+} \frac{1}{t}(\Phi(w + tv) - \Phi(w)).$$

The generalized gradient of Φ at $u \in X$ is the subset of the dual space X^* given by

$$\partial\Phi(u) := \left\{ u^* \in X^* : \langle u^*, v \rangle \leq \Phi^0(u; v), \quad \forall v \in X \right\}.$$

A continuous and convex function $\Phi : X \rightarrow \mathbb{R}$ is locally Lipschitz and its generalized gradient $\partial\Phi : X \rightarrow 2^{Y^*}$ coincides with the subdifferential of Φ in the sense of convex analysis. If $\Phi : X \rightarrow \mathbb{R}$ is a continuously differentiable function, its generalized gradient is just the differential $D\Phi$ of Φ .

The notion of generalized gradient is needed to handle the multivalued term $[\underline{f}(u), \overline{f}(u)]$ in problem (1). Given $f \in L^\infty_{\text{loc}}(\mathbb{R})$, we introduce

$$F(s) = \int_0^s f(t) dt \quad \text{for all } s \in \mathbb{R}.$$

The function $F : \mathbb{R} \rightarrow \mathbb{R}$ is locally Lipschitz and the generalized gradient $\partial F(s)$ of F at any $s \in \mathbb{R}$ is the compact interval in \mathbb{R}

$$\partial F(s) = [\underline{f}(s), \overline{f}(s)],$$

where $\underline{f}(s)$ and $\overline{f}(s)$ are the functions in (2) and (3), respectively (see, e.g., [5, Example 2.2.5]).

We also recall the definition of Palais–Smale condition and the statement of mountain-pass theorem in our nonsmooth setting. We say that a locally Lipschitz function $\Phi : X \rightarrow \mathbb{R}$ on a Banach space X satisfies the Palais–Smale condition at level $c \in \mathbb{R}$ (in short, $(PS)_c$ condition) if each sequence $\{u_n\}$ in X such that $\Phi(u_n) \rightarrow c$ and

$$\min_{\xi \in \partial\Phi(u_n)} \|\xi\| \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

possesses a convergent subsequence.

Theorem 1 Assume that a locally Lipschitz function $\Phi : X \rightarrow \mathbb{R}$ on a Banach space X satisfies the following conditions:

- (i) $\Phi(0) = 0$;
- (ii) there are constants $\rho > 0$ and $r > 0$ such that $\Phi(u) \geq r$ for all $u \in X$ with $\|u\|_X = \rho$;
- (iii) there exists an element $e \in X$ such that $\|e\|_X > \rho$ and $\Phi(e) < r$;
- (iv) Φ satisfies the $(PS)_c$ condition for

$$c = \inf_{\gamma \in \Gamma} \max_{t \in [0,1]} I(\gamma(t)),$$

where

$$\Gamma := \{ \gamma \in C([0, 1]; X) : \gamma(0) = 0 \text{ and } \gamma(1) = e \}.$$

Then $c \geq r$ and c is a critical value of Φ meaning that there exists $u \in X$ with $0 \in \partial\Phi(u)$ and $\Phi(u) = c$.

3 Statement of Main Result

We formulate our hypotheses on the data in problem (1).

- (m) $m : [0, +\infty) \rightarrow \mathbb{R}$ is a continuous function such that for the numbers $\{t_0, t_1, \dots, t_L\}$ with $0 = t_0 < t_1 < t_2 < \dots < t_L$ there hold
 - (i) $m(t_k) = 0$ for all $k = 0, 1, \dots, L$;
 - (ii) $m(t) > 0$ for all $t \in [0, t_L] \setminus \{t_0, t_1, \dots, t_L\}$.
- (f) $f \in L^\infty_{loc}(\mathbb{R})$ satisfies the conditions:

$$|f(t)| \leq \rho_f |t|^{q-1} + c_f, \quad \forall t \in \mathbb{R}, \tag{7}$$

with constants $q \in [1, p_s^*]$, $\rho_f \geq 0$, and $c_f \geq 0$;

$$F(t) := \int_0^t f(s) ds \leq C_f, \quad \forall t \in \mathbb{R}, \tag{8}$$

with a constant $C_f \geq 0$;

$$\mu_k := \frac{1}{p} \int_{t_{k-1}}^{t_k} m(s) ds > \frac{\rho_f}{q} S(q)^q t_k^{\frac{q}{p}} - c_f S(1) t_k^{\frac{1}{p}}, \quad \forall k \in \{1, 2, \dots, L\}; \tag{9}$$

$$t\eta > 0 \text{ for all } \eta \in [\underline{f}(t), \overline{f}(t)], \quad \forall t \in \mathbb{R} \setminus \{0\}; \tag{10}$$

$$|\Omega| \int_0^{+\infty} f(s) ds > \mu := \max\{\mu_1, \dots, \mu_L\}. \tag{11}$$

Remark 1 Hypothesis (f) requires to have a suitable balance between the functions f and m . The strict inequality $q < p_s^*$ and assumption (8) are needed for the Palais–Smale condition. Assumptions (9), (10), and (11) are helpful for the mountain-pass geometry.

Definition 1 We say that $u \in X_0$ is a (weak) solution of problem (1) if

$$\begin{aligned} & m(\|u\|_{X_0}^p) \int_{\mathbb{R}^{2N}} |u(x) - u(y)|^{p-2} (u(x) - u(y))(v(x) - v(y)) K(x - y) dx dy \\ & \geq \int_{\Omega} \min\{\underline{f}(u(x))v(x), \bar{f}(u(x))v(x)\} dx \text{ for all } v \in X_0. \end{aligned} \tag{12}$$

By replacing $v \in X_0$ with $-v$ it is seen that (12) is equivalent to

$$\begin{aligned} & m(\|u\|_{X_0}^p) \int_{\mathbb{R}^{2N}} |u(x) - u(y)|^{p-2} (u(x) - u(y))(v(x) - v(y)) K(x - y) dx dy \\ & \leq \max\{\underline{f}(u(x))v(x), \bar{f}(u(x))v(x)\} dx \text{ for all } v \in X_0. \end{aligned} \tag{13}$$

Due to $u \in X_0$ and (7) the integrals in Definition 1 exist. Now we are in a position to state the main result of the paper.

Theorem 2 *Assume that the conditions (K), (m) and (f) are fulfilled. Then problem (1) has at least L nontrivial solutions $u_1, \dots, u_L \in X_0$ satisfying the arrangement*

$$0 < \|u_1\|_{X_0}^p < t_1 < \|u_2\|_{X_0}^p < t_2 < \dots < t_{L-1} < \|u_L\|_{X_0}^p < t_L.$$

Theorem 2 will be proven through a special variational approach. The starting point is the energy functional $I : X_0 \rightarrow \mathbb{R}$ given by

$$I(u) := \frac{1}{p} M(\|u\|_{X_0}^p) - \int_{\Omega} F(u) dx \text{ for all } u \in X_0, \tag{14}$$

where $M : [0, \infty) \rightarrow \mathbb{R}$ is the function

$$M(t) := \int_0^t m(\tau) d\tau, \forall t \geq 0.$$

Here we prove that it is locally Lipschitz and determine its generalized gradient.

Lemma 1 *Under the growth condition (7) with $q \in [1, p_s^*]$, the functional I in (14) verifies:*

- (i) $I : X_0 \rightarrow \mathbb{R}$ is locally Lipschitz;
- (ii) if $\xi \in \partial I(u)$, with $u \in X_0$, then there exists $g \in L^q(\Omega)$ such that

$$\begin{aligned}
 & \langle \xi, v \rangle \tag{15} \\
 & = m(\|u\|_{X_0}^p) \int_Q |u(x) - u(y)|^{p-2} (u(x) - u(y))(v(x) - v(y)) K(x - y) \, dx dy \\
 & - \int_{\Omega} g v \, dx, \quad \forall v \in X_0
 \end{aligned}$$

and

$$g(x) \in [\underline{f}(u(x)), \overline{f}(u(x))] \text{ for a.e. } x \in \Omega. \tag{16}$$

Proof Define $J : X_0 \rightarrow \mathbb{R}$ by

$$J(u) := \frac{1}{p} M(\|u\|_{X_0}^p) \text{ for all } u \in X_0. \tag{17}$$

Let us show that J is Gâteaux differentiable. For $u, v \in X_0$ and $t \in (0, 1)$, we note

$$J(u + tv) - J(u) = \frac{1}{p} \left(M(\|u + tv\|_{X_0}^p) - M(\|u\|_{X_0}^p) \right).$$

The mean value theorem provides

$$\sigma_t \in \left(\min\{\|u\|_{X_0}^p, \|u + tv\|_{X_0}^p\}, \max\{\|u\|_{X_0}^p, \|u + tv\|_{X_0}^p\} \right)$$

such that

$$M(\|u + tv\|_{X_0}^p) - M(\|u\|_{X_0}^p) = m(\sigma_t) \left(\|u + tv\|_{X_0}^p - \|u\|_{X_0}^p \right).$$

We infer that

$$\begin{aligned}
 & \frac{1}{t} \left(M(\|u + tv\|_{X_0}^p) - M(\|u\|_{X_0}^p) \right) \\
 & = m(\sigma_t) \int_Q \frac{1}{t} \left[|u(x) - u(y) + t(v(x) - v(y))|^p - |u(x) - u(y)|^p \right] K(x - y) \, dx dy.
 \end{aligned}$$

Since $p > 1$, we have

$$\begin{aligned}
 & \lim_{t \rightarrow 0} \frac{1}{t} \left[|u(x) - u(y) + t(v(x) - v(y))|^p - |u(x) - u(y)|^p \right] \\
 & = p|u(x) - u(y)|^{p-2} (u(x) - u(y))(v(x) - v(y)) \text{ for a.e. } (x, y) \in Q.
 \end{aligned}$$

By a well-known convexity inequality, we obtain

$$\begin{aligned} & [|u(x) - u(y) + t(v(x) - v(y))|^p - |u(x) - u(y)|^p] K(x - y) \\ & \leq \xi(x, y) := 2^{p-1} |v(x) - v(y)|^p K(x - y) \text{ for all } t \in (0, 1), \end{aligned}$$

with $\xi \in L^1(Q)$ thanks to $v \in X_0$. Applying Lebesgue’s dominated convergence theorem in conjunction with the continuity of m yields

$$\begin{aligned} & \lim_{t \rightarrow 0} \frac{1}{t} \left(M(\|u + tv\|_{X_0}^p) - M(\|u\|_{X_0}^p) \right) \tag{18} \\ & = pm(\|u\|_{X_0}^p) \int_Q |u(x) - u(y)|^{p-2} (u(x) - u(y))(v(x) - v(y)) K(x - y) dx dy. \end{aligned}$$

By (17) and (18) we conclude that $J : X_0 \rightarrow \mathbb{R}$ is Gâteaux differentiable.

Next we prove that the Gâteaux differential $J' : X_0 \rightarrow X_0^*$ is continuous. Let $u_n \rightarrow u$ in X_0 for some $u \in X_0$. Thus $u_n \rightarrow u$ in $L^p(\Omega)$ and along a relabeled subsequence $u_n(x) \rightarrow u(x)$ for a.e. $x \in \Omega$ and

$$\begin{aligned} h_n(x, y) & := |u_n(x) - u_n(y)|^{p-2} (u_n(x) - u_n(y)) K(x - y)^{\frac{1}{p'}} \rightarrow \\ h(x, y) & := |u(x) - u(y)|^{p-2} (u(x) - u(y)) K(x - y)^{\frac{1}{p'}} \end{aligned}$$

for a.e. $(x, y) \in Q$. Since the sequence $\{h_n\}$ is bounded in $L^{p'}(Q)$, Brézis–Lieb lemma in [3] ensures that

$$\lim_{n \rightarrow \infty} (\|h_n\|_{L^{p'}(Q)}^{p'} - \|h_n - h\|_{L^{p'}(Q)}^{p'}) = \|h\|_{L^{p'}(\Omega)}^{p'}.$$

Taking into account the strong convergence $u_n \rightarrow u$ in X_0 , it turns out

$$\begin{aligned} & \lim_{n \rightarrow \infty} \int_Q \left| |u_n(x) - u_n(y)|^{p-2} (u_n(x) - u_n(y)) - |u(x) - u(y)|^{p-2} (u(x) - u(y)) \right|^{p'} \\ & \times K(x - y) dx dy = \lim_{n \rightarrow \infty} \|h_n - h\|_{L^{p'}(Q)}^{p'} = \lim_{n \rightarrow \infty} (\|u_n\|_{X_0}^p - \|u\|_{X_0}^p) = 0. \tag{19} \end{aligned}$$

By Hölder’s inequality we get

$$\begin{aligned} & \left| \int_Q |u(x) - u(y)|^{p-2} (u(x) - u(y))(v(x) - v(y)) K(x - y) dx dy \right| \\ & \leq \int_Q |u(x) - u(y)|^{p-1} K(x - y)^{\frac{p-1}{p}} |v(x) - v(y)| K(x - y)^{\frac{1}{p}} dx dy \leq \|u\|_{X_0}^{p-1} \|v\|_{X_0}. \end{aligned}$$

Then, in view of (17) and (18), we arrive at

$$\begin{aligned} \|J'(u_n) - J'(u)\|_{X_0^*} &= \sup_{\|v\|_{X_0} \leq 1} |\langle J'(u_n) - J'(u), v \rangle| \\ &\leq C_1 \left(\int_Q \left| |u_n(x) - u_n(y)|^{p-2} (u_n(x) - u_n(y)) - |u(x) - u(y)|^{p-2} (u(x) - u(y)) \right|^{p'} \right. \\ &\quad \left. \times K(x - y) dx dy \right)^{\frac{1}{p'}} + \left| m(\|u_n\|_{X_0}^p) - m(\|u\|_{X_0}^p) \right| \|u\|_{X_0}^{p-1}, \end{aligned} \tag{20}$$

with a constant $C_1 > 0$. Combining (19) and (20) allows us to conclude that

$$\lim_{n \rightarrow \infty} \|J'(u_n) - J'(u)\|_{X_0^*} = 0,$$

thereby $J \in C^1(X_0, \mathbb{R})$.

Define $\Psi : L^q(\Omega) \rightarrow \mathbb{R}$ by

$$\Psi(u) := \int_{\Omega} F(u) dx \text{ for all } u \in L^q(\Omega). \tag{21}$$

We show that Ψ is Lipschitz continuous on the bounded subsets of $L^q(\Omega)$. To this end, let S be a bounded subset of $L^q(\Omega)$ and let $u, v \in L^q(\Omega)$. Since the function F is locally Lipschitz, we can use Lebourg’s mean value theorem (see [5]) obtaining

$$F(u(x)) - F(v(x)) = \omega(u(x) - v(x)) \text{ for a.e. } x \in \Omega,$$

with a real number $\omega = \omega(x)$ belonging to the open real interval determined by $u(x)$ and $v(x)$. Then the growth condition (7) implies

$$|F(u(x)) - F(v(x))| \leq (\rho_f \max\{|u(x)|^{q-1}, |v(x)|^{q-1}\} + c_f) |u(x) - v(x)| \text{ for a.e. } x \in \Omega.$$

Assumption (7), Hölder’s inequality, and (21) imply

$$\begin{aligned} |\Psi(u) - \Psi(v)| &= \left| \int_{\Omega} (F(u) - F(v)) dx \right| \\ &\leq \int_{\Omega} (\rho_f \max\{|u(x)|^{q-1}, |v(x)|^{q-1}\} + c_f) |u(x) - v(x)| dx \\ &\leq \rho_f \max\{\|u\|_{L^q(\Omega)}^{q-1}, \|v\|_{L^q(\Omega)}^{q-1}\} \|u - v\|_{L^q(\Omega)} + c_f |\Omega|^{\frac{1}{q'}} \|u - v\|_{L^q(\Omega)}. \end{aligned}$$

Since S is a bounded subset of $L^q(\Omega)$, it results that the function Ψ in (21) is Lipschitz continuous on the bounded subsets of $L^q(\Omega)$, in particular locally Lipschitz.

Denote by $i : X_0 \rightarrow L^q(\Omega)$ the embedding $i(u) = u$ for all $u \in X_0$, which is linear and continuous. Therefore the restriction $\Psi|_{X_0} = \Psi \circ i$ is a locally Lipschitz function on X_0 . It suffices to note from (14) that $I = J - \Psi|_{X_0}$ for achieving the proof of part (i).

Taking into account that X_0 is dense in $L^q(\Omega)$, we have the following formula regarding the generalized gradient:

$$\partial(\Psi|_{X_0})(u) = i^* \partial \Psi(u) \text{ for all } u \in X_0,$$

where $i^* : L^q(\Omega) \rightarrow X_0^*$ stands for the adjoint of i . The growth condition (7) allows us to invoke Aubin–Clarke theorem (see [5]) to deduce that for every $\xi \in \partial(\Psi|_{X_0})(u)$ there exists $g \in L^{q'}(\Omega)$ such that

$$\langle \xi, v \rangle = \int_{\Omega} g(x)v(x) dx \text{ for all } v \in X_0.$$

Hence (15) and (16) hold true. □

4 Truncated Problems

Our approach relies on the truncations m_k of the function m in (1) given by

$$m_k(t) = \begin{cases} m(t) & \text{if } t_{k-1} \leq t < t_k \\ 0 & \text{elsewhere,} \end{cases} \tag{22}$$

for $k \in \{1, 2, \dots, L\}$. Corresponding to the truncations in (22) we formulate the intermediate fractional Kirchhoff problems

$$\begin{cases} m_k(\|u\|_{X_0}^p) \mathfrak{L}_K^p u \in [f(u), \bar{f}(u)] & \text{in } \Omega \\ u = 0 & \text{in } \mathbb{R}^N \setminus \Omega. \end{cases} \tag{23}$$

The solutions of (23) are understood in the weak sense of Definition 1. The energy functional $I_k : X_0 \rightarrow \mathbb{R}$ for problem (23) is defined by

$$I_k(u) = \frac{1}{p} M_k(\|u\|_{X_0}^p) - \int_{\Omega} F(u) dx, \tag{24}$$

with

$$M_k(t) = \int_0^t m_k(\tau) d\tau. \tag{25}$$

For problem (23), Lemma 1 reads as follows.

Lemma 2 *Under the growth condition (7) with $q \in [1, p_s^*]$, the functional I_k in (24) verifies:*

- (i) $I_k : X_0 \rightarrow \mathbb{R}$ is locally Lipschitz;
- (ii) if $\xi \in \partial I_k(u)$, with $u \in X_0$, then there exists $g_k \in L^q(\Omega)$ such that

$$\langle \xi, v \rangle \tag{26}$$

$$= m_k(\|u\|_{X_0}^p) \int_{\Omega} |u(x)-u(y)|^{p-2}(u(x)-u(y))(v(x)-v(y))K(x-y) dx dy - \int_{\Omega} g_k v dx, \forall v \in X_0$$

and

$$g_k(x) \in [\underline{f}(u_k(x)), \overline{f}(u_k(x))] \text{ for a.e. } x \in \Omega. \tag{27}$$

A priori estimates for solutions to problem (23) are available as shown below.

Proposition 1 *Under the assumptions of Lemma 2 and assuming in addition (10), if $u_k \in X_0$ is a nontrivial critical point of I_k , that is $0 \in \partial I_k(u_k)$ with $u_k \neq 0$, then it holds*

$$t_{k-1}^{\frac{1}{p}} < \|u_k\|_{X_0} < t_k^{\frac{1}{p}}. \tag{28}$$

Proof Suppose by contradiction that $\|u_k\|_{X_0} \geq t_k^{\frac{1}{p}}$ or $\|u_k\|_{X_0} \leq t_{k-1}^{\frac{1}{p}}$. Then from (23) we infer that $m_k(\|u_k\|_{X_0}^p) = 0$, which by $0 \in \partial I_k(u_k)$ and (26) results in

$$\int_{\Omega} g_k(x)v(x) dx = 0 \text{ for all } v \in X_0,$$

so $g_k(x) = 0$ for a.e. $x \in \Omega$. By hypothesis (10) and in conjunction with (27) we find that $u_k(x) = 0$ for a.e. $x \in \Omega$. This contradicts the fact that u_k is nontrivial, thus completing the proof. □

Corollary 1 *Assume the conditions of Proposition 1. If u_k is a nontrivial critical point of I_k , then u_k is a nontrivial weak solution to problem (23) and a nontrivial weak solution to problem (1). Moreover, if $k \neq j$ with $k, j \in \{1, \dots, t_L\}$, then we get distinct solutions $u_k \neq u_j$ of problem (1).*

Proof Let $u_k \in X_0$ be a nontrivial critical point of I_k . From Proposition 1 we know that (28) holds true. Then Lemmas 1(ii), 2(ii), and (22) guarantee that $\partial I_k(u_k) = \partial I(u_k)$, which proves that $0 \in \partial I(u_k)$. Hence it holds

$$\begin{aligned}
 & m(\|u_k\|_{X_0}^p) \int_Q |u_k(x) - u_k(y)|^{p-2} (u_k(x) - u_k(y))(v(x) - v(y)) K(x - y) \, dx dy \\
 & \geq \int_{\Omega} g_k v \, dx, \quad \forall v \in X_0,
 \end{aligned}$$

with some $g_k \in L^{q'}(\Omega)$ satisfying (28), which leads to

$$\begin{aligned}
 & m(\|u_k\|_{X_0}^p) \int_{\mathbb{R}^{2N}} |u_k(x) - u_k(y)|^{p-2} (u_k(x) - u_k(y))(v(x) - v(y)) K(x - y) \, dx dy \\
 & \geq \int_{\Omega} \min\{\underline{f}(u_k(x))v(x), \overline{f}(u_k(x))v(x)\} \, dx \quad \text{for all } v \in X_0.
 \end{aligned}$$

Consequently, u_k solves (1) and a fortiori (23). The second assertion in the statement of corollary is the direct consequence of (28) and the partition $t_0 = 0 < t_1 < \dots < t_L$. □

5 Mountain-Pass Geometry

Now we focus on the geometry of mountain-pass theorem for the functional I_k in (24), with $k = 1, \dots, L$.

Lemma 3 *Assume the conditions of Lemma 2 and in addition (9), (10), and (11). Then there hold:*

(i) *there exist positive constants ϑ_k and r_k such that*

$$I_k(u) \geq \vartheta_k \quad \text{for all } u \in X_0 \text{ with } \|u\|_{X_0} = r_k;$$

(ii) *there exists $e_k \in X_0$ satisfying $e_k(x) \geq 0$ for a.e. $x \in \Omega$, $I_k(e_k) \leq 0 < \vartheta_k$ and $\|e_k\|_{X_0} > r_k$.*

Proof

(i) Let $u \in X_0$ with $\|u\|_{X_0}^p = t_k$. We invoke (24), (25), (22), (7), and (9) to obtain

$$\begin{aligned}
 I_k(u) &= \frac{1}{p} \int_{t_{k-1}}^{t_k} m(s) \, ds - \int_{\Omega} \int_0^{u(x)} f(s) \, ds \, dx \\
 &\geq \mu_k - \frac{\rho_f}{q} \|u\|_{L^q(\Omega)}^q - c_f \|u\|_{L^q(\Omega)} \geq \mu_k - \frac{\rho_f}{q} S(q)^q t_k^{\frac{q}{p}} - c_f S(1) t_k^{\frac{1}{p}} > 0.
 \end{aligned}$$

Assertion (i) is verified by taking $\vartheta_k = \mu_k - \frac{\rho f}{q} S(q)^q t_k^{\frac{q}{p}} - c_f S(1) t_k^{\frac{1}{p}}$ and $r_k = t_k^{\frac{1}{p}}$.

(ii) By assumption (11) there are an open set Ω_0 in \mathbb{R}^N with $\overline{\Omega_0} \subset \Omega$ and a constant $\beta > 0$ for which one has

$$|\Omega_0| \int_0^\beta f(s) ds \geq \mu. \tag{29}$$

Fix $\xi \in C_0^\infty(\Omega)$ with $\xi \geq 0$ on Ω and $\xi \equiv 1$ on Ω_0 . Then, for each $k = 1, 2, \dots, L$, we can find $l_k \geq \beta$ such that for $e_k = l_k \xi$ it holds

$$\|e_k\|_{X_0} > t_k^{\frac{1}{p}}.$$

Then (24), (25), and (22), the properties of e_k , (10), and (29) imply

$$I_k(e_k) = \mu_k - \int_\Omega \int_0^{e_k(s)} f(s) ds dx \leq \mu_k - |\Omega_0| \int_0^\beta f(s) ds \leq 0,$$

thus part (ii) holds true. □

Set

$$c_k := \inf_{\gamma \in \Gamma_k} \max_{t \in [0,1]} I_k(\gamma(t)), \tag{30}$$

where

$$\Gamma_k := \{ \gamma \in C([0, 1], X_0) : \gamma(0) = 0 \text{ and } \gamma(1) = e_k \} \text{ for } k \in \{1, 2, \dots, L\},$$

with e_k in Lemma 3(ii). We estimate from above the minimax values c_k .

Lemma 4 *Assume the conditions of Lemma 3. Then we have*

$$c_k < \mu_k \text{ for all } k \in \{1, 2, \dots, L\}.$$

Proof For each $k \in \{1, 2, \dots, L\}$, consider the path $\gamma_k^*: [0, 1] \rightarrow X_0$ defined by $\gamma_k^*(t) = t e_k$ for all $t \in [0, 1]$, with e_k in Lemma 3(ii), which evidently belongs to the set Γ_k . There exists $t_k^* \in (0, 1)$ such that

$$I_k(t_k^* e_k) = \max_{t \in [0,1]} I_k(\gamma_k^*(t)) = \max_{t \in [0,1]} I_k(t e_k).$$

We have $t_k^* \in (0, 1)$ because $I_k(0) = 0$, $I_k(e_k) \leq 0$ and $I_k(u) \geq \vartheta_k > 0$ for all $u \in X_0$ with $\|u\|_{X_0} = r_k$ (see Lemma 3). Recall that $t_k^* e_k(x) \geq 0$ for a.e. $x \in \Omega$

and $t_k^* e_k \neq 0$. Then by (30), (24), (25), (22), and (10) we infer that

$$\begin{aligned}
 c_k &\leq \max_{t \in [0,1]} I_k(\gamma_k^*(t)) = I_k(t_k^* e_k) = \frac{1}{p} M_k(\|t_k^* e_k\|_{X_0}^p) - \int_{\Omega} F(t_k^* e_k) dx \\
 &< \frac{1}{p} M_k(\|t_k^* e_k\|_{X_0}^p) \leq \frac{1}{p} \int_{t_{k-1}}^{t_k} m(s) ds = \mu_k,
 \end{aligned}$$

which completes the proof. □

6 Palais–Smale Condition

The next statement concerns the validity of $(PS)_{c_k}$ condition.

Lemma 5 *Assume the conditions in Theorem 2. Then, for each $k \in \{1, 2, \dots, L\}$, the functional I_k satisfies the $(PS)_{c_k}$ condition.*

Proof Fix $k \in \{1, 2, \dots, L\}$ and let $\{u_n\} \subset X_0$ be a $(PS)_{c_k}$ sequence for the functional I_k in (24). This reads as

$$I_k(u_n) = \frac{1}{p} M_k(\|u_n\|_{X_0}^p) - \int_{\Omega} F(u_n) ds \rightarrow c_k \text{ as } n \rightarrow \infty \tag{31}$$

and

$$\min_{\xi \in \partial(I_k)(u_n)} \|\xi\| \rightarrow 0 \text{ as } n \rightarrow \infty. \tag{32}$$

Since $\partial(I_k)(u_n)$ is a nonempty weakly compact subset of X_0^* , the minimum in (32) is attained giving rise to a sequence $\xi_n \in \partial(I_k)(u_n)$ with $\xi_n \rightarrow 0$ in X_0^* . Combining $\xi_n \in \partial(I_k)(u_n)$ with Aubin–Clarke theorem (see [5]) ensures that there exists $h_n \in L^{q'}(\Omega)$ such that

$$\langle \xi_n, v \rangle \tag{33}$$

$$\begin{aligned}
 &= m_k(\|u_n\|_{X_0}^p) \int_Q |u_n(x) - u_n(y)|^{p-2} (u_n(x) - u_n(y))(v(x) - v(y)) K(x - y) dx dy \\
 &- \int_{\Omega} h_n v dx, \quad \forall v \in X_0
 \end{aligned}$$

and

$$h_n(x) \in \partial F(u_n(x)) = [\underline{f}(u_n(x)), \overline{f}(u_n(x))] \text{ for a.e. } x \in \Omega. \tag{34}$$

We claim that the sequence $\{u_n\}$ is bounded in X_0 . Suppose by contradiction that for a relabeled subsequence of $\{u_n\}$ we have

$$\|u_n\|_{X_0} \rightarrow \infty \text{ as } n \rightarrow \infty. \tag{35}$$

According to (35), we may admit that $\|u_n\|_{X_0} \geq t_k^{\frac{1}{p}}$. Thus by (22) and (25) it holds $M_k(\|u_n\|_{X_0}^p) = \mu_k$. Owing to (35), we see that (31) reduces to

$$\lim_{n \rightarrow \infty} \int_{\Omega} F(u_n) dx = \mu_k - c_k. \tag{36}$$

From (22) we know that $m_k(t) = 0$ for all $t > t_k$. Hence $\xi_n \rightarrow 0$ in X_0^* , (33) and (35) imply that $h_n \rightarrow 0$ in $L^q(\Omega)$ as $n \rightarrow \infty$, thus up to a subsequence $h_n(x) \rightarrow 0$ for a.e. $x \in \Omega$. This ensures that $u_n(x) \rightarrow 0$ as $n \rightarrow \infty$ for a.e. $x \in \Omega$. Indeed, arguing pointwise a.e. admit for an $x \in \Omega$ that along a subsequence $u_n(x) \rightarrow l$ with $l \neq 0$. By (7) and (34) (see also (2) and (3)) we have that the sequence $(h_n(x))$ is bounded. Since the graph of ∂F is closed in $\mathbb{R} \times \mathbb{R}$, we deduce from $h_n(x) \rightarrow 0$ and (34) that

$$0 \in \partial F(l) = [\underline{f}(l), \overline{f}(l)]$$

with $l \neq 0$, which contradicts hypothesis (10). This contradiction confirms that $u_n(x) \rightarrow 0$ as $n \rightarrow \infty$ for a.e. $x \in \Omega$.

By the continuity of F we have

$$F(u_n(x)) \rightarrow F(0) = 0 \text{ for a.e. } x \in \Omega, \tag{37}$$

while assumption (8) entails

$$F(u_n(x)) \leq C_f \text{ for a.e. } x \in \Omega, \text{ all } n \in \mathbb{N}. \tag{38}$$

Relying on (37) and (38) we are allowed to apply Fatou’s Lemma with limit superior that gives

$$\lim_{n \rightarrow \infty} \int_{\Omega} F(u_n) dx \leq 0. \tag{39}$$

The existence of the limit in (39) is guaranteed by (36). Consequently, from (36) and (39), we deduce that $\mu_k \leq c_k$ contradicting Lemma 4. Therefore the sequence $\{u_n\}$ is bounded in X_0 .

Through the reflexivity of X_0 there exists a subsequence of $\{u_n\}$, again denoted $\{u_n\}$, such that

$$u_n \rightarrow w_k \text{ weakly in } X_0 \text{ and } \|u_n\|_{X_0}^p \rightarrow d_k \text{ as } n \rightarrow \infty, \tag{40}$$

for some $w_k \in X_0$ and for a constant $d_k \geq 0$. Let us prove that

$$t_{k-1} < d_k \leq t_k, \quad \forall k = 1, \dots, L, \tag{41}$$

for which we argue indirectly. Assume $t_k < d_k$. By (40) we may suppose $\|u_n\|_{X_0} \geq t_k^{\frac{1}{p}}$, so $M_k(\|u_n\|_{X_0}^p) = \mu_k$ according to (22), (25). Note that (31) leads to (36). Carrying out the argument as before, a contradiction with Lemma 4 arises. If we admit that $d_k \leq t_{k-1}$, by (22) and (25) we get

$$\lim_{n \rightarrow \infty} M_k(\|u_n\|_{X_0}^p) = M_k(d_k) = 0.$$

As above, on the basis of (32), (33), and (34) we derive that $u_n(x) \rightarrow 0$ for a.e. $x \in \Omega$ along a relabeled subsequence. Complying with assumption (f) we have $q < p_s^*$, which renders from (40) that $u_n \rightarrow w_k$ strongly in $L^q(\Omega)$ and $F(u_n) \rightarrow F(w_k)$ strongly in $L^1(\Omega)$. Altogether we derive that $w_k = 0$ and $I(u_n) \rightarrow I(w_k) = 0$. Then (31) forces $c_k = 0$, which contradicts $c_k \geq \vartheta_k > 0$ (see Lemma 3(i)). Thus (41) holds true.

In order to complete the proof it suffices to show that (a subsequence of) $\{u_n\}$ converges strongly to w_k in X_0 . Passing to a relabeled subsequence, in view of (40) and $q < p_s^*$, we have that $u_n \rightarrow w_k$ in $L^q(\Omega)$, whereas (34) and (7) provide $h_n \rightarrow h$ weakly in $L^{q'}(\Omega)$ for some $h \in L^{q'}(\Omega)$. Due to (34) and the (strong \times weak)-closedness of the graph of the generalized gradient $\partial\Psi$ for Ψ in (21), we can infer that

$$h(x) \in \partial F(w_k(x)) = [\underline{f}(w_k(x)), \overline{f}(w_k(x))] \text{ for a.e. } x \in \Omega. \tag{42}$$

Insert $v = u_n$ and $v = w_k$ in (33) and let $n \rightarrow \infty$ in the resulting equalities. Using $\xi_n \rightarrow 0$ in X_0^* , (40), $u_n \rightarrow w_k$ strongly in $L^q(\Omega)$ and $h_n \rightarrow h$ weakly in $L^{q'}(\Omega)$, we find that

$$m_k(d_k)d_k = \int_{\Omega} h w_k \, dx \quad \text{and} \quad m_k(d_k)\|w_k\|_{X_0}^p = \int_{\Omega} h w_k \, dx. \tag{43}$$

It turns out

$$m_k(d_k)(\|w_k\|_{X_0}^p - d_k) = 0. \tag{44}$$

If $m_k(d_k) = 0$, then (43) renders

$$\int_{\Omega} h w_k \, dx = 0.$$

Taking into account (42) and assumption (10) results in $h w_k = 0$. Using again assumption (10), it follows that $w_k = 0$. Then from (36) we infer that $\mu_k = c_k$,

which contradicts Lemma 4. This contradiction enables us to deduce from (44) the equality $d_k = \|w_k\|_{X_0}^p$. The uniform convexity of the space X_0 and (40) with $d_k = \|w_k\|_{X_0}^p$ imply that $u_n \rightarrow w_k$ strongly in X_0 , thus completing the proof. \square

7 Proof of Theorem 2 and Example

Now we are able to prove Theorem 2. For every $k \in \{1, 2, \dots, L\}$, consider the intermediate problem (23) and its associated Euler functional $I_k : X_0 \rightarrow \mathbb{R}$ given by (24), (25). The critical points of I_k coincide with the (weak) solutions of (23). We observe that the hypotheses of Theorem 1 are verified for the functional I_k on the Banach space X_0 . Precisely, Lemma 2 sets forth that $I_k \in C^1(X_0, \mathbb{R})$, whereas the conditions (i), (ii), (iii) in Theorem 1 are fulfilled for I_k due to Lemma 3. Condition (iv) in Theorem 1 is satisfied for I_k with $c = c_k$ because of Lemma 5. We are thus allowed to apply Theorem 1 in the case of the locally Lipschitz functional $I_k : X_0 \rightarrow \mathbb{R}$ introduced in (24), (25). Applying Theorem 1 provides the existence of a nontrivial critical point $u_k \in X_0$ of I_k in the sense that $0 \in \partial I(u_k)$ with $u_k \neq 0$, so the existence of a nontrivial solution u_k of the intermediate problem (23) is guaranteed. Furthermore, as known from Proposition 1, the a priori estimate (28) is valid. Now, by Corollary 1, it appears that u_k is a solution to problem (1) satisfying the estimate $t_{k-1} < \|u_k\|_{X_0}^p < t_k$. Since this is true for each $k \in \{1, 2, \dots, L\}$, the proof of Theorem 2 is complete.

Here is a simple example permitting to identify a class of problems (1) for which the conditions required in Theorem 2 are fulfilled. A rather general procedure to construct examples is in this way indicated. Different other schemes can be built up.

Example 1 Fix a measurable function $K : \mathbb{R}^N \setminus \{0\} \rightarrow (0, +\infty)$ satisfying hypothesis (K), for instance $K(x) = |x|^{-(N+ps)}$ and fix any function $f \in L_{loc}^\infty(\mathbb{R})$ verifying the growth and sign requirements in (7), (8), and (10) (their range of applicability is large). In particular, by (10) we have that

$$\int_0^{+\infty} f(s) ds > 0.$$

Without loss of generality we may suppose in (7) that $q > 1$, $\rho_f > 0$, and $c_f > 0$. Take finitely many positive numbers $0 = t_0 < t_1 < t_2 < \dots < t_L$ (with any positive integer L) such that

$$\frac{\rho_f}{q} S(q)^q t_k^{\frac{q-1}{p}} < c_f S(1), \quad \forall k \in \{1, 2, \dots, L\}, \tag{45}$$

which can be achieved provided each $t_k > 0$ is sufficiently small.

Corresponding to the numbers $\{t_0, t_1, \dots, t_L\}$ let us choose a continuous function $\tilde{m} : [0, +\infty) \rightarrow \mathbb{R}$ vanishing at $\{t_0, t_1, \dots, t_L\}$ and being positive on $[0, t_L] \setminus$

$\{t_0, t_1, \dots, t_L\}$, that is, \tilde{m} complies with hypothesis (m). We note from (45) that setting $m = c\tilde{m}$, with a positive constant $c > 0$, one has that (9) is automatically satisfied. In addition, there holds condition (11), i.e.,

$$|\Omega| \int_0^{+\infty} f(s) ds > \frac{c}{p} \int_{t_{k-1}}^{t_k} \tilde{m}(s) ds, \quad k = 1, \dots, L,$$

provided the constant $c > 0$ is sufficiently small. We conclude for K , f , m and $0 = t_0 < t_1 < t_2 < \dots < t_L$ chosen as above that we can apply Theorem 2 to the corresponding problem (1).

References

1. G. Autuori, P. Pucci, M.C. Salvatori, Global nonexistence for nonlinear Kirchhoff systems. *Arch. Ration. Mech. Anal.* **196**, 489–516 (2010)
2. G. Autuori, A. Fiscella, P. Pucci, Stationary Kirchhoff problems involving a fractional elliptic operator and a critical nonlinearity. *Nonlinear Anal.* **125**, 699–714 (2015)
3. H. Brézis, E. Lieb, A relation between pointwise convergence of functions and convergence of functionals. *Proc. Am. Math. Soc.* **88**, 486–490 (1983)
4. K.C. Chang, Variational methods for non-differentiable functionals and their applications to partial differential equations. *J. Math. Anal. Appl.* **80**, 102–129 (1981)
5. F.H. Clarke, *Optimization and Nonsmooth Analysis* (Wiley, New York, 1983)
6. E. Di Nezza, G. Palatucci, E. Valdinoci, Hitchhiker’s guide to the fractional Sobolev spaces. *Bull. Sci. Math.* **136**, 521–573 (2012)
7. G.M. Figueiredo, J.R. Santos Júnior, Existence of a least energy nodal solution for a Schrödinger–Kirchhoff equation with potential vanishing at infinity. *J. Math. Phys.* **56**, 051506 (2015)
8. L. Gasiński, J.R. Santos Júnior, Multiplicity of positive solutions for an equation with degenerate nonlocal diffusion. *Comput. Math. Appl.* **78**, 136–143 (2019)
9. Z.J. Guo, Ground states for Kirchhoff equations without compact condition. *J. Differ. Equ.* **259**, 2884–2902 (2015)
10. G. Kirchhoff, in *Vorlesungen ueber Mathematische Physik, Mechanik*. Lecture vol. 19 (Teubner, Leipzig, 1877)
11. Z.H. Liu, D. Motreanu, S. Zeng, Existence of solutions to Kirchhoff-type problem with vanishing nonlocal term and fractional p -Laplacian. Preprint
12. Z.H. Liu, J.G. Tan, Nonlocal elliptic hemivariational inequalities. *Electron. J. Qual. Theory Differ. Equ.* Paper 66 (2017), p. 7
13. D.F. Lü, S.J. Peng, Existence and asymptotic behavior of vector solutions for coupled nonlinear Kirchhoff-type systems. *J. Differ. Equ.* **263**, 8947–8978 (2017)
14. S. Migórski, S.D. Zeng, Mixed variational inequalities driven by fractional evolutionary equations. *Acta Math. Sci.* **39**, 461–468 (2019)
15. S. Migórski, V.T. Nguyen, S.D. Zeng, Nonlocal elliptic variational-hemivariational inequalities. *J. Integr. Equ. Appl.* **32**, 51–58 (2020)
16. O.H. Miyagaki, D. Motreanu, F.R. Pereira, Multiple solutions for a fractional elliptic problem with critical growth. *J. Differ. Equ.* **269**, 5542–5572 (2020)
17. G. Molica Bisci, V.D. Rădulescu, R. Servadei, *Variational Methods for Nonlocal Fractional Problems* (Cambridge University Press, Cambridge, 2016)
18. D. Motreanu, P.D. Panagiotopoulos, *Minimax Theorems and Qualitative Properties of the Solutions of Hemivariational Inequalities*. *Nonconvex Optimization and its Applications*, vol.

- 29 (Kluwer Academic Publishers, Dordrecht, 1999)
19. N. Pan, B.L. Zhang, J. Cao, Degenerate Kirchhoff-type diffusion problems involving the fractional p -Laplacian. *Nonlinear Anal. Real World Appl.* **37**, 56–70 (2017)
 20. K. Perera, Z. Zhang, Nontrivial solutions of Kirchhoff-type problems via the Yang index. *J. Differ. Equ.* **221**, 246–255 (2006)
 21. B. Ricceri, Energy functionals of Kirchhoff-type problems having multiple global minima. *Nonlinear Anal.* **115**, 130–136 (2015)
 22. J.R. Santos Júnior, G. Siciliano, Positive solutions for Kirchhoff problems with vanishing nonlocal term. *J. Differ. Equ.* **265**, 2034–2043 (2018)
 23. M.Q. Xiang, V.D. Rădulescu, B.L. Zhang, Nonlocal Kirchhoff diffusion problems: local existence and blow-up of solutions. *Nonlinearity* **31**, 3228–3250 (2018)
 24. M.Q. Xiang, B.L. Zhang, M. Ferrara, Existence of solutions for Kirchhoff type problem involving the non-local fractional p -Laplacian. *J. Math. Anal. Appl.* **424**, 1021–1041 (2015)
 25. S.D. Zeng, Z.H. Liu, S. Migórski, A class of fractional differential hemivariational inequalities with application to contact problem. *Z. Angew. Math. Phys.* **69**, 36 (2018)

Competition for Medical Supplies Under Stochastic Demand in the Covid-19 Pandemic: A Generalized Nash Equilibrium Framework



Anna Nagurney, Mojtaba Salarpour, June Dong, and Pritha Dutta

Abstract The Covid-19 pandemic has negatively impacted virtually all economic and social activities across the globe. Presently, since there is still no vaccine and no curative treatments for this disease, medical supplies in the form of Personal Protective Equipment and ventilators are sorely needed for healthcare workers and certain patients, respectively. The fact that this healthcare disaster is not limited in time and space has resulted in intense global competition for medical supplies. In this paper, we construct the first Generalized Nash Equilibrium model with stochastic demands to model competition among organizations at demand points for medical supplies. The model includes multiple supply points and multiple demand points, along with prices of the medical items and generalized costs associated with transportation. The theoretical constructs are provided and a Variational Equilibrium utilized to enable alternative variational inequality formulations. A qualitative analysis is presented and an algorithm proposed, along with convergence results. Illustrative examples are detailed as well as numerical examples that are solved with the implemented algorithm. The results reveal the impacts of the addition of supply points as well as of demand points on the medical item product flows. The formalism may be adapted to multiple medical items both in the near term and in the longer term (such as for vaccines).

A. Nagurney (✉) · M. Salarpour
Department of Operations and Information Management, Isenberg School of Management,
University of Massachusetts, Amherst, MA, USA
e-mail: nagurney@isenberg.umass.edu

J. Dong
Department of Marketing & Management, School of Business, State University of New York,
Oswego, NY, USA

P. Dutta
Department of Management and Management Science, Lubin School of Business, Pace
University, New York, NY, USA

1 Introduction

The Covid-19 pandemic, which was declared a pandemic by the World Health Organization on March 11, 2020 (cf. Secon et al. [60]), has disrupted the globe, altering economic and social activities and negatively impacting education, travel, work, and even leisure. Healthcare systems around the world continue to face great challenges as the battle against the novel coronavirus that causes this disease continues. The great need for medical items from Personal Protective Equipment (PPEs) to ventilators and, now, even convalescent plasma has led to intense competition for medical supplies among healthcare institutions and even regions, including states, as well as nations. Although a vaccine is not yet available and a cure does not yet exist, scientific advances are adding to knowledge regarding possible treatments. However, even when a vaccine becomes available, one can expect, because of the great demands and potential insufficiency of manufacturing capacity as well as vaccine components for distribution, that competition will be a reality for the foreseeable future even for vaccines. The same holds for medicinal treatments for patients suffering from Covid-19.

Indeed, the competition for PPEs, to start, is reasonable, since it has been scientifically established that one of the most effective ways to mitigate contagion associated with the novel coronavirus [26] is to use Personal Protective Equipment (PPE), for healthcare and other essential workers (see Jacobs et al. [25]) as well as those in social proximity [8, 23]. China has historically produced half of the world's face masks, but with the coronavirus originating in Wuhan, China, the country dedicated the majority of the supply for their own citizens, whereas other countries, such as Germany, even banned the export of PPEs [34]. The intense competition for PPEs led to a dramatic increase in the price, with some prices rising by more than 1000%, according to the report by The Society for Healthcare Organization Procurement Professionals [61]. For example (cf. Diaz et al. [10] and Berklan [5]), the price of N95 masks grew from \$0.38 to \$5.75 each (a 1413% increase); isolation protective gowns experienced a price increase from \$0.25 to \$5.00 (a 1900% increase), with the price of reusable face shields going from \$0.50 to \$4.00 (a 700% increase). According to Glenza [17], demand and prices for PPEs, as of the end of June 2020, are dramatically increasing again across the United States as coronavirus cases continue to rise in more than half of states. Furthermore, shortages of PPEs are again being reported in the United States in July as medical and dental practices reopen and with the reopening of some schools also on the horizon.

In addition, because the coronavirus SARS-CoV-2 that causes Covid-19 may result in severe respiratory problems in certain individuals, various healthcare organizations, including hospitals, were clamoring for ventilators for their patients [16, 52]. This is an example of, yet, another medical item for which there was and continues to be intense competition globally, and with limited supply availability (see [18, 28, 56, 58, 59]). The supply chain for ventilators is quite complex, with components sourced from different countries.

There is a growing demand for another medical product that has become critical in the healthcare system due to the pandemic. It is the plasma or liquid part of blood obtained from recovered Covid-19 patients, also known as convalescent plasma. It contains antibodies that can fight the virus SARS-CoV-2 causing the Covid-19 disease [22]. The pandemic has given rise to a rather unique competitive market for convalescent plasma, as blood banks and hospitals are seeking this antibody-rich serum to directly transfuse and treat critically ill patients, while pharmaceutical companies are collecting it to produce plasma-derived medicine such as hyperimmune globulin that can act as a cure for Covid-19 patients [2, 42]. Even though both the efficacy and safety of the treatments are still under investigation worldwide, there exist studies on patients with other infectious diseases, and severe acute viral respiratory infections, including those caused by related coronaviruses (SARS-CoV and MERS-CoV) that found therapeutic benefits of convalescent plasma [13, 35, 62, 64]. Both the non-profit and profit-making organizations competing in this market for convalescent plasma are taking measures to raise awareness, to generate confidence regarding the safety of the donation process, and to recruit donors [3, 19].

In the United States, according to the guidelines issued by the [15], individuals who have fully recovered from Covid-19 and have shown no symptoms for at least 2 weeks prior to donation are eligible to donate plasma. In addition to meeting the regular donor criteria, convalescent plasma donors need to provide documentation of prior Covid-19 diagnosis. According to Harvard Health Publishing [21], one donor can produce sufficient plasma to treat three patients. As the world continues to wait for the availability of vaccines, and more studies show promising results of convalescent plasma therapy [7, 12, 27], the demand for this product and competition among hospitals, medical facilities, and pharmaceutical companies for the limited donor pool is going to become more prominent.

2 Literature Review and Our Contributions

Since the pandemic was declared only several months ago, although for many it feels like an eternity, the research is nascent, but ongoing and vigorous. Queiroz et al. [57] presented a research agenda through a structured literature review of Covid-19-related work and supply chain research on earlier epidemics. Ivanov [24], in turn, discussed simulation-based research focused on the potential impacts on global supply chains of the Covid-19 pandemic. Nagurney [39] developed a supply chain network optimization model for perishable food in the Covid-19 pandemic, which included the critical labor resource. The model can be used to investigate the impacts of labor disruptions, due to illnesses, death, etc., on prices and product flows.

In this paper, we construct a competitive game theory network model for medical supplies inspired by the Covid-19 pandemic. It features salient characteristics of the realities of this pandemic in terms of competition among organizations/institutions for supplies under limited capacities globally as well as uncertain demands due to

the fact that so much about this novel coronavirus remains unknown and is yet to be discovered. Since the organizations, notably, healthcare ones such as hospitals and nursing homes but also medical practices, etc., compete with one another for the limited supplies, given the prices and their associated logistical costs as well as the expected loss due to possible shortages or surpluses, the model is a Generalized Nash Equilibrium (GNE) model (cf. Debreu [9]; see also Arrow and Debreu [4]) rather than a Nash equilibrium one (cf. Nash [53, 54]). To date, there have been very few GNE models in the setting of disaster relief. Here we are dealing with a global healthcare disaster on a monumental scale, which, unlike other disasters (cf. Nagurney and Qiang [46], Kotsireas et al. [31, 32]), is not limited in space and time. Furthermore, our model has stochastic elements.

We emphasize that in the case of Generalized Nash Equilibrium models not only do the objective functions of the players in the game depend on the strategies of the other players but the feasible sets do as well (see, e.g., Fischer et al. [14]). Nagurney et al. [40] constructed the first disaster relief GNE model integrating financial and logistical aspects of humanitarian organizations' activities and demonstrated that, because of the underlying functions, an optimization reformulation was possible. Subsequently, Nagurney et al. [41] generalized the results to a broader class of functions and used the concept of a Variational Equilibrium (cf. Kulkarni and Shabhang [33]), which enabled a finite-dimensional variational inequality formulation and solution procedures. However, these models were deterministic. The first stochastic GNE model for disaster relief was constructed by Nagurney et al. [48] with each humanitarian organization facing a two-stage stochastic optimization problem and with the common, that is, the shared, constraints being on the demand side and associated with relief items to be delivered to the victims at the various demand points. There were no bounds on the availability of supplies.

In this paper, in contrast, and as is vividly occurring in the Covid-19 pandemic, the supplies of the items, which in our model are medical items, are constrained. Also, the demand for the medical items is uncertain with associated penalties for shortages or surpluses, with the former expected to be much higher due to potential loss of life, increased pain and suffering, etc. The constructs that we utilize for handling the uncertain demands for medical items are based on results of [11], who introduced a supply chain equilibrium model with random demands, and on the results of [49] and [43, 44], who focused on optimization models in disaster relief and healthcare. Nagurney and Nagurney [45] developed a supply chain network model for disaster relief under cost and demand uncertainty, but again, therein, there was a single decision-maker and, hence, game theory was not needed. Mete and Zabinsky [36] introduced a two-stage stochastic optimization model for storage and distribution of medical supplies but also considered a single decision-maker. Adida et al. [1] considered hospital stockpiling of medical supplies with a focus on shortages in the system and a common population. The authors because of their assumptions could derive closed form expressions for solutions. In our model, there are multiple independent demand points and they compete for the medical item supplies with one another. Our model also includes general transportation costs, and each demand point is subject to uncertain demand for the medical supplies.

Moreover, our model is a Generalized Nash Equilibrium model and not a Nash equilibrium model.

Muggy and Heier Stamm [37] provided a review of game theory in humanitarian operations to that date, and note that there remain many unexplored modeling research opportunities. The excellent survey article of [20] on multicriteria optimization in humanitarian aid includes references to both deterministic and stochastic models. The authors, in their future research directions section, emphasize the need for papers that consider the diverging interests of multiple and sometimes competing stakeholders. Such a research gap is addressed in this paper.

This remainder of the paper is organized as follows. In Section 3, we present the Generalized Nash Equilibrium network model for medical supplies and provide alternative variational inequality formulations of the governing equilibrium conditions. In Section 4, we discuss some qualitative properties of the model as well as the function that enters the variational inequality that we utilize to solve the numerical examples in Section 5. Section 6 summarizes our results, presents our conclusions, and also gives suggestions for future research.

3 The Generalized Nash Equilibrium Network Model for Medical Supplies Under Stochastic Demand

We consider m locations that are supply locations for the medical supplies, with a typical supply point denoted by i , and n locations that are demand points, with a typical demand point denoted by j . Note that supply points can be locations in different regions, states, or even countries. Demand points are locations where the medical supplies are needed such as hospitals, nursing homes, medical clinics, prisons, etc. The bipartite structure of the game theory problem is depicted in Figure 1. The notation for the model is given in Table 1. All vectors are column vectors.

The demand for the medical item at the demand points is uncertain due to the unpredictability of the actual demand at the demand points. The literature contains examples of supply chain network models with uncertain demand and associated shortage and surplus penalties (see, e.g., Dong et al. [11], Nagurney et al. (2011), Nagurney and Masoumi [43], Nagurney et al. [44]). Nagurney and Nagurney [45] developed a model for disaster relief under cost and demand uncertainty. The probability distribution of demand for PPEs can be obtained using census data and/or information gathered during the pandemic disaster preparedness phase.

Before constructing the objective function, we present some needed preliminaries.

Since d_j denotes the actual (uncertain) demand at destination point j , we have

$$P_j(D_j) = P_j(d_j \leq D_j) = \int_0^{D_j} \mathcal{F}_j(t)dt, \quad j = 1, \dots, n, \tag{1}$$

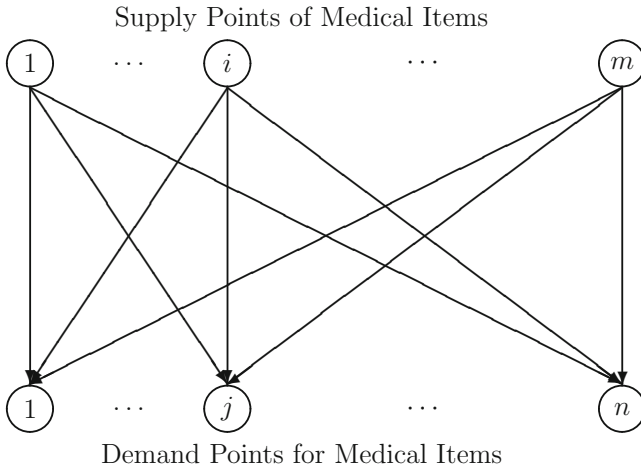


Fig. 1 The network structure of the competitive game theory model for medical supplies

where P_j and \mathcal{F}_j denote the probability distribution function and the probability density function of demand at point j , respectively.

Recall from Table 1 that v_j is the “projected demand” for the medical item at demand point j ; $j = 1, \dots, n$. The amounts of shortage and surplus at demand point j are calculated, respectively, according to:

$$\Delta_j^- \equiv \max\{0, d_j - v_j\}, \quad j = 1, \dots, n, \tag{2a}$$

$$\Delta_j^+ \equiv \max\{0, v_j - d_j\}, \quad j = 1, \dots, n. \tag{2b}$$

The expected values of shortage and surplus at each demand point are, hence:

$$E(\Delta_j^-) = \int_{v_j}^{\infty} (t - v_j) \mathcal{F}_j(t) dt, \quad j = 1, \dots, n, \tag{3a}$$

$$E(\Delta_j^+) = \int_0^{v_j} (v_j - t) \mathcal{F}_j(t) dt, \quad j = 1, \dots, n. \tag{3b}$$

The expected penalty incurred by demand point j due to the shortage and surplus of the medical item is equal to:

$$E(\lambda_j^- \Delta_j^- + \lambda_j^+ \Delta_j^+) = \lambda_j^- E(\Delta_j^-) + \lambda_j^+ E(\Delta_j^+), \quad j = 1, \dots, n. \tag{4}$$

We assume that $\lambda_j^+ + \lambda_j^-$ is greater than zero, for each demand point j .

The projected demand at demand point j , v_j , is equal to the sum of flows of the medical item to j , that is:

Table 1 Notation for the Generalized Nash Equilibrium network model

Notation	Definition
q_{ij}	the amount of the medical item purchased from supply location i by j . We first group all the i elements $\{q_{ij}\}$ into the vector q_j , and then we group such vectors for all j into the vector $q \in R_+^{mn}$
v_j	the projected demand at demand point j ; $j = 1, \dots, n$
d_j	the actual (uncertain) demand for the medical item at demand location j ; $j = 1, \dots, n$
Δ_j^-	the amount of shortage of the medical item at demand point j ; $j = 1, \dots, n$
Δ_j^+	the amount of surplus of the medical item at demand point j ; $j = 1, \dots, n$
λ_j^-	the unit penalty associated with a shortage of the medical item at demand point j ; $j = 1, \dots, n$
λ_j^+	the unit penalty associated with a surplus of the medical item at demand point j ; $j = 1, \dots, n$
ρ_i	the price of the medical item at supply location i ; $i = 1, \dots, m$
$c_{ij}(q)$	the generalized cost of transportation associated with transporting the medical item from supply location i to demand location j , which includes the financial cost, any tariffs/taxes, time, and risk. We group all the generalized costs into the vector $c(q) \in R^{mn}$
S_i	the nonnegative amount of the medical item available for purchase at supply location i ; $i = 1, \dots, m$
μ_i	the nonnegative Lagrange multiplier associated with the supply constraint at supply location i . We group the Lagrange multipliers into the vector $\mu \in R_+^m$

$$v_j \equiv \sum_{i=1}^m q_{ij}, \quad j = 1, \dots, n. \tag{5}$$

Each demand location j seeks to minimize the total costs associated with the purchasing of the medical item plus the total cost of transportation plus the expected cost due to a shortage or surplus at j .

The objective function of each demand point j is, hence, given by

$$\text{Minimize} \quad \sum_{i=1}^m \rho_i q_{ij} + \sum_{i=1}^m c_{ij}(q) + \lambda_j^- E(\Delta_j^-) + \lambda_j^+ E(\Delta_j^+) \tag{6}$$

subject to the following constraints:

$$\sum_{j=1}^n q_{ij} \leq S_i, \quad i = 1, \dots, m, \tag{7}$$

$$q_{ij} \geq 0, \quad i = 1, \dots, m. \tag{8}$$

The first term in the objective function (6) represents the purchasing costs, whereas the second term represents the generalized total transportation costs. The third term in (6) captures the expected cost due to shortage or surplus of the medical items at the demand point of the organization. We expect that the weight λ_j^- would be significantly higher than the weight λ_j^+ for each j since a shortage of the medical items can result in greater suffering and loss of life.

The constraint (7) represents common, that is, a shared, constraint in that the demand locations compete for the medical items that are available for purchase at the supply locations at a maximum available supply. The constraints in (8) are the nonnegativity assumption on the medical item purchase volumes.

We assume that the total generalized transportation cost functions are continuously differentiable and convex. Note that, in our model, the transportation costs can, in general, depend upon the vector of medical item flows since there is competition for freight service provision in the pandemic.

We now present some preliminaries that allow us to express the partial derivatives of the expected total shortage and discarding costs of the medical items at the demand points only in terms of the medical item flow variables. We then prove that the third term in the Objective Function (6) is also convex.

Note that for each demand point j :

$$\frac{\partial E(\Delta_j^-)}{\partial q_{ij}} = \frac{\partial E(\Delta_j^-)}{\partial v_j} \times \frac{\partial v_j}{\partial q_{ij}}, \quad \forall i. \tag{9}$$

By Leibniz’s integral rule, we have

$$\begin{aligned} \frac{\partial E(\Delta_j^-)}{\partial v_j} &= \frac{\partial}{\partial v_j} \left(\int_{v_j}^{\infty} (t - v_j) \mathcal{F}_j(t) d(t) \right) = \int_{v_j}^{\infty} \frac{\partial}{\partial v_j} (t - v_j) \mathcal{F}_j(t) d(t) \\ &= P_j(v_j) - 1, \quad j = 1, \dots, n. \end{aligned} \tag{10a}$$

Therefore,

$$\frac{\partial E(\Delta_j^-)}{\partial v_j} = P_j \left(\sum_{i=1}^m q_{ij} \right) - 1, \quad j = 1, \dots, n. \tag{10b}$$

On the other hand, we have:

$$\frac{\partial v_j}{\partial q_{ij}} = \frac{\partial}{\partial q_{ij}} \sum_{l=1}^m q_{lj} = 1, \quad \forall i; j = 1, \dots, n. \tag{11}$$

Therefore, from (10b) and (11), we conclude that

$$\frac{\partial E(\Delta_j^-)}{\partial q_{ij}} = \left[P_j \left(\sum_{i=1}^m q_{ij} \right) - 1 \right], \quad \forall i; j = 1, \dots, n. \tag{12}$$

Analogously, for the surplus, we have

$$\frac{\partial E(\Delta_j^+)}{\partial q_{ij}} = \frac{\partial E(\Delta_j^+)}{\partial v_j} \times \frac{\partial v_j}{\partial q_{ij}}, \quad \forall i; j = 1, \dots, n, \tag{13}$$

$$\begin{aligned} \frac{\partial E(\Delta_j^+)}{\partial v_j} &= \frac{\partial}{\partial v_j} \left(\int_0^{v_j} (v_j - t) \mathcal{F}_j(t) d(t) \right) = \int_0^{v_j} \frac{\partial}{\partial v_j} (v_j - t) \mathcal{F}_j(t) d(t) \\ &= P_j(v_j), \quad j = 1, \dots, n. \end{aligned} \tag{14a}$$

Thus,

$$\frac{\partial E(\Delta_j^+)}{\partial v_j} = P_j \left(\sum_{i=1}^m q_{ij} \right), \quad j = 1, \dots, n. \tag{14b}$$

From (14b) and (11), we have

$$\frac{\partial E(\Delta_j^+)}{\partial q_{ij}} = P_j \left(\sum_{i=1}^m q_{ij} \right), \quad \forall i; j = 1, \dots, n. \tag{15}$$

Lemma 1 *The expected shortage and surplus cost function $\lambda_j^- E(\Delta_j^-) + \lambda_j^+ E(\Delta_j^+)$ is convex.*

Proof We have

$$\frac{\partial^2}{\partial q_{ij}^2} \left[\lambda_j^- E(\Delta_j^-) + \lambda_j^+ E(\Delta_j^+) \right] = \lambda_j^- \frac{\partial^2 E(\Delta_j^-)}{\partial q_{ij}^2} + \lambda_j^+ \frac{\partial^2 E(\Delta_j^+)}{\partial q_{ij}^2}, \quad \forall i; j = 1, \dots, n. \tag{16a}$$

Substituting the first order derivatives from (12) and (15) into (16a) yields

$$\begin{aligned} \frac{\partial^2}{\partial q_{ij}^2} \left[\lambda_j^- E(\Delta_j^-) + \lambda_j^+ E(\Delta_j^+) \right] &= \lambda_j^- \frac{\partial}{\partial q_{ij}} \left[P_j \left(\sum_{i=1}^m q_{ij} \right) - 1 \right] + \lambda_j^+ \frac{\partial}{\partial q_{ij}} P_j \left(\sum_{i=1}^m q_{ij} \right) \\ &= (\lambda_j^- + \lambda_j^+) \mathcal{F}_j \left(\sum_{i=1}^m q_{ij} \right) \geq 0, \quad \forall i; j = 1, \dots, n. \end{aligned} \tag{16b}$$

The above inequality holds provided that $(\lambda_j^- + \lambda_j^+)$, i.e., the sum of shortage and surplus penalties, is positive. Hence, $\lambda_j^- E(\Delta_j^-) + \lambda_j^+ E(\Delta_j^+)$, and, as a consequence, the objective function in (6) are also convex. \square

We refer to the objective function (6) for j as the disutility of j and denote it by $DU_j(q)$; $j = 1, \dots, n$.

We define the feasible sets $K_j \equiv \{q_j \geq 0\}$; $j = 1, \dots, n$. We define $K \equiv \prod_{i=1}^I K_i$. We also define the feasible set $\mathcal{S} \equiv \{q|q \text{ satisfying (7)}\}$, which consist of the shared constraints.

Definition 1 (Generalized Nash Equilibrium for Medical Items) A vector of medical items $q^* \in K \cap \mathcal{S}$ is a Generalized Nash Equilibrium if for each demand point j ; $j = 1, \dots, n$:

$$DU_j(q_j^*, \hat{q}_j^*) \leq DU_j(q_j, \hat{q}_j^*), \quad \forall q_j \in K_j \cap \mathcal{S}, \tag{17}$$

where $\hat{q}_j^* \equiv (q_1^*, \dots, q_{j-1}^*, q_{j+1}^*, \dots, q_n^*)$.

According to (17), an equilibrium is established if no demand point has any incentive to unilaterally change its vector of medical item purchases/shipments. Observe that in our model not only does the objective function of a demand point depend not only on the vector of strategies of its own strategies and on those of the other demand points, but the feasible set does as well. Hence, this model is not a Nash [53, 54] model, but, rather, it is a Generalized Nash Equilibrium model. Our model captures the reality of the intense competitive landscape in the Covid-19 pandemic.

Here, we utilize the concept of a *Variational Equilibrium*, which allows us to formulate the above GNE conditions as the solution to a finite-dimensional variational inequality problem. Hence, rather than making use of quasi-variational inequalities, for which the algorithms are not as advanced, we can apply variational inequality algorithms to solve numerically the model. Indeed, as emphasized in Nagurney et al. [50], Nagurney, et al. [47], and Nagurney et al. [39], we can define a Variational Equilibrium, which is a refinement and a specific type of GNE (cf. Kulkarni and Shabhang [33]) that enables a variational inequality formulation.

We define the feasible set $\mathcal{K} \equiv K \cap \mathcal{S}$.

Definition 2 (Variational Equilibrium) A vector of medical items $q^* \in \mathcal{K}$ is a Variational Equilibrium of the above Generalized Nash Equilibrium problem if it is a solution to the following variational inequality:

$$\sum_{j=1}^n \sum_{i=1}^m \frac{\partial DU_j(q^*)}{q_{ij}} \times (q_{ij} - q_{ij}^*) \geq 0, \quad \forall q \in \mathcal{K}, \tag{18}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in mn -dimensional Euclidean space.

In expanded form, the variational inequality in (18) is: determine $q^* \in \mathcal{K}$ such that

$$\sum_{j=1}^n \sum_{i=1}^m \left[\rho_i + \sum_{l=1}^m \frac{\partial c_{lj}(q^*)}{\partial q_{ij}} + \lambda_j^+ P_j(\sum_{l=1}^m q_{lj}^*) - \lambda_j^- (1 - P_j(\sum_{l=1}^m q_{lj}^*)) \right] \times [q_{ij} - q_{ij}^*] \geq 0, \quad \forall q \in \mathcal{H}. \tag{19}$$

Note that the variational equilibrium guarantees that the Lagrange multipliers associated with the common constraints are the same for all the demand points. This feature yields an elegant fairness and equity interpretation, which is very relevant during this pandemic.

We now put variational inequality (19) into standard form. Recall (cf. Nagurney [38]) that the finite-dimensional variational inequality problem, VI(F, \mathcal{H}), is to determine a vector $X^* \in \mathcal{H} \subset R^N$, such that

$$\langle F(X^*), X - X^* \rangle \geq 0, \quad \forall X \in \mathcal{H}, \tag{20}$$

where F is a given continuous function from \mathcal{H} to R^N , and \mathcal{H} is a given closed, convex set.

We let $X \equiv q$ and $F(X)$ be the vector with elements: $\{\frac{\partial DU_j(q^*)}{\partial q_{ij}}\}, \forall j, i$ with \mathcal{H} as originally defined and $N = mn$. Then, clearly, variational inequality (19) can be put into standard form (20), under our assumptions.

Also it is worth noting that the existence of a solution q^* to variational inequality (19) is guaranteed under the classical theory (see Kinderlehrer and Stampacchia [29]) since the function that enters the variational inequality is continuous and the feasible set \mathcal{H} is not only convex but also compact because the supplies of the medical items are bounded. Hence, the following theorem is immediate.

Theorem 1 (Existence) *A solution to variational inequality (19) exists.*

We now provide an alternative variational inequality to (18) (and (19)). We associate a nonnegative Lagrange multiplier μ_i with constraint (7), for each supply location $i = 1, \dots, m$. We group all the Lagrange multipliers into the vector $\mu \in R_+^m$. We define the feasible set $\mathcal{H}^2 \equiv \{(q, \mu) | q \geq 0, \mu \geq 0\}$.

Then, using arguments as in [47], an alternative variational inequality for (19) is: determine $(q^*, \mu^*) \in \mathcal{H}^2$ such that

$$\begin{aligned} & \sum_{j=1}^n \sum_{i=1}^m \left[\rho_i + \sum_{l=1}^m \frac{\partial c_{lj}(q^*)}{\partial q_{ij}} + \lambda_j^+ P_j(\sum_{l=1}^m q_{lj}^*) - \lambda_j^- (1 - P_j(\sum_{l=1}^m q_{lj}^*)) + \mu_i^* \right] \times [q_{ij} - q_{ij}^*] \\ & + \sum_{i=1}^m \left[S_i - \sum_{j=1}^n q_{ij}^* \right] \times [\mu_i - \mu_i^*] \geq 0, \quad \forall (q, \mu) \in \mathcal{H}^2. \end{aligned} \tag{21}$$

Variational inequality (21) can also be put into standard form (20) if we define $X \equiv (q, \mu)$ and $F(X) \equiv (F^1(X), F^2(X))$, where $F^1(X)$ has as its (i, j) -th

component: $\rho_i + \sum_{l=1}^m \frac{\partial c_{lj}(q)}{\partial q_{ij}} + \lambda_j^+ P_j(\sum_{l=1}^m q_{lj}) - \lambda_j^-(1 - P_j(\sum_{l=1}^m q_{lj})) + \mu_i$; $i = 1, \dots, m$; $j = 1, \dots, n$, and the i -th component of $F^2(X)$ is $S_i - \sum_{j=1}^n q_{ij}$, for $i = 1, \dots, m$. Furthermore, $\mathcal{K} \equiv \mathcal{K}^2$ and $N = mn + m$.

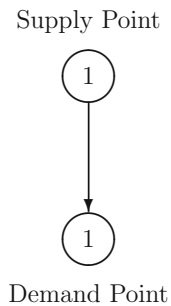
3.1 Illustrative Examples

In this subsection, we present three small numerical examples for illustrative purposes. These examples are inspired by the Covid-19 pandemic and the associated challenges in procuring N95 face masks, which are among the most needed medical products in dealing with this healthcare disaster. We emphasize that the equilibrium Lagrange multipliers provide valuable information since they represent the shadow prices of the supply constraints. In particular, if an equilibrium Lagrange multiplier is positive, then this is the amount of the cost (or the loss) that could be saved with an extra unit of the supply of the medical item.

Illustrative Example 1 (One Supply Point and One Demand Point) In this example, there is a single supply point and a single demand point, as depicted in Figure 2.

The supply point sells 20-pack N95 masks in the form of large bulks of 1000 packs each; therefore, one unit of item flow from the supply point to a demand point, q_{ij} , represents 1000 of 20-pack N95 masks. The demand at the demand point is uniformly distributed between 100 and 1000 of large bulks. To determine the price of a unit item flow, ρ_i at supply point i , we assume that the price of each 20-pack N95 mask during the pandemic is \$25, so that the purchase price of each large bulk is $\rho_1 = 25,000$. Although a face mask is not, under normal conditions, an expensive product, it has been proved to be essential in reducing the spread of the virus. Based on this, we assume that, for every 2000 people who do not use the face mask, one person would die due to the disease. Although it is not easy to value people’s lives, we assume a \$200,000 equivalent for each loss. As a result,

Fig. 2 Network topology for illustrative Example 1



the penalty, λ_1^- , on the shortage of one item flow, which is equivalent to 20,000 N95 masks, is set at \$2,000,000. Also, since the supply chain has been severely disrupted at the time of the declaration of the pandemic, overloading can cause many problems in transportation and processing at entry points for countries. To prevent this, we also consider a penalty of $\lambda_1^+ = 100,000$ on surplus item flows. The data for this example is as follows:

$$\rho_1 = 25,000, \quad S_1 = 1000, \quad c_{11}(q) = q_{11}^2 + 3q_{11}, \quad \lambda_1^- = 2,000,000, \quad \lambda_1^+ = 100,000.$$

We can rewrite variational inequality (21) for this example as: determine $(q^*, \mu^*) \in \mathcal{X}^2$ such that:

$$\left[25,000 + 2q_{11}^* + 3 + 100,000\left(\frac{q_{11}^* - 1000}{900}\right) - 2,000,000\left(\frac{1000 - q_{11}^*}{900}\right) + \mu_1^* \right] \\ \times [q_{11} - q_{11}^*] + [1000 - q_{11}^*] \times [\mu_1 - \mu_1^*] \geq 0, \quad \forall (q, \mu) \in \mathcal{X}^2$$

The solution to the above variational inequality, which we obtained analytically, is

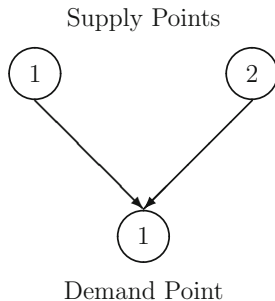
$$q_{11}^* = 945.62, \quad \mu_1^* = 0.00.$$

Observe that the organization at the demand point procures a huge number of masks because of the great importance of PPEs in preventing the further spread of the virus and the potential damage that could be caused by an insufficient number of N95 face masks. The projected demand value $v_1 = 945.62$ lies between the lower and the upper bounds of the uniform distribution range. Note that the projected demand is very close to the upper bound. The decision-makers at the organization at the demand point are aware of the importance of the masks and have assigned a much larger penalty on a shortage as compared to the surplus penalty. The disutility of the organization in this logistical operation is equal to 67,543,534.04.

Illustrative Example 2 (Two Supply Points and One Demand Point) In the second illustrative example, a new supply point has been added to the supply chain network, as depicted in Figure 3.

Hence, now, the decision-makers at the demand point have two options for procuring the face masks. The new supply point offers masks for less than half the price of the other supply point, but its supply capacity is half that of the previous one. Also, the generalized transportation cost rate from the origin of the N95 masks of the new supply point to the demand point is higher than the rate of the other supply point. The data on the new supply point is as follows:

Fig. 3 Network topology for illustrative Example 2



$$\rho_2 = 10,000, \quad S_2 = 500, \quad c_{21}(q) = 2q_{21}^2 + 4q_{21}.$$

Variational inequality (21) can be rewritten as follows for this example: determine $(q^*, \mu^*) \in \mathcal{K}^2$ such that

$$\begin{aligned} & \left[25,000 + 2q_{11}^* + 3 + 100,000\left(\frac{q_{11}^* + q_{21}^* - 100}{900}\right) - 2,000,000\left(\frac{1000 - q_{11}^* - q_{21}^*}{900}\right) + \mu_1^* \right] \times [q_{11} - q_{11}^*] \\ & + \left[10,000 + 4q_{21}^* + 4 + 100,000\left(\frac{q_{11}^* + q_{21}^* - 100}{900}\right) - 2,000,000\left(\frac{1000 - q_{11}^* - q_{21}^*}{900}\right) + \mu_2^* \right] \times [q_{21} - q_{21}^*] \\ & + [1000 - q_{11}^*] \times [\mu_1 - \mu_1^*] + [500 - q_{21}^*] \times [\mu_2 - \mu_2^*] \geq 0, \quad \forall (q, \mu) \in \mathcal{K}^2. \end{aligned}$$

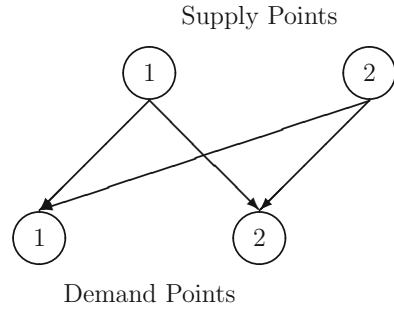
The solution to the above variational inequality, obtained analytically, is

$$q_{11}^* = 446.05, \quad q_{21}^* = 500.00, \quad \mu_1^* = 0.00, \quad \mu_2^* = 13891.80.$$

Observe that, with the addition of a new supply point, the decision-makers' strategy has changed. Since the price of the product offered by the new supply point is much lower than that at the first supply point, the decision-makers purchase more items from supply point 2, despite the fact that the generalized transportation cost to the demand point from supply point 2 is higher than that from supply point 1. However, the supply capacity of the new supply point is half that of the first supply point, and we see that all its capacity has been used. Therefore, the associated equilibrium Lagrange multiplier is positive. Again, the projected demand falls between the lower and the upper bounds of the uniform distribution and is closer to the upper bound for the same reason as in the previous example. But, now, with greater flexibility in the supply chain due to the addition of a new supply point, the disutility of the organization at the demand point has declined, dropping to 59,860,548.75.

Illustrative Example 3 (Two Supply Points and Two Demand Points) This example is constructed from the previous examples, with the difference that now

Fig. 4 Network topology for illustrative Example 3



there are two demand points trying to procure N95 masks and competing over limited supplies; see Figure 4.

The demand for the new demand point is uniformly distributed between 100 and 500. The generalized transportation cost functions and the penalty coefficients associated with the second demand point are

$$c_{12}(q) = 2q_{12}^2 + 3q_{12}, \quad c_{22}(q) = 3q_{22}^2 + 4q_{22}, \quad \lambda_2^- = 2,000,000, \quad \lambda_2^+ = 100,000.$$

Variational inequality (21) for this example is as below: determine $(q^*, \mu^*) \in \mathcal{K}^2$ such that

$$\begin{aligned} & \left[25,000 + 2q_{11}^* + 3 + 100,000\left(\frac{q_{11}^* + q_{21}^* - 100}{900}\right) - 2,000,000\left(\frac{1000 - q_{11}^* - q_{21}^*}{900}\right) + \mu_1^* \right] \times [q_{11} - q_{11}^*] \\ & + \left[10,000 + 4q_{21}^* + 4 + 100,000\left(\frac{q_{11}^* + q_{21}^* - 100}{900}\right) - 2,000,000\left(\frac{1000 - q_{11}^* - q_{21}^*}{900}\right) + \mu_2^* \right] \times [q_{21} - q_{21}^*] \\ & + \left[25,000 + 4q_{12}^* + 3 + 100,000\left(\frac{q_{12}^* + q_{22}^* - 100}{400}\right) - 2,000,000\left(\frac{500 - q_{12}^* - q_{22}^*}{400}\right) + \mu_1^* \right] \times [q_{12} - q_{12}^*] \\ & + \left[10,000 + 6q_{22}^* + 4 + 100,000\left(\frac{q_{12}^* + q_{22}^* - 100}{400}\right) - 2,000,000\left(\frac{500 - q_{12}^* - q_{22}^*}{400}\right) + \mu_2^* \right] \times [q_{22} - q_{22}^*] \\ & + [1000 - q_{11}^*] \times [\mu_1 - \mu_1^*] + [500 - q_{21}^*] \times [\mu_2 - \mu_2^*] \geq 0, \quad \forall (q, \mu) \in \mathcal{K}^2. \end{aligned}$$

The solution to this variational inequality, again, obtained analytically, is

$$q_{11}^* = 634.14, \quad q_{21}^* = 311.74, \quad q_{12}^* = 287.71, \quad q_{22}^* = 188.26, \quad \mu_1^* = 0.00, \quad \mu_2^* = 15020.30.$$

With the addition of another demand point, there is increased competition for the valuable N95 masks. The strategies of the organization at demand point 1 have changed as compared to the previous example. It can be seen that the full capacity

of supply point 2 has not been assigned to demand point 1, since the organization at demand point 1 now competed with the organization at demand point 2. As a result, the major part of the demand point 1’s procurement of the N95 masks is from supply point 1 that has a larger capacity as compared to supply point 2. And, similar to the previous example, the equilibrium Lagrange multiplier associated with the supply capacity of supply point 2 is positive since it has sold all its available supply of N95 masks, while the other supply point has not exhausted its capacity. Both demand points receive a large amount of face masks and their projected demands lie in their respective uniform probability distribution range. Both projected demands are closer to the upper bound since the penalty on shortage is much higher than the penalty on surplus. The addition of a new demand point to the competition has changed the strategies of the organization at demand point 1, and we can see the impact on its disutility. Its disutility has now increased to 62,580,546.57. The disutility of the second demand point is 28,457,845.74.

4 Qualitative Properties and the Algorithm

We now discuss some properties of the model, specifically, those that guarantee that the conditions for convergence of the modified projection method (cf. Korpelevich [30] and Nagurney [38]) that we use to compute solutions to numerical examples in this next section are met. The algorithm is guaranteed to converge to a solution of variational inequality (21) if the function $F(X)$ that enters the variational inequality is monotone and Lipschitz continuous and that a solution exists. It was recently applied to compute solutions to a stochastic game theory model for disaster relief by Nagurney [39].

Recall that the function $F(X)$ is said to be monotone, if

$$\langle F(X^1) - F(X^2), X^1 - X^2 \rangle \geq 0, \quad \forall X^1, X^2 \in \mathcal{X}. \tag{22}$$

Theorem 2 (Monotonicity) *The function $F(X)$ is monotone, for all $X \in \mathcal{X}$, if all the generalized transportation cost functions c_{ij} , $i = 1, \dots, m$; $j = 1, \dots, n$, are convex.*

Proof $\forall X^1, X^2 \in \mathcal{X}$, let $v_j^1 = \sum_{i=1}^m q_{ij}^1$ and $v_j^2 = \sum_{i=1}^m q_{ij}^2$.

$$\begin{aligned} & \langle F(X^1) - F(X^2), X^1 - X^2 \rangle \\ &= \sum_{j=1}^n \sum_{i=1}^m \left[\sum_{l=1}^m \frac{\partial c_{lj}(q^1)}{\partial q_{ij}} - \sum_{l=1}^m \frac{\partial c_{lj}(q^2)}{\partial q_{ij}} \right] \times (q_{ij}^1 - q_{ij}^2) \end{aligned} \tag{23}$$

$$+ \sum_{j=1}^n (\lambda_j^+ + \lambda_j^-) \times (P_j(v_j^1) - P_j(v_j^2)) \times (v_j^1 - v_j^2). \tag{24}$$

Given the convexity of the generalized transportation cost functions, equation (23) is greater or equal to zero. Since a probability function $P_j, \forall j$, is an increasing function, the expression in equation (24) is greater than or equal to zero. Hence, $F(X)$ is monotone. \square

If the conditions in Theorem 1 are slightly strengthened so that the vector function that enters into the variational inequality problem (21) is strictly monotone, then its solution is unique (see, e.g., Nagurney [38]).

Theorem 3 (Uniqueness) *The function $F(X)$ is strictly monotone for all $X \in \mathcal{X}$, if all the generalized transportation cost functions $c_{ij}; i = 1, \dots, m; j = 1, \dots, n$, are strictly convex. Then the variational inequality (21) has a unique solution in \mathcal{X} .*

Theorem 4 (Lipschitz Continuity) *If the generalized transportation cost functions c_{ij} , for all i and j , have bounded second order partial derivatives, then the function $F(X)$ that enters the variational inequality problem (21) is Lipschitz continuous; that is, there exists a constant $L > 0$, known as the Lipschitz constant, such that*

$$\|F(X^1) - F(X^2)\| \leq L\|X^1 - X^2\|, \quad \forall X^1, X^2 \in \mathcal{X}. \tag{25}$$

Proof Since each probability function $P_j; j = 1, \dots, n$, is always less than or equal to 1, the result is direct by applying a mid-value theorem from calculus to the vector function $F(X)$ that enters the variational inequality problem (21). See also Nagurney and Zhang [51] and Nagurney [38]. \square

The iterative steps of the modified projection method, with τ denoting an iteration counter, are as follows:

The Modified Projection Method

Step 0: Initialization

Initialize with $X^0 \in \mathcal{X}$. Set the iteration counter $\tau := 1$ and let β be a scalar such that $0 < \beta \leq \frac{1}{L}$, where L is the Lipschitz constant.

Step 1: Computation

Compute \bar{X}^τ by solving the variational inequality subproblem:

$$\langle \bar{X}^\tau + \beta F(X^{\tau-1}) - X^{\tau-1}, X - \bar{X}^\tau \rangle \geq 0, \quad \forall X \in \mathcal{X}. \tag{26}$$

Step 2: Adaptation

Compute X^τ by solving the variational inequality subproblem:

$$\langle X^\tau + \beta F(\bar{X}^\tau) - X^{\tau-1}, X - X^\tau \rangle \geq 0, \quad \forall X \in \mathcal{X}. \tag{27}$$

Step 3: Convergence Verification

If $|X^\tau - X^{\tau-1}| \leq \epsilon$, with $\epsilon > 0$, a pre-specified tolerance, then stop; otherwise, set $\tau := \tau + 1$ and go to Step 1.

The modified projection method for the model governed by variational inequality (21) yields closed form expressions for the medical item flows and for the Lagrange multipliers in both Steps (26) and (27). This is a nice feature for computer implementation.

Theorem 5 (Convergence) *Assume that the function that enters the variational inequality (21) (or (19)) has at least one solution and all the generalized transportation cost functions are convex, then the modified projection method described above converges to the solution of the variational inequality (21) (or (19)).*

Proof According to Korpelevich [30], the modified projection method converges to the solution of the variational inequality problem of the form (20), provided that the function F that enters the variational inequality is monotone and Lipschitz continuous and that a solution exists. Existence of a solution follows from Theorem 1. Monotonicity follows Theorem 2. Lipschitz continuity, in turn, follows from Theorem 4. □

We now provide the explicit formulae for the medical item flows and the Lagrange multipliers at iteration τ for Step 1. The analogues for Step 2 can be easily derived accordingly.

Specifically, we have:

Explicit Formula for the Medical Item Flow for Each i, j at Iteration τ of Step 1

Determine \bar{q}_{ij}^τ for each i, j at Step 1 iteration τ according to:

$$\bar{q}_{ij}^\tau = \max\{0, q_{ij}^{\tau-1} + \beta(-\rho_i - \sum_{l=1}^m \frac{\partial c_{lj}(q^{\tau-1})}{\partial q_{ij}} - \lambda_j^+ P_j(\sum_{l=1}^m q_{lj}^{\tau-1}) + \lambda_j^- (1 - P_j(\sum_{l=1}^m q_{lj}^{\tau-1})) - \mu_i^{\tau-1})\}. \tag{28}$$

Explicit Formula for the Lagrange Multiplier for Each i at Iteration τ of Step 1

Determine $\bar{\mu}_i^\tau$ for each i at Step 1 iteration τ according to:

$$\bar{\mu}_i^\tau = \max\{0, \mu_i^{\tau-1} + \beta(-S_i + \sum_{j=1}^n q_{ij}^{\tau-1})\}. \tag{29}$$

5 Numerical Examples

In this section, we apply the modified projection method to compute solutions to numerical examples. The algorithm was implemented in FORTRAN, and the computer system used was a Linux system at the University of Massachusetts Amherst. We initialized the algorithm by setting all the medical item flows and the Lagrange multipliers to 0.00. The convergence condition for all the examples was that the absolute value of two successive variable iterates was less than or equal to 10^{-8} . The β parameter in the modified projection method was set to: 0.1.

The examples are of increasing complexity. We report all the input and the output data for transparency purposes and reproducibility.

In this section, we focus on procurement of N95 masks but in the scenario of increasing demand will among smaller healthcare organizations in the form of medical practices. With the pandemic in the USA continuing in the summer of 2020 and with the opening of schools and universities to a certain degree on the horizon, there are increased pressures on the procurement of PPEs. In particular, we reference the following news article by O’Connell [55]; see also Wan [63].

Numerical Example 1 (One Supply Point and One Demand Point) In the first numerical example, for which we computed the solution using the code that we implemented, there is a single supply point and a single demand point as in the network in Figure 2. The q_{ij} s are in units since these medical practices are small relative to hospitals, etc. We assumed a uniform probability distribution in the range [100, 1000] at the demand point. The additional data for this example are

$$\rho_1 = 2, \quad S_1 = 1000, \quad c_{11}(q) = .005q_{11}^2 + .01q_{11}, \quad \lambda_1^- = 1000, \quad \lambda_1^+ = 10.$$

The computed equilibrium solution is

$$q_{11}^* = 980.56, \quad \mu_1^* = 0.00.$$

The projected demand of 980.56 is close to the upper bound of the probability distribution at the demand point.

Numerical Example 2 (One Supply Point and Two Demand Points) This example has the same data as those in Numerical Example 1 except for added data for the second demand point. The network topology is as in Figure 5.

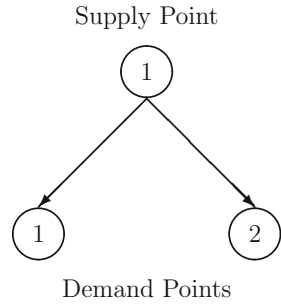
The probability distribution at the second demand point had the same lower and upper bounds as in the first demand point.

This example has the same data as Numerical Example 1 except for the following additional data for the new demand point:

$$c_{12}(q) = .01q_{12}^2 + .02, \quad \lambda_2^- = 1000, \quad \lambda_2^+ = 10.$$

The network topology for this example is as in Figure 5.

Fig. 5 Network topology for Numerical Example 2



The modified projection method converged to the following equilibrium solution:

$$q_{11}^* = 502.20, \quad q_{12}^* = 497.80, \quad \mu_1^* = 541.61.$$

With increased competition for N95 mask supplies from the second demand point, the first demand point has a large reduction in procured supplies, as compared to the volume received in Numerical Example 1. The available supply of 1000 N95 masks is exhausted between the two demand points, and, hence, the associated Lagrange multiplier μ_1^* is positive. The equilibrium conditions hold with excellent accuracy.

Numerical Example 3 (Two Supply Points and Two Demand Points) In Numerical Example 3, we considered the impacts of the addition of a second supply point to Numerical Example 2. The topology was as in Figure 4. Hence, the data are as above with the following additions:

$$S_2 = 500, \quad \rho_2 = 3, \quad c_{21}(q) = .015q_{21}^2 + .03, \quad c_{22}(q) = .02q_{22}^2 + .04q_{22}.$$

The modified projection method yielded the following equilibrium solution:

$$q_{11}^* = 526.31, \quad q_{12}^* = 473.69, \quad q_{21}^* = 225.57, \quad q_{22}^* = 274.43, \quad \mu_1^* = 261.17, \quad \mu_2^* = 258.65.$$

With the addition of a new supply point, both demand points gain significantly in terms of the volume of N95 that each procures and the supplies at each supply point are fully sold out. As a result, both equilibrium Lagrange multipliers are positive.

Numerical Example 4 (Two Supply Points and Three Demand Points) Numerical Example 4 was constructed from Numerical Example 3 with demand point 3 added, as in Figure 6.

Numerical Example 4 has the same data as Numerical Example 3 but with the addition of data for demand point 3 as follows:

$$c_{13}(q) = .01q_{13}^2 + .02q_{13}, \quad c_{23}(q) = .015q_{23}^2 + .03q_{23}, \quad \lambda_3^- = 1000, \quad \lambda_3^+ = 10.$$

Fig. 6 Network topology for numerical Example 4

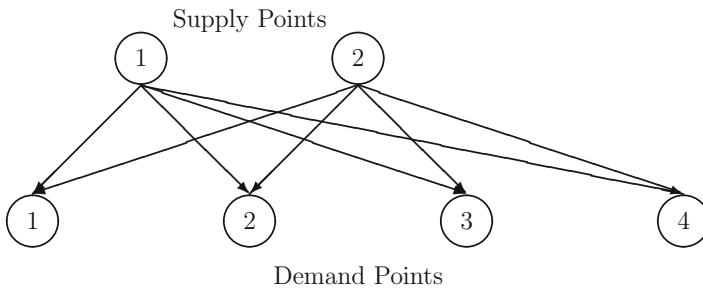
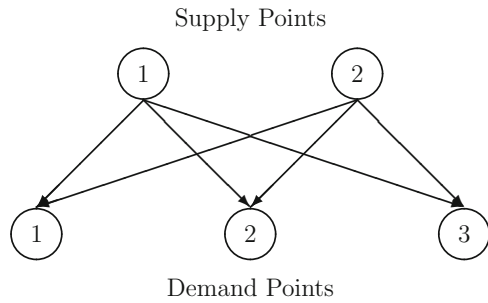


Fig. 7 Network topology for Numerical Example 5

The probability distribution for the N95 masks associated with demand point 3 is uniform with a lower bound of 200 and an upper bound of 1000.

The modified projection method yielded the following equilibrium solution:

$$q_{11}^* = 360.11, \quad q_{12}^* = 318.83, \quad q_{13}^* = 321.06,$$

$$q_{21}^* = 122.29, \quad q_{22}^* = 161.10, \quad q_{23}^* = 216.62, \quad \mu_1^* = 565.25, \quad \mu_2^* = 564.16.$$

Observe that with increasing competition for the N95 masks with another demand point, both demand points 1 and 2 experience decreases in procurement of supplies. The two supply points again fully sell out of their N95 masks, and the associated equilibrium Lagrange multipliers are both positive.

Numerical Example 5 (Two Supply Points and Four Demand Points) In the final example, Numerical Example 5, we consider yet another demand point in addition to the demand points in Numerical Example 4. Please refer to Figure 7. Smaller medical practices are increasingly concerned about being able to secure the much needed PPEs to protect the health of their employees and the viability of their practices.

The data for this example is as the data for Numerical Example 4, and the probability distribution structure for the demand at demand point is the same, with the following additional data for the new demand point 4:

$$c_{14}(q) = .015q_{14}^2 + .03q_{14}, \quad c_{24}(q) = .025q_{24}^2 + .05q_{24}, \quad \lambda_4^- = 1000, \quad \lambda_4^+ = 10.$$

The modified projection method now yielded the following equilibrium solution:

$$q_{11}^* = 260.73, \quad q_{12}^* = 229.36, \quad q_{13}^* = 251.22, \quad q_{14}^* = 258.69,$$

$$q_{21}^* = 79.57, \quad q_{22}^* = 109.17, \quad q_{23}^* = 160.46, \quad q_{24}^* = 150.81, \quad \mu_1^* = 725.71, \quad \mu_2^* = 724.91.$$

Again, the equilibrium conditions hold with excellent accuracy for this example, as was the case for all the other numerical example computed solutions. The suppliers of the N95 sell out their supplies. However, the demand points lose in terms of supply procurement for their organizations with the increased demand and competition from and yet another demand point.

We emphasize that although the above numerical examples are stylized, our mathematical, computational framework enables the investigation of numerous scenarios and sensitivity analyses. For example, one can consider the impacts of the removal of supply points and/or demand points; the addition of supply and/or demand points; changes in the prices of the medical item under study, as well as changes to the generalized transportation costs. Furthermore, one can investigate the impacts of alternative probability distribution functions.

The above numerical results are consistent with what one can expect to observe in reality in terms of how organizations would procure critical medical supplies such as N95 masks under demand unpredictability and competition. The findings confirm that more supply points with sufficient supplies are needed to ensure that organizations are not deprived of critical supplies due to competition. As a result of this competition and limited local availability, in particular, in the case of supplies such as masks and even coronavirus test kits, we are seeing several countries now setting up local production sites [6].

6 Summary and Conclusions and Suggestions for Future Research

Medical supplies are essential in the battle against the coronavirus that causes Covid-19. The demand for medical supplies globally from PPEs to ventilators has created an intense competition. PPEs are essential in protecting healthcare workers, and it now has been recognized that masks can reduce the transmission of the novel coronavirus. Ventilators, on the other hand, can be life saving for patients with severe cases of Covid-19, and convalescent plasma has become a possible interim treatment. With the pandemic, supply chains, including those for medical items, have been disrupted adding to the intense competition for such supplies.

The Covid-19 pandemic is not limited to space or time, and, therefore, there have been many shortages of medical items. In order to elucidate the competition for such

supplies in this pandemic, we developed a Generalized Nash Equilibrium model that consists of multiple supply points for the medical items and multiple demand points with the demand at the latter being stochastic. Using some recently introduced machinery, we were able to provide alternative variational inequality formulations of the equilibrium conditions. We then utilized the variational inequality with not only medical item product flows as variables but also the Lagrange multipliers associated with the supply capacities of the medical items at the supply point. We studied the model both qualitatively and quantitatively—the latter through illustrative examples that we were able to solve analytically as well as via numerical examples for which we utilized an algorithm that we proposed. The algorithm, for which we also provided convergence results, resolved the variational inequality problem into a series of subproblems for which closed form expressions in the variables were identified.

This work adds to the literature on game theory models for disaster relief with the specific features of the Covid-19 pandemic. It can be applied to study the network economics of a spectrum of medical items, both in the near term and in the longer term, as when vaccines as well as medicines for Covid-19 become available. We also highlight possible extensions of this work. For example, the model is amenable to extension to multiple medical items. It would also be very interesting to have the supplies be elastic, that is, as a function of price. We leave such research endeavors for the future.

Acknowledgments This paper is dedicated to all essential workers, including: healthcare workers, first responders, freight service providers, grocery store workers, farmers, and educators, who sacrificed so much in the Covid-19 pandemic. Your dedication and courage have graced our planet and we salute you. We also remember all those who perished because of insufficient supplies of PPEs.

References

1. E. Adida, P.-C.C. DeLaurentis, M.A. Lawley, Hospital stockpiling for disaster planning. *IIE Trans.* **43**, 348–362 (2011)
2. J. Aleccia, Market for blood plasma from COVID-19 survivors heats up (2020). NPR, May 11. <https://www.npr.org/sections/health-shots/2020/05/11/852354920/market-for-blood-plasma-from-covid-19-survivors-heats-up>
3. American Red Cross, FAQ: COVID-19 convalescent plasma (2020). <https://www.redcrossblood.org/faq.html#donating-blood-covid-19-convalescent-plasma>
4. K.J. Arrow, G. Debreu, Existence of an equilibrium for a competitive economy. *Econometrica* **22**, 265–290 (1954)
5. J.M. Berklan, Analysis: PPE costs increase over 1,000% during COVID-19 crisis (2020). <https://www.mcknights.com/news/analysis-ppe-costs-increase-over-1000-during-covid-19-crisis/>
6. K. Bradsher, China dominates medical supplies, in this outbreak and the next (2020). July 5, The New York Times. <https://www.nytimes.com/2020/07/05/business/china-medical-supplies.html?smid=tw-nytimes&smtyp=cur>
7. A. Casadevall, L.A. Pirofski, The convalescent sera option for containing COVID-19. *J. Clin. Invest.* **130**(4), 1545–1548 (2020)

8. CDC, Using Personal Protective Equipment (PPE) (2020). <https://www.cdc.gov/coronavirus/2019-ncov/hcp/using-ppe.html>
9. G. Debreu, A social equilibrium existence theorem. *Proc. Nat. Acad. Sci.* **38**(10), 886–893 (1952)
10. D. Diaz, G. Sands, C. Alesci, Protective equipment costs increase over 1,000% amid competition and surge in demand (2020). https://www.cnn.com/2020/04/16/politics/ppe-price-costs-rising-economy-personal-protective-equipment/index.html?utm_medium=social&utm_content=2020-04-16T20%3A45%3A12&utm_term=image&utm_source=twCNNp
11. J. Dong, D. Zhang, A. Nagurney, Supply chain supernetworks with random demands. *Eur. J. Oper. Res.* **156**, 194–212 (2004.)
12. K. Duan, B. Liu, C. Li, H. Zhang, T. Yu, J. Qu, M. Zhou, L. Chen, S. Meng, Y. Hu, C. Peng, Effectiveness of convalescent plasma therapy in severe COVID-19 patients. *Proc. Nat. Acad. Sci.* **117**(17), 9490–9496 (2020)
13. N. Ferguson, D. Laydon, G. Nedjati Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z.U.L.M.A. Cucunuba Perez, G. Cuomo-Dannenburg, A. Dighe, Report 9: impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand (2020)
14. A. Fischer, M. Herrich, K. Schonefeld, Generalized Nash equilibrium problems—recent advances and challenges. *Pesquisa Oper.* **34**(3), 521–558 (2014)
15. Food and Drug Administration, Donate COVID-19 plasma (2020). <https://www.fda.gov/emergency-preparedness-and-response/coronavirus-disease-2019-covid-19/donate-covid-19-plasma>
16. K. Gelles G. Petras, How ventilators work and why COVID-19 patients need them to survive coronavirus (2020). <https://www.usatoday.com/in-depth/news/2020/04/10/coronavirus-ventilator-how-works-why-covid-19-patients-need/2942996001/>
17. J. Glenza, ‘The new gold’: demand for PPE soars again amid shortage as US cases rise (2020). *The Guardian*, June 29
18. C. Goudie, B. Markoff, C. Tressel, R. Weidner, Coronavirus USA: federal fix sought for ‘Wild West’ COVID-19 PPE competition (2020). <https://abc7chicago.com/coronavirus-cases-update-map/6072209/>
19. Grifols, Frequently Asked Questions: Why should potential donors donate at a Grifols plasma donor center? (2020) <https://www.grifolspasma.com/en/donation-resources/plasma-donation-faqs>
20. W.J. Gutjahr, P.C. Nolz, Multicriteria optimization in humanitarian aid. *Eur. J. Oper. Res.* **252**, 351–366 (2016)
21. Harvard Health Publishing, Treatments for COVID-19 (2020). <https://www.health.harvard.edu/diseases-and-conditions/treatments-for-covid-19#:~:text=In%20order%20to%20donate%20plasma,reinfected%20with%20the%20virus>
22. T. Hererra, What is convalescent blood plasma, and why do we care about it? (2020). April 24, *The New York Times*. <https://www.nytimes.com/2020/04/24/smarter-living/coronavirus-convalescent-plasma-antibodies.html>.
23. J.B.T. Herron, A.G.C. Hay-David, A.D. Gilliam, P.A. Brennan, Personal protective equipment and Covid 19—a risk to healthcare staff? *Br. J. Oral Maxillofac. Surg.* **58**(5), 500–502 (2020)
24. D. Ivanov, Predicting the impacts of epidemic outbreaks on global supply chains: a simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case. *Transport. Res. E-Log* **136**, 101922 (2020)
25. A. Jacobs, M. Richtel, M. Baker, ‘At war with no ammo’: doctors say shortage of protective gear is dire. *New York Times* (2020), pp. 1547–1548. <https://www.nytimes.com/2020/03/19/health/coronavirus-masks-shortage.html>
26. John Hopkins Medicine, What is coronavirus? (2020). <https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus>
27. C.Y. Johnson, Blood plasma from people who recovered is a safe covid-19 treatment, study says (2020). June 18, *The Washington Post*. <https://www.washingtonpost.com/health/2020/06/18/blood-plasma-people-who-recovered-is-safe-covid-19-treatment-study-says/>

28. D. Kamdar, Global contest for medical equipment amidst the COVID19 pandemic (2020). <https://www.orfonline.org/expert-speak/global-contest-for-medical-equipment-amidst-the-covid19-pandemic-66438/>
29. D. Kinderlehrer, G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications* (Academic Press, New York, 1980)
30. G.M. Korpelevich, The extragradient method for finding saddle points and other problems. *Matekon* **13**, 35–49 (1977)
31. I.S. Kotsireas, A. Nagurney, P.M. Pardalos (eds.), *Dynamics of Disasters: Key Concepts, Models, Algorithms, and Insights* (Springer, Cham, 2016)
32. I.S. Kotsireas, A. Nagurney, P.M. Pardalos (eds.), *Dynamics of Disasters: Algorithmic Approaches and Applications* (Springer, Cham, 2018)
33. A.A. Kulkarni, U.V. Shanbhag, On the variational equilibrium as a refinement of the generalized Nash equilibrium. *Automatica* **48**(1), 45–55 (2012)
34. G. Lopez, Why America ran out of protective masks—and what can be done about it (2020). <https://www.vox.com/policy-and-politics/2020/3/27/21194402/coronavirus-masks-n95-respirators-personal-protective-equipment-ppe>
35. J. Mair-Jenkins, M. Saavedra-Campos, J.K. Baillie, P. Cleary, F.M. Khaw, W.S. Lim, S. Makki, K.D. Rooney, Convalescent Plasma Study Group, J.S. Nguyen-Van-Tam, C.R. Beck, The effectiveness of convalescent plasma and hyperimmune immunoglobulin for the treatment of severe acute respiratory infections of viral etiology: a systematic review and exploratory meta-analysis. *J. Infect. Dis.* **211**(1), 80–90 (2015)
36. H.O. Mete, Z.B. Zabinsky, Stochastic optimization of medical supply location and distribution in disaster management. *Int. J. Prod. Econ.* **126**(1), 76–84 (2010)
37. L. Muggy J.L. Heier Stamm, Game theory applications in humanitarian operations: a review. *J. Humanit. Logist. Supply Chain Manag.* **4**(1), 4–23 (2014)
38. A. Nagurney, *Network Economics: A Variational Inequality Approach*, 2nd and Revised Edition. (Kluwer Academic Publishers, Dordrecht, 1999)
39. A. Nagurney, Perishable food supply chain networks with labor in the Covid-19 pandemic, in *Dynamics of Disasters—Impact, Risk, Resilience, and Solutions*, ed. by I.S. Kotsireas, A. Nagurney, P.M. Pardalos (Springer, Cham, 2020). Accepted for publication.
40. A. Nagurney, E. Alvarez Flores, C. Soylu, A Generalized Nash Equilibrium model for post-disaster humanitarian relief. *Transp. Res. E* **95**, 1–18 (2016)
41. A. Nagurney, P. Daniele, E. Alvarez Flores, V. Caruso, A variational equilibrium network framework for humanitarian organizations in disaster relief: Effective product delivery under competition for financial funds, in *Dynamics of Disasters: Algorithmic Approaches and Applications*, I.S. Kotsireas, A. Nagurney, P.M. Pardalos (Springer, Cham, 2018), pp. 109–133
42. A. Nagurney, P. Dutta, *A Multiclass, Multiproduct Covid-19 Convalescent Plasma Donor Equilibrium Model* (Isenberg School of Management, University of Massachusetts Amherst, 2020)
43. A. Nagurney, A.H. Masoumi, M. Yu, Supply chain network operations management of a blood banking system with cost and risk minimization. *Comput. Manag. Sci.* **9**(2), 205–231 (2012)
44. A. Nagurney, A.H. Masoumi, M. Yu, An integrated disaster relief supply chain network model with time targets and demand uncertainty, in *Regional Science Matters: Studies Dedicated to Walter Isard*, ed. by P. Nijkamp, A. Rose, K. Kourtit (Springer, Cham, 2015), pp. 287–318
45. A. Nagurney, L.S. Nagurney, A mean-variance disaster relief supply chain network model for risk reduction with stochastic link costs, time targets, and demand uncertainty, in *Dynamics of Disasters: Key Concepts, Models, Algorithms, and Insights*, ed. by I.S. Kotsireas, A. Nagurney, P.M. Pardalos (Springer, Cham, 2016), pp. 231–255
46. A. Nagurney, Q. Qiang, *Fragile Networks: Identifying Vulnerabilities and Synergies in an Uncertain World* (Wiley, Hoboken, 2009)
47. A. Nagurney, M. Salarpour, P. Daniele, An integrated financial and logistical game theory model for humanitarian organizations with purchasing costs, multiple freight service providers, and budget, capacity, and demand constraints. *Int. J. Prod. Econ.* **212**, 212–226 (2019)

48. A. Nagurney, M. Salarpour, J. Dong, L.S. Nagurney, A stochastic disaster relief game theory network model. *SN Oper. Res. Forum* **1**(10), 1–33 (2020)
49. A. Nagurney, M. Yu, Q. Qiang, Supply chain network design for critical needs with outsourcing. *Papers in Regional Science* **90**, 123–142 (2011)
50. A. Nagurney, M. Yu, D. Besik, Supply chain network capacity competition with outsourcing: a variational equilibrium framework. *J. Global Optim.* **69**(1), 231–254 (2017)
51. A. Nagurney, D. Zhang, *Projected Dynamical Systems and Variational Inequalities with Applications* (Kluwer Academic Publishers, Boston, 1996)
52. S.A. Namendys-Silva, Respiratory support for patients with COVID-19 infection. *Lancet Respir. Med.* **8**(4), e18 (2020)
53. J.F. Nash, Equilibrium points in n-person games. *Proc. Nat. Acad. Sci.* **36**(1), 48–49 (1950)
54. J.F. Nash, Non-cooperative games. *Ann. Math.* **54**, 286–295 (1951)
55. J. O’Connell, Doctors say their PPE supply could run dry in weeks. *The Times-Tribune*, July 6 (2020)
56. R. Pifer, 7 states team up to buy \$5B in medical equipment, supplies for COVID-19 (2020). <https://www.healthcarediver.com/news/7-states-team-up-to-buy-5b-in-medical-equipment-supplies-for-covid-19/577263/>
57. M.M. Queiroz, D. Ivanov, A. Dolgui, S.F. Wamba, Impacts of epidemic outbreaks on supply chains: mapping a research agenda amid the COVID-19 pandemic through a structured literature review. *Ann. Oper. Res.* 1–38 (2020). <https://doi.org/10.1007/s10479-020-03685-7>
58. SCCM, United States resource availability for COVID-19 (2020). <https://sccm.org/Blog/March-2020/United-States-Resource-Availability-for-COVID-19>
59. Z. Schlanger, Begging for thermometers, body bags, and gowns: U.S. health care workers are dangerously ill-equipped to fight COVID-19 (2020). <https://time.com/5823983/coronavirus-ppe-shortage/>
60. H. Secon A. Woodward D. Mosher A comprehensive timeline of the new coronavirus pandemic, from China’s first COVID-19 case to the present (2020). <https://www.businessinsider.com/coronavirus-pandemic-timeline-history-major-events-2020-3>
61. The Society for Healthcare Organization Procurement Professionals, SHOPP Covid PPD Costs analysis (2020). <http://cdn.cnn.com/cnn/2020/images/04/16/shopp.covid.ppd.costs.analysis.pdf>
62. J. Van Griensven, T. Edwards, X. de Lamballerie, M.G. Semple, P. Gallian, S. Baize, P.W. Horby, H. Raoul, N.F. Magassouba, A. Antierens, C. Lomas, Evaluation of convalescent plasma for Ebola virus disease in Guinea. *N. Engl. J. Med.* **374**(1), 33–42 (2016)
63. W. Wan, America is running short on masks, gowns and gloves. Again. *The Washington Post*, July 8 (2020)
64. A.M. Winkler, S.A. Koepsell, The use of convalescent plasma to treat emerging infectious diseases: focus on Ebola virus disease. *Curr. Opin. Hematol.* **22**(6), 521–526. (2015)

Relative Strongly Exponentially Convex Functions



Muhammad Aslam Noor, Khalida Inayat Noor, and Themistocles M. Rassias

Abstract In this paper, we define and consider some new concepts of the strongly exponentially convex functions involving an arbitrary negative bifunction. Some properties of these strongly exponentially convex functions are investigated under suitable conditions. It is shown that the difference of strongly exponentially convex functions and strongly exponentially affine functions is again an exponentially convex function. Results obtained in this paper can be viewed as refinement and improvement of previously known results

1 Introduction

Convexity theory has played a fundamental role in the development of various mathematical and engineering sciences. This theory provides us a powerful tool to tackle unrelated complicated problems in a unified and general framework. Convex functions and convex sets have been generalized and extended in different directions using novel and innovative ideas. This theory has applications in almost every field and continues to stimulate research in every direction. Polyak [27] introduced the strongly convex functions in the optimization theory, which inspired a great deal of interest. Karmardian [14] used the strongly convex functions to discuss the unique existence of a solution of the nonlinear complementarity problems. Zu and Marcotte [30] used strongly convex functions in the convergence analysis of the iterative methods for solving variational inequalities and equilibrium problems. Nikodem and Pales [19] investigated the characterization of the inner product spaces using the strongly convex functions, which can be viewed as a novel and innovative application. It is also known that the minimum of the strongly convex functions

M. A. Noor (✉) · K. I. Noor
COMSATS University Islamabad, Islamabad, Pakistan

T. M. Rassias
National Technical University of Athens, Athens, Greece
e-mail: trassias@math.ntua.gr

is unique. Qu and Li [28] investigated the exponentially stability of primal-dual gradient dynamics using the concept of strongly convex functions. Awan et al. [6–9] and Noor et al. [24] have derived Hermite–Hadamard type inequalities for various classes of strongly convex functions, which provide upper and lower estimates for the integrand. For more applications and properties of the strongly convex functions, cf. [1–4, 6–11, 14–20, 22–24, 27, 28, 30] and the references therein.

Closely related to the log-convex functions, we have the concept of exponentially convex (concave) functions, and the origin of exponentially convex functions can be traced back to Bernstein [11]. Avriel [5] introduced and studied the concept of r -convex functions. For further properties of the r -convex functions, see Zhao et al. [30] and the references therein. Antczak [4] also explored the application in the mathematical programming and optimization. The exponentially convex functions have important applications in information theory, big data analysis, machine learning, and statistic (cf. [2, 4, 25]) and the references therein. Exponentially convex (concave) functions can be considered as a significant extension of the convex functions. Pal and Wong [25] have discussed its role in information geometry and statistics. Antczak [4] introduced these exponentially convex functions implicitly and discussed their role in mathematical programming. Alirazaie and Mathur [2], Dragomir and Gomm [13], Noor et al. [20, 21], and Rashid et al. [29] have derived Hermite–Hadamard type integral inequalities for these exponentially convex functions. Noor and Noor [22] introduced and considered some new class of strongly exponentially convex functions involving an arbitrary bifunction, where several concepts of monotonicity were discussed.

Inspired by the work of Adamek [1], Nikodem et al. [19], and Noor et al. [20, 22, 23], we introduce a new class of strongly exponentially convex functions involving a exponentially bifunction, which is called the relative strongly exponentially convex function. We establish the relationship between these classes and derive some new results under some mild conditions. We have also investigated the optimality conditions for the relative strongly exponentially convex functions. It is shown that the minimum of the differentiable exponentially convex function can be characterized by a class of variational inequalities, which is called exponential variational inequality. It is shown that the difference of strongly exponentially convex functions and strongly exponentially affine functions is again an exponentially convex function. It is expected that relative strongly exponentially convex functions can have similar applications in optimization theory and variational inequalities to the ones that strongly convex functions have.

2 Preliminary Results

Let K be a nonempty closed set in a real Hilbert space H . We denote by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ the inner product and norm, respectively. Let $F : K \rightarrow \mathbb{R}$ be a continuous function.

We now recall some well known and basic concepts and results, cf. [12, 18, 26].

Definition 1 ([18]) The set K in H is said to be convex set, if

$$u + t(v - u) \in K, \quad \forall u, v \in K, t \in [0, 1].$$

We now consider a class of exponentially convex functions.

Definition 2 A function F is said to be exponentially convex function, if

$$e^{F((1-t)a+tb)} \leq (1-t)e^{F(a)} + te^{F(b)}, \quad \forall a, b \in K, \quad t \in [0, 1]. \quad (1)$$

Definition 2 can be stated in the following equivalent form, which is due to Antczak [4].

Definition 3 A function F is said to be exponentially convex function, if

$$e^{F((1-t)a+tb)} \leq \log[(1-t)e^{F(a)} + te^{F(b)}], \quad \forall a, b \in K, \quad t \in [0, 1]. \quad (2)$$

We would like to mention that the function $f(x) = e^x$ is not a convex function, but it is an exponentially convex function.

We now define the exponentially convex functions on the interval $I[a, b]$, which is mainly due to Noor and Noor [21–23].

Definition 4 Let $I = [a, b]$. Then F is exponentially convex function, if and only if,

$$\begin{vmatrix} 1 & 1 & 1 \\ a & x & b \\ e^{F(a)} & e^{F(x)} & e^{F(b)} \end{vmatrix} \geq 0; \quad a \leq x \leq b.$$

One can easily show that the following are equivalent:

1. F is exponentially convex function.
2. $e^{F(x)} \leq e^{F(a)} + \frac{e^{F(b)} - e^{F(a)}}{b-a}(x - a)$.
3. $e^{F(x)} \leq \frac{b-x}{b-a}e^{F(a)} + \frac{x-a}{b-a}e^{F(b)}$.
4. $\frac{e^{F(x)} - e^{F(a)}}{x-a} \leq \frac{e^{F(b)} - e^{F(a)}}{b-a}$.
5. $(b-x)e^{F(a)} + (a-b)e^{F(x)} + (x-a)e^{F(b)} \geq 0$.
6. $\frac{e^{F(a)}}{(b-a)(a-x)} + \frac{e^{F(x)}}{(x-b)(a-x)} + \frac{e^{F(b)}}{(b-a)(x-b)} \geq 0$,

where

$$x = (1-t)a + tb \in [0, 1].$$

Remark 1 If the exponentially convex function F is differentiable, then, from

$$e^{F(x)} \leq e^{F(a)} + \frac{e^{F(b)} - e^{F(a)}}{b-a}(x - a),$$

we have

$$\langle F'(a)e^{F(a)}, b - a \rangle \leq e^{F(b)} - e^{F(a)},$$

where $F'(\cdot)$ is the differential of the function F .

For the applications of the exponentially convex (concave) functions in the mathematical programming and information theory, see Antczak [4] and Alirezaei and Mathar[2].

We now introduce the concept of the strongly exponentially convex functions involving an arbitrary bifunction, which is the main motivation of this paper.

Definition 5 The function F on the convex set K is said to be relative strongly exponentially convex with respect to an arbitrary bifunction $F(\cdot, \cdot)$, if there exists a constant $\mu > 0$, such that

$$e^{F(u+t(v-u))} \leq (1-t)e^F(u) + te^F(v) - \mu t(1-t)e^{F(v,u)}, \forall u, v \in K, t \in [0, 1]. \quad (3)$$

The function F is said to be relative strongly exponentially concave, if and only if, $-F$ is strongly exponentially convex.

If $t = \frac{1}{2}$ and $\mu = 1$, then

$$e^{F(\frac{u+v}{2})} \leq \frac{e^F(u) + e^F(v)}{2} - \frac{1}{4}e^{F(v,u)}, \quad \forall u, v \in K. \quad (4)$$

The function F is called the relative strongly exponentially J -convex function.

We also introduce the concept of strongly exponentially affine convex functions.

Definition 6 The function F on the convex set K is said to be relative strongly exponentially affine convex involving the arbitrary bifunction $F(\cdot, \cdot)$, if there exists a constant $\mu > 0$, such that

$$e^{F(u+t(v-u))} = (1-t)e^{F(u)} + te^{F(v)} - \mu t(1-t)e^{F(v,u)}, \quad \forall u, v \in K, t \in [0, 1]. \quad (5)$$

Also, we say that the function F is relative strongly exponentially affine J -convex function, if

$$e^{F(\frac{u+v}{2})} = \frac{e^{F(u)} + e^{F(v)}}{2} - \frac{1}{4}\mu e^{F(v,u)}, \quad \forall u, v \in K. \quad (6)$$

We now discuss some special cases of strongly exponentially convex functions.

I. If $e^{F(v-u)} = \|v - u\|^2$, then

$$e^{F(u+t(v-u))} \leq (1-t)e^{F(u)} + te^{F(v)} - \mu t(1-t)\|v-u\|^2, \quad \forall u, v \in K, t \in [0, 1],$$

which is called the strongly exponentially convex function introduced and studied by Noor and Noor [23].

II. If $e^{F(v-u)} = F(v, u)$, then

$$e^{F(u+t(v-u))} \leq (1-t)e^{F(u)} + te^{F(v)} - \mu t(1-t)F(v-u), \quad \forall u, v \in K, t \in [0, 1], \quad (7)$$

which is known as strongly exponentially convex function involving the bifunction $F(v, u)$, see Noor and Noor [22].

For the properties of the strongly convex functions in optimization, Inequalities, and equilibrium problems, cf. [1, 3–8, 10–22, 25–28, 30] and the references therein.

Definition 7 The function F on the convex set K is said to be relative strongly exponentially quasi-convex involving the bifunction $F(v, u)$, if there exists a constant $\mu > 0$ such that

$$e^{F(u+t(v-u))} \leq \max\{e^{F(u)}, e^{F(v)}\} - \mu t(1-t)e^{F(v,u)}, \quad \forall u, v \in K, t \in [0, 1].$$

Definition 8 The function F on the convex set K is said to be relative strongly exponentially log-convex with respect to an arbitrary bifunction $F(v, u)$, if there exist a constant $\mu > 0$ such that

$$e^{F(u+t(v-u))} \leq (e^{F(u)})^{1-t}(e^{F(v)})^t - \mu t(1-t)e^{F(v,u)}, \quad \forall u, v \in K, t \in [0, 1],$$

where $F(\cdot) > 0$.

From this definition, we have

$$\begin{aligned} e^{F(u+t(v-u))} &\leq (e^{F(u)})^{1-t}(e^{F(v)})^t - \mu t(1-t)e^{F(v,u)}, \\ &= (1-t)e^{F(u)} + te^{F(v)} - \mu t(1-t)e^{F(v,u)}. \end{aligned}$$

This shows that every strongly exponentially convex function is a strongly exponentially convex function, but the converse is not true.

In fact, we have

$$\begin{aligned} e^{F(u+t(v-u))} &\leq (e^{F(u)})^{1-t}(e^{F(v)})^t - \mu t(1-t)e^{F(v,u)} \\ &\leq (1-t)e^{F(u)} + te^{F(v)} - \mu t(1-t)e^{F(v,u)} \\ &\leq \max\{e^{F(u)}, e^{F(v)}\} - \mu t(1-t)e^{F(v,u)}. \end{aligned}$$

This shows that every strongly exponentially log-convex function is a strongly exponentially convex function, and every strongly exponentially convex function is a strongly exponentially quasi-convex function. However, the converse is not true.

Definition 9 A differentiable function F on the convex set K is said to be strongly exponentially pseudo-convex function, if and only if there exists a constant $\mu > 0$ such that

$$\langle e^{F(u)} F'(u), v - u \rangle + \mu e^{F(v,u)} \geq 0$$

$$\Rightarrow e^{F(v)} - e^{F(u)} \geq 0, \quad \forall u, v \in K.$$

We also need the following assumptions regarding the bifunction $F(., .)$, which plays a crucial role in the derivation of our results.

Condition N. Assume that the function $F(., .)$ satisfies these assumptions:

$$\begin{aligned} e^{F(u, u+t(v-u))} &= t^2 e^{F(v, u)} \\ e^{F(v, u+t(v-u))} &= (1-t)^2 e^{F(v, u)}, \quad \forall u, v \in K, t \in [0, 1]. \end{aligned}$$

Definition 10 An exponential function F is said to be homogeneous of degree 2, if

$$e^{F(\lambda x)} = \lambda^2 e^{F(x)}, \quad \forall \lambda \in \mathbb{R}^n.$$

3 Main Results

In this section, we consider some basic properties of generalized strongly convex functions.

Theorem 1 Let F be a differentiable function on the convex set K and condition N holds. Then the function F is relative strongly exponentially convex function, if and only if,

$$e^{F(v)} - e^{F(u)} \geq \langle e^{F(u)} F'(u), v - u \rangle + \mu e^{F(v, u)}, \quad \forall v, u \in K. \tag{8}$$

Proof Let F be a relative strongly exponentially convex function on the convex set K . Then

$$e^{F(u+t(v-u))} \leq (1-t)e^{F(u)} + te^{F(v)} - t(1-t)\mu e^{F(v, u)}, \quad \forall u, v \in K,$$

which can be written as

$$e^{F(v)} - e^{F(u)} \geq \left\{ \frac{e^{F(u+t(v-u))} - e^{F(u)}}{t} \right\} + (1-t)\mu e^{F(v, u)}.$$

Taking the limit in the above inequality as $t \rightarrow 0$, we have

$$e^{F(v)} - e^{F(u)} \geq \langle e^{F(u)} F'(u), v - u \rangle + \mu e^{F(v, u)},$$

which is (8), the required result.

Conversely, let (8) hold. Then $\forall u, v \in K, t \in [0, 1], v_t = u + t(v - u) \in K$, and using condition N, we have

$$\begin{aligned} e^{F(v)} - e^{F(v_t)} &\geq \langle e^{F(v_t)} F'(v_t), v - v_t \rangle + \mu e^{F(v_t-u)} \\ &= (1-t) \langle e^{F(v_t)} F'(v_t), v - u \rangle + \mu(1-t)^2 e^{F(v,u)}. \end{aligned} \quad (9)$$

In a similar way, we have

$$\begin{aligned} e^{F(u)} - e^{F(v_t)} &\geq \langle e^{F(v_t)} F'(v_t), u - v_t \rangle + \mu e^{F(u-v_t)} \\ &= -t \langle e^{F(v_t)} F'(v_t), v - u \rangle + \mu t^2 e^{F(v,u)}. \end{aligned} \quad (10)$$

Multiplying (9) by t and (10) by $(1-t)$ and adding the resultant, we have

$$e^{F(u+t(v-u))} \leq (1-t)e^{F(u)} + te^{F(v)} - t(1-t)\mu e^{F(v,u)},$$

showing that F is a relative strongly exponentially convex function. \square

Theorem 2 Let F be a differentiable relative strongly exponentially convex function and condition N holds. Then, (8) holds, if and only if,

$$\langle e^{F(u)} F'(u) - e^{F(v)} F'(v), u - v \rangle \geq 2\mu e^{F(v,u)}, \quad \forall u, v \in K. \quad (11)$$

Proof Let F be a relative strongly exponentially convex function on the convex set K . Then, from Theorem 3.1, we have

$$e^{F(v)} - e^{F(u)} \geq \langle e^{F(u)} F'(u), v - u \rangle + \mu e^{F(v,u)}, \quad \forall u, v \in K. \quad (12)$$

Changing the role of u and v in (12), we have

$$e^{F(u)} - e^{F(v)} \geq \langle e^{F(v)} F'(v), u - v \rangle + \mu e^{F(v,u)}, \quad \forall u, v \in K. \quad (13)$$

Adding (12) and (13), we have

$$\langle e^{F(u)} F'(u) - e^{F(v)} F'(v), u - v \rangle \geq 2\mu e^{F(v,u)},$$

the required (11).

Conversely, let F' satisfy (11). Then, from (17), we have

$$\langle e^{F(v)} F'(v), u - v \rangle \leq \langle e^{F(u)} F'(u), u - v \rangle - 2\mu e^{F(v,u)}. \quad (14)$$

Since K is a convex set, $\forall u, v \in K, t \in [0, 1], v_t = u + t(v - u) \in K$.

Taking $v = v_t$ in (14) and using condition N , we have

$$\begin{aligned} \langle e^{F(v_t)} F'(v_t), u - v_t \rangle &\leq \langle e^{F(u)} F'(u), u - v_t \rangle - 2\mu e^{F(u,v_t)} \\ &= -t \langle e^{F(u)} F'(u), v - u \rangle - 2t^2 \mu e^{F(v,u)}, \end{aligned}$$

which implies that

$$\langle e^{F(v_t)} F'(v_t), v - u \rangle \geq \langle e^{F(u)} F'(u), v - u \rangle + 2t\mu e^{F(v,u)}. \tag{15}$$

Consider the auxiliary function

$$g(t) = e^{F(u+t(v-u))},$$

from which, we have

$$g(1) = e^{F(v)}, \quad g(0) = e^{F(u)}.$$

Then, from (15), we have

$$\begin{aligned} g'(t) &= \langle e^{F(v_t)} F'(v_t), v - u \rangle \\ &\geq \langle e^{F(u)} F'(u), v - u \rangle + 2\mu t e^{F(v,u)}. \end{aligned} \tag{16}$$

Integrating (16) between 0 and 1, we have

$$g(1) - g(0) = \int_0^1 g'(t) dt \geq \langle e^{F(u)} F'(u), v - u \rangle + \mu e^{F(v,u)}.$$

Thus, it follows that

$$e^{F(v)} - e^{F(u)} \geq \langle e^{F(u)} F'(u), v - u \rangle + \mu e^{F(v,u)},$$

which is (8) as required. □

Theorems 1 and 2 enable us to introduce the following new concepts.

Definition 11 The differential $F'(\cdot)$ of the strongly exponentially convex functions is said to be relative strongly exponentially monotone, if

$$\langle e^{F(u)} F'(u) - e^{F(v)} F'(v), u - v \rangle \geq \mu e^{F(v,u)}, \forall u, v \in H.$$

Definition 12 The differential $F'(\cdot)$ of the exponentially convex functions is said to be exponentially monotone, if

$$\langle e^{F(u)} F'(u) - e^{F(v)} F'(v), u - v \rangle \geq 0, \forall u, v \in H.$$

Definition 13 The differential $F'(\cdot)$ of the relative strongly exponentially convex functions is said to be relative strongly exponentially pseudomonotone, if

$$\langle e^{F(u)} F'(u), v - u \rangle \geq 0$$

implies that

$$\langle e^{F(v)} F'(v), v - u \rangle \geq \mu e^{F(v,u)}, \quad \forall u, v \in H. \tag{17}$$

We now give a necessary condition for the relative strongly exponentially pseudo-convex function.

Theorem 3 *Let F' be a relative strongly exponentially pseudomonotone and condition N holds. Then, F is a relative strongly exponentially pseudo-invex function.*

Proof Let F' be a relative strongly exponentially pseudomonotone operator. Then, $\forall u, v \in K$,

$$\langle e^{F(u)} F'(u), v - u \rangle \geq 0$$

implies that

$$\langle e^{F(v)} F'(v), v - u \rangle \geq \mu e^{F(v,u)}. \tag{18}$$

Since K is a convex set, $\forall u, v \in K, t \in [0, 1], v_t = u + t(v - u) \in K$. Taking $v = v_t$ in (18) and using condition N, we have

$$\langle e^{F(v_t)} F'(v_t), v - u \rangle \geq t\mu e^{F(v,u)}. \tag{19}$$

Consider the auxiliary function

$$g(t) = e^{F(u+t(v-u))} = e^{F(v_t)}, \quad \forall u, v \in K, t \in [0, 1],$$

which is differentiable, since F is a differentiable function. Then, using (19), we have

$$g'(t) = \langle e^{F(v_t)} F'(v_t), v - u \rangle \geq t\mu e^{F(v,u)}.$$

Integrating the above relation between 0 and 1, we have

$$g(1) - g(0) = \int_0^1 g'(t) dt \geq \frac{\mu}{2} e^{F(v,u)},$$

that is,

$$e^{F(v)} - e^{F(u)} \geq \frac{\mu}{2} e^{F(v,u)},$$

showing that F is a relative strongly exponentially pseudo-convex function. □

Definition 14 The function F is said to be sharply relative strongly exponentially pseudo-convex, if there exists a constant $\mu > 0$, such that

$$\langle e^{F(u)} F'(u), v - u \rangle \geq 0$$

$$\Rightarrow F(v) \geq e^{F(v+t(u-v))} + \mu t(1-t)e^{F(v,u)} \quad \forall u, v \in K, t \in [0, 1].$$

Theorem 4 *Let F be a sharply relative strongly exponentially pseudo-convex function on K with a constant $\mu > 0$. Then*

$$\langle e^{F(v)} F'(v), v - u \rangle \geq \mu e^{F(v,u)} \quad \forall u, v \in K.$$

Proof Let F be a sharply relative strongly exponentially pseudo-convex function. Then

$$e^{F(v)} \geq e^{F(v+t(u-v))} + \mu t(1-t)e^{F(v,u)}, \quad \forall u, v \in K, t \in [0, 1],$$

from which, we have

$$\left\{ \frac{e^{F(v+t(u-v))} - e^{F(v)}}{t} \right\} + \mu t(1-t)e^{F(v,u)} \leq 0.$$

Taking the limit in the above inequality, as $t \rightarrow 0$, we have

$$\langle e^{F(v)} F'(v), v - u \rangle \geq \mu e^{F(v,u)},$$

which is the required result. □

We now discuss the optimality condition for the differentiable relative strongly exponentially convex functions, which is the main motivation of our next result.

Theorem 5 *Let F be a differentiable relative strongly exponentially convex function with modulus $\mu > 0$. If $u \in K$ is the minimum of the function F , then*

$$e^{F(v)} - e^{F(u)} \geq \mu e^{F(v,u)}, \quad \forall u, v \in K. \tag{20}$$

Proof Let $u \in K$ be a minimum of the function F . Then

$$F(u) \leq F(v), \forall v \in K,$$

from which, we have

$$e^{F(u)} \leq e^{F(v)}, \forall v \in K. \tag{21}$$

Since K is a convex set, so, $\forall u, v \in K, \quad t \in [0, 1],$

$$v_t = (1-t)u + tv \in K.$$

Taking $v = v_t$ in (21), we have

$$0 \leq \lim_{t \rightarrow 0} \left\{ \frac{e^{F(u+t(v-u))} - e^{F(u)}}{t} \right\} = \langle e^{F(u)} F'(u), v - u \rangle. \quad (22)$$

Since F is a differentiable relative strongly exponentially convex function, so

$$e^{F(u+t(v-u))} \leq e^{F(u)} + t(e^{F(v)} - e^{F(u)}) - \mu t(1-t)e^{F(v,u)}, \quad u, v \in K, t \in [0, 1],$$

from which, using (22), we have

$$\begin{aligned} e^{F(v)} - e^{F(u)} &\geq \lim_{t \rightarrow 0} \left\{ \frac{e^{F(u+t(v-u))} - e^{F(u)}}{t} \right\} + \mu e^{F(v,u)}. \\ &= \langle e^{F(u)} F'(u), v - u \rangle + \mu e^{F(v,u)} \\ &\geq \mu e^{F(v,u)}, \end{aligned}$$

the required result (20). \square

Remark 2 We would like to mention that, if

$$\langle e^{F(u)} F'(u), v - u \rangle + \mu e^{F(v,u)} \geq 0, \quad \forall u, v \in K,$$

then $u \in K$ is the minimum of the function F .

We would like to emphasize that the minimum $u \in K$ of the exponentially convex functions can be characterized by the inequality

$$\langle e^{F(u)} F'(u), v - u \rangle \geq 0, \quad \forall v \in K, \quad (23)$$

which is called the exponential variational inequality and appears to be a new one. It is an interesting problem to study the existence of a unique solution of the inequality (23) and its applications.

Definition 15 A function F is said to be an exponentially pseudo-convex function, if there exists a strictly positive bifunction $b(\cdot, \cdot)$, such that

$$\begin{aligned} e^{F(v)} &< e^{F(u)} \\ &\Rightarrow \\ e^{F(u+t(v-u))} &< e^{F(u)} + t(t-1)b(v, u), \quad \forall u, v \in K, t \in [0, 1]. \end{aligned}$$

Theorem 6 If the function F is an exponentially convex function such that $e^{F(v)} < e^{F(u)}$, then F is a relative strongly exponentially pseudo-convex function.

Proof Since $e^{F(v)} < e^{F(u)}$ and F is a relative strongly exponentially convex function, so $\forall u, v \in K, t \in [0, 1]$, we have

$$\begin{aligned}
 e^{F(u+t(v-u))} &\leq e^{F(u)} + t(e^{F(v)} - e^{F(u)}) - \mu t(1-t)e^{F(v,u)} \\
 &< e^{F(u)} + t(1-t)(e^{F(v)} - e^{F(u)}) - \mu t(1-t)e^{F(v,u)} \\
 &= e^{F(u)} + t(t-1)(e^{F(u)} - e^{F(v)}) - \mu t(1-t)e^{F(v,u)} \\
 &< e^{F(u)} + t(t-1)b(u, v) - \mu t(1-t)e^{F(v,u)},
 \end{aligned}$$

where $b(u, v) = e^{F(u)} - e^{F(v)} > 0$, the required result. This shows that the function F is the relative strongly exponentially convex function. \square

It is well known that each strongly convex function is of the form $f \pm \|\cdot\|^2$, where f is a convex function. A similar result is proved for relative strongly exponentially convex functions. In this direction, we have:

Theorem 7 *Let f be a relative strongly exponentially affine function. Then F is a relative strongly exponentially convex function, if and only if, $g = F - f$ is an exponentially convex function.*

Proof Let f be a relative strongly exponentially affine function. Then

$$e^{f((1-t)u+tv)} = (1-t)e^{f(u)} + te^{f(v)} - \mu t(1-t)e^{f(v,u)}. \tag{24}$$

From the relative strongly exponentially convexity of F , we have

$$e^{F((1-t)u+tv)} \leq (1-t)e^{F(u)} + te^{F(v)} - \mu t(1-t)e^{F(v,u)}. \tag{25}$$

From (24) and (25), we have

$$e^{F((1-t)u+tv)} - e^{f((1-t)u+tv)} \leq (1-t)(e^{F(u)} - e^{f(u)}) + t(e^{F(v)} - e^{f(v)}), \tag{26}$$

from which, it follows that

$$\begin{aligned}
 e^{g((1-t)u+tv)} &= e^{F((1-t)u+tv)} - e^{f((1-t)u+tv)} \\
 &\leq (1-t)(e^{F(u)} - e^{f(u)}) + t(e^{F(v)} - e^{f(v)}),
 \end{aligned}$$

which shows that $g = F - f$ is an exponentially convex function.

The inverse implication is obvious. \square

We would like to remark that one can show that a function F is a relative strongly exponentially convex function, if and only if, F is a relative strongly exponentially affine function essentially using the technique of Adamek [1] and Noor et al. [17].

We would like to note that the relative strongly exponentially convex function is also a Wright strongly convex function. From Definition 5, we have

$$e^{F((1-t)u+tv)} + e^{F(tu+(1-t)v)} \leq \frac{1}{2}\{e^{F(u)} + e^{F(v)}\} - 2\mu t(1-t)e^{F(v,u)}, \forall u, v \in K, t \in [0, 1],$$

which is called the relative Wright strongly exponentially convex function. It is an interesting problem to study the properties and applications of the relative Wright strongly exponentially convex functions.

Definition 16 For a given function F , we consider the exponentially quadratic equation:

$$e^{F(x+y)} + e^{F(y-x)} = 2e^{F(x)} + 2e^{F(y)}, \quad \forall x, y \in R^n. \tag{27}$$

We now prove the equivalence between the strongly exponentially J -convex functions and the exponentially quadratic equations.

Theorem 8 Let F be a even and exponentially homogeneous function of degree 2. Then the function F is a strongly exponentially J -convex function, that is

$$e^{F(\frac{u+v}{2})} \leq \frac{1}{2}[e^{Fu} + e^{Fv}] - \frac{1}{4}e^{F(v-u)}, \quad \forall u, v \in R^n,$$

if and only if, F is an exponentially quadratic equation (27).

Proof Let F be a strongly exponentially J -convex function. Then

$$e^{F(\frac{u+v}{2})} \leq \frac{e^F(u) + e^F(v)}{2} - \frac{1}{4}e^{F(v-u)}, \quad \forall u, v \in R^n,$$

from which, it follows that

$$e^{F(u+v)} + e^{F(v-u)} \leq 2e^{F(u)} + 2e^{F(v)}, \quad \forall u, v \in R^n, \tag{28}$$

where we have used the fact that F is a exponentially homogenous function of degree 2.

Taking $u + v = w$, $v - u = z$, in (28) and using the fact that F is exponentially homogeneous, we have

$$2e^{F(w)} + 2e^{F(z)} \leq e^{F(w+z)} + e^{F(w-x)}, \quad \forall w, z \in R^n. \tag{29}$$

From the inequalities (28) and (29), we have

$$e^{F(u+v)} + e^{F(v-u)} = 2e^{F(u)} + 2e^{F(v)}, \quad \forall u, v \in R^n,$$

which is the exponentially quadratic equation (27).

Converse is obvious. □

Remark 3 If $F(u) = \|u\|^2$, then the exponentially quadratic equations (27) reduce to

$$e^{\|u+v\|^2} + e^{\|v-u\|^2} = 2e^{\|u\|^2} + 2e^{\|v\|^2}, \quad \forall u, v \in R^n,$$

which is called the exponentially parallelogram law. This means that the space R^n is an exponentially inner product space, see Nikodem and Pales [19].

4 Conclusion

In this paper, we have introduced and studied a new class of exponentially convex functions involving an arbitrary bifunction, which is called the relative strongly exponentially convex function. We have studied the basic properties of these functions. Several new and interesting results have been obtained. It is shown that the optimality conditions of the differentiable relative strongly exponentially convex functions can be characterized by a class of variational inequalities, which are called exponentially variational inequalities. The qualitative study of the exponentially variational inequalities is an interesting problem for future research. It is expected that the ideas and techniques of this paper may motivate further research.

Acknowledgment The authors would like to thank the Rector, COMSATS University Islamabad, Pakistan, for providing excellent research and academic environments.

References

1. M. Adamek, On a problem connected with strongly convex functions. *Math. Inequal. Appl.* **19**(4), 1287–1293 (2016)
2. G. Alirezaei, R. Mazhar, On exponentially concave functions and their impact in information theory. *J. Inform. Theory Appl.* **9**(5), 265–274 (2018)
3. H. Angulo, J. Gimenez, A.M. Moeos, K. Nikodem, On strongly h -convex functions. *Ann. Funct. Anal.* **2**(2), 85–91 (2011)
4. T. Antczak, On (p, r) -invex sets and functions. *J. Math. Anal. Appl.* **263**, 355–379 (2001)
5. M. Avriel, r -Convex functions. *Math. Program.* **2**, 309–323 (1972)
6. M.U. Awan, M.A. Noor, K.I. Noor, F. Safdar, On strongly generalized convex functions. *Filomat* **31**(18), 5783–5790 (2017)
7. M.U. Awan, M.A. Noor, K.I. Noor, Hermite-Hadamard inequalities for exponentially convex functions. *Appl. Math. Inform. Sci.* **12**(2), 405–409 (2018)
8. M.U. Awan, M.A. Noor, M.V. Mihai, K.I. Noor, N. Akhtar, On approximately harmonic h -convex functions depending on a given function. *Filomat* **33**(12), 3783–3793 (2019)
9. M.U. Awan, M.A. Noor, T.-S. Du, K.I. Noor, New refinements of fractional Hermite-Hadamard inequality. *RACSAM* **113**, 21–29 (2019)
10. A. Azcar, J. Gimenez, K. Nikodem, J.L. Sanchez, On strongly midconvex functions. *Opuscula Math.* **31**(1), 15–26 (2011)
11. S.N. Bernstein, Sur les fonctions absolument monotones. *Acta Math.* **52**, 1–66 (1929)
12. G. Cristescu, L. Lupsa, *Non-Connected Convexities and Applications* (Kluwer Academic Publishers, Dordrecht, 2002)
13. S.S. Dragomir, I. Gomm, Some Hermite-Hadamard type inequalities for functions whose exponentials are convex. *Stud. Univ. Babeş-Bolyai Math.* **60**(4), 527–534 (2015)
14. S. Karamardian, The nonlinear complementarity problems with applications, part 2. *J. Optim. Theory Appl.* **4**(3), 167–181 (1969)

15. T. Lara, N. Merentes, K. Nikodem, Strongly h -convexity and separation theorems. *Int. J. Anal.* **5** (2016). Article ID 7160348
16. N. Merentes, K. Nikodem, Remarks on strongly convex functions. *Aequationes Math.* **80**(1–2)(2010) 193–199
17. S.K. Mishra, N. Sharma, On strongly generalized convex functions of higher order. *Math. Inequal. Appl.* **22**(1), 111–121 (2019)
18. C.P. Niculescu, L.E. Persson, *Convex Functions and Their Applications* (Springer, New York, 2018)
19. K. Nikodem, Z.S. Pales, Characterizations of inner product spaces by strongly convex functions. *Banach J. Math. Anal.* **1**, 83–87 (2011)
20. M.A. Noor, K.I. Noor, On generalized strongly convex functions involving bifunction. *Appl. Math. Inform. Sci.* **13**(3), 411–416 (2019)
21. M.A. Noor, K.I. Noor, Exponentially convex functions. *J. Orisa Math. Soc.* **39** (2019)
22. M.A. Noor, K.I. Noor, On strongly exponentially preinvex functions. *U.P.B. Sci. Bull. Ser. A* **81** (2019)
23. M.A. Noor, K.I. Noor, Strongly exponentially convex functions and their properties. *J. Adv. Math. Stud.* **12**(2), 177–185 (2019)
24. M.A. Noor, K.I. Noor, S. Iftikhar, F. Safdar, Some properties of generalized strongly harmonic convex functions. *Inter. J. Anal. Appl.* **16**(3), 427–436 (2018)
25. S. Pal, T.K. Wong, On exponentially concave functions and a new information geometry. *Annal. Prob.* **46**(2), 1070–1113 (2018)
26. J. Pecaric, F. Proschan, Y.L. Tong, *Convex Functions, Partial Ordering and Statistical Applications* (Academic Press, New York, 1966)
27. B.T. Polyak, Existence theorems and convergence of minimizing sequences in extremum problems with restrictions. *Soviet Math. Dokl.* **7**, 2–75 (1966)
28. G. Qu, N. Li, On the exponentially stability of primal-dual gradient dynamics. *IEEE Control Syst. Lett.* **3**(1), 43–48 (2019)
29. S. Rashid, M.A. Noor, K.I. Noor, Fractional exponentially m -convex functions and inequalities. *Inter. J. Anal. Appl.* **17**(3), 464–478 (2019)
30. D.L. Zu, P. Marcotte, Co-coercivity and its role in the convergence of iterative schemes for solving variational inequalities. *SIAM J. Optim.* **6**(3), 714–726 (1996)

Properties of Exponentially m -Convex Functions



Muhammad Aslam Noor and Khalida Inayat Noor

Abstract In this paper, we define and introduce some new concepts of the exponentially m -convex functions involving a fixed constant $m \in (0, 1]$. We investigate several properties of the exponentially m -convex functions and discuss their relations with convex functions. Optimality conditions are characterized by a class of variational inequalities. Several interesting results characterizing the exponentially m -convex functions are obtained. Results obtained in this paper can be viewed as significant improvement of previously known results.

1 Introduction

Convexity theory describes a broad spectrum of very interesting developments involving a link among various fields of mathematics, physics, economics, and engineering sciences. Some of these developments have made mutually enriching contacts with other fields. This theory provides us several new and powerful techniques to solve a wide class of linear and nonlinear problems. It reveals the fundamental facts on the qualitative behavior of solutions regarding its existence, uniqueness, and regularity to important classes of problems. Convexity also enabled us to develop highly efficient and powerful new numerical methods to solve nonlinear problems, see [1–9, 11–18, 23]. In recent years, various extensions and generalizations of convex functions and convex sets have been considered and studied using innovative ideas and techniques. It is known that more accurate inequalities can be obtained using the logarithmically convex functions than the convex functions. Closely related to the log-convex functions, we have the concept

Nonlinear Analysis and Global Optimization (Edits: Panos Pardalos and Themistocles M. Rassias), Springer Volume.

M. A. Noor (✉) · K. I. Noor
COMSATS University Islamabad, Islamabad, Pakistan

© Springer Nature Switzerland AG 2021
T. M. Rassias, P. M. Pardalos (eds.), *Nonlinear Analysis and Global Optimization*,
Springer Optimization and Its Applications 167,
https://doi.org/10.1007/978-3-030-61732-5_17

of exponentially convex(concave) functions, and the origin of exponentially convex functions can be traced back to Bernstein [6]. Avriel [4] introduced and studied the concept of r -convex functions, where as the (r, p) -convex functions were studied by Antczak [3]. For further properties of the r -convex functions, see Zhao et al. [25] and the references therein. Exponentially convex functions have important applications in information theory, big data analysis, machine learning, and statistic. See [1, 2, 20, 21, 24, 25] and the references therein. Motivated and inspired by the ongoing research in this interesting, applicable, and dynamic field, Noor and Noor [13–15] considered the concept of exponentially convex functions and discussed the basic properties of the exponentially convex functions. It is worth mentioning that these exponentially convex functions[13–15] are distinctly different from the exponentially convex functions considered and studied by Berstein [6], Awan et al.[5], and Pecaric et al.[19, 21]. For example, the definition of exponential convexity in Noor and Noor [13–15] is quite different from Bernstein's since, for example, $F(x) = \log x$ is exponentially convex in Noor's sense but not in Bernstein's since it is not convex. The sum of two exponentially convex functions may not be exponentially convex functions. For example, functions $\log x$ and $-x$ are exponentially convex on $(0, 2)$, but their sum $\log x - x$ is not.

Toader[25] introduced the concept of m -convex sets and m -convex functions, which inspired a great deal of research activities. The m -convex functions unify the convex functions and starlike convex functions. For the properties, generalizations, applications, and other aspects of m -convex functions, see [18].

We would like to mention that the concepts of exponentially convex functions and m -convex are two different generalizations of the convex functions. Motivated by this fact, we introduce a new class of convex functions involving a mixed constant $m \in (0, 1]$, which is called the exponentially m -convex functions. We have shown that the exponentially m -convex(m -concave) functions have nice properties which convex functions enjoy. Several new concepts have been introduced and investigated. We show that the local minimum of the exponentially m -convex functions is the global minimum. The optimal conditions of the differentiable exponentially convex functions can be characterized by a class of variational inequalities, which is itself an interesting outcome of our main results. The difference (sum) of the exponentially convex function and exponentially affine convex function is again an exponentially convex function. The ideas and techniques of this paper may be starting point for further research in these areas.

2 Preliminary Results

Let K be a non-empty closed set in a real inner product space H . We denote by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ the inner product and norm, respectively. Let $F : K \rightarrow H$ be a continuous function. We denote by $m \in (0, 1]$, unless otherwise specified.

Definition 1 ([24]) The set K in H is said to be m -convex set, if

$$(1 - t)u + tmv \in K, \quad \forall u, v \in K, t \in [0, 1].$$

We now introduce new concepts of m -convex functions and exponentially m -convex functions.

Definition 2 A function F is said to be m -convex function, if

$$F((1 - t)u + tmv) \leq (1 - t)F(u) + tF(mv), \quad \forall u, v \in K, \quad t \in [0, 1]. \quad (1)$$

Definition 3 A function F is said to be exponentially m -convex function, if

$$e^{F((1-t)u+tmv)} \leq (1-t)e^{F(u)} + te^{F(mv)}, \quad \forall u, v \in K, \quad t \in [0, 1]. \quad (2)$$

We remark that Definition 3 can be rewritten in the following equivalent form.

Definition 4 A function F is said to be exponentially m -convex function, if

$$e^{F((1-t)u+tmv)} \leq \log[(1-t)e^{F(u)} + te^{F(mv)}], \quad \forall u, v \in K, \quad t \in [0, 1]. \quad (3)$$

A function is called the exponentially m -concave function f , if $-f$ is an exponentially m -convex function. It is obvious that these two concepts are equivalent. These equivalent formulations have been used to discuss various aspects of the exponentially convex functions.

For $m = 1$, we obtain the classes of exponentially convex functions investigated by Antczak[3]. It is worth mentioning that one can also deduce the concept of exponentially convex functions introduced by Bernstein[6] and Avriel [4] from Antczak [3]. For the applications of the exponentially convex functions in the mathematical programming and information theory, see Antczak [3], Alirezai and Mathar[2], and Pal et al. [17]. For the applications of the exponentially concave function in the communication and information theory, we have the following example.

Example [2]: The error function

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

becomes an exponentially concave function in the form $erf(\sqrt{x})$, $x \geq 0$, which describes the bit/symbol error probability of communication systems depending on the square root of the underlying signal-to-noise ratio. This shows that the exponentially concave functions can play important part in communication theory and information theory.

Definition 5 ([2]) A function F is said to be exponentially affine m -convex function, if

$$e^{F((1-t)u+tmv)} = (1-t)e^{F(u)} + te^{F(mv)}, \quad \forall u, v \in K, \quad t \in [0, 1].$$

Definition 6 The function F on the convex set K is said to be exponentially quasi m -convex, if

$$e^{F(u+t(mv-u))} \leq \max\{e^{F(u)}, e^{F(mv)}\}, \quad \forall u, v \in K, \quad t \in [0, 1].$$

Definition 7 The function F on the convex set K is said to be exponentially log m -convex, if

$$e^{F(u+t(mv-u))} \leq (e^{F(u)})^{1-t}(e^{F(v)})^t, \quad \forall u, v \in K, \quad t \in [0, 1].$$

From the above definitions, we have

$$\begin{aligned} e^{F(u+t(mv-u))} &\leq (e^{F(u)})^{1-t}(e^{F(v)})^t \\ &\leq (1-t)e^{F(u)} + te^{F(mv)} \\ &\leq \max\{e^{F(u)}, e^{F(mv)}\}. \end{aligned}$$

This shows that

every exponentially log m -convex function
 \implies exponentially m -convex function \implies exponentially m -convex function.

However, the converse is not true.

Let $K = I = [a, mb]$ be the interval. We now define the exponentially convex functions on I .

Definition 8 Let $I = [a, mb]$. Then, F is an exponentially m -convex function, if and only if,

$$\begin{vmatrix} 1 & 1 & 1 \\ a & x & mb \\ e^{F(a)} & e^{F(x)} & e^{F(mb)} \end{vmatrix} \geq 0; \quad a \leq x \leq b.$$

One can easily show that the following are equivalent:

1. F is an exponentially convex function.
2. $e^{F(x)} \leq e^{F(a)} + \frac{e^{F(mb)} - e^{F(a)}}{mb-a}(x-a)$.
3. $\frac{e^{F(x)} - e^{F(a)}}{x-a} \leq \frac{e^{F(mb)} - e^{F(a)}}{mb-a}$.
4. $(mb-x)e^{F(a)} + (a-mb)e^{F(x)} + (x-a)e^{F(mb)} \geq 0$.
5. $\frac{e^{F(a)}}{(mb-a)(a-x)} + \frac{e^{F(x)}}{(x-mb)(a-x)} + \frac{e^{F(mb)}}{(mb-a)(x-mb)} \leq 0$,

where $x = (1-t)a + tmb \in [a, mb]$.

If F is a differentiable exponentially m -convex function, then

$$e^{F(mv)} - e^{F(u)} \geq \langle e^{F(u)} F'(u), mv - u \rangle, \quad \forall u, mv \in [a, mb].$$

We would like to point that the exponentially m -convex is also Wright strongly m -convex functions. From the definition of exponentially m -convex functions, we have

$$e^{F((1-t)u+tmv)} + e^{F(tu+(1-t)mv)} \leq \{e^F(u) + e^F(mv)\}, \quad \forall u, v \in K, t \in [0, 1],$$

which is called the Wright exponentially m -convex function. It is an interesting problem to study the properties and applications of the Wright exponentially convex functions.

Using the technique of Toader[24], Rashid et al. [22] introduced the following concept of exponentially m -convex functions:

Definition 9 ([22]) A function F is said to be an exponentially m -convex function in the Toader’s sense, if

$$e^{F((1-t)u+tmv)} \leq (1-t)e^{F(u)} + tme^{F(v)}, \quad \forall u, v \in K, \quad t \in [0, 1]. \quad (4)$$

We remark that Definition 9 can be rewritten in the following equivalent form.

Definition 10 ([22]) A function F is said to be an exponentially m -convex function in the Toader’s sense, if

$$e^{F((1-t)u+tmv)} \leq \log[(1-t)e^{F(u)} + tme^{F(v)}], \quad \forall u, v \in K, \quad t \in [0, 1]. \quad (5)$$

We would like to point out that these two concepts defined in Definitions 3 and 9 are equivalent, if the function $e^{F(mv)} = me^{F(v)}$, that is, the function F is exponentially homogeneous. Consequently, all the results proved in this paper can be extended for the exponentially m -convex functions in the Toader’s sense with suitable modifications.

3 Main Results

In this section, we consider some basic properties of generalized strongly convex functions.

Theorem 1 Let $I = [a, b] \subset R$ be an interval containing zero, and let $m \in (0,]$ be a constant. Let $a, b, c \in I$ be points such that $a \leq mc \leq b$. Then, the exponentially m -convex function satisfies the inequality

$$\int_a^b e^{F(x)} dx \leq \frac{mc - a}{2} e^{F(a)} + \frac{b - mc}{2} e^{F(b)} + \frac{b - a}{2} e^{F(mc)}. \quad (6)$$

Proof Assume that $a \leq x \leq mc$. Then, from (4), we have

$$\begin{aligned} \int_a^{mc} e^{F(x)} dx &\leq \int_a^{mc} \frac{mc-x}{mc-a} e^{F(a)} dx + \int_a^{mc} \frac{x-a}{mc-a} e^{F(mc)} dx \\ &= \frac{mc-a}{2} (f(a) + f(mc)). \end{aligned} \tag{7}$$

In a similar way, for $mc \leq x \leq b$, we have

$$\begin{aligned} \int_{mc}^b e^{F(x)} dx &\leq \int_{mc}^b \frac{mc-x}{mc-a} e^{F(mc)} dx + \int_{mc}^b \frac{x-a}{mc-a} e^{F(b)} dx \\ &= \frac{mc-a}{2} (F(b) + F(mc)). \end{aligned} \tag{8}$$

From (7) and (8), we have

$$\begin{aligned} \int_a^b e^{F(x)} dx &= \int_a^{mc} e^{F(x)} dx + \int_{mc}^b e^{F(x)} dx \\ &= \frac{mc-a}{2} (F(a) + F(mc)) + \frac{mc-a}{2} (F(b) + F(mc)) \\ &= \frac{mc-a}{mc-a} e^{F(a)} + \frac{b-mc}{2} e^{F(b)} + \frac{b-a}{2} e^{F(mc)}, \end{aligned}$$

the required (6).

Remark 1 For the interval $I = [a, b]$ containing zero, one can choose a point $c \in [a, b]$ in (6), since $mc \in [a, b]$. Using this information, we can obtain the following inequality for the exponentially m -convex functions for the case $c = a$ or $b = c$.

$$\int_a^b e^{F(x)} dx \leq \frac{mb-a}{2} e^{F(a)} + \frac{b-ma}{2} e^{F(b)},$$

from which, we can have

$$\int_a^b e^{F(x)} dx \leq \frac{b-a}{2} \{e^{F(a)} + e^{F(b)}\}.$$

Theorem 2 Let F be a strictly exponentially m -convex function. Then, any local minimum of F is a global minimum.

Proof Let the exponentially convex function F have a local minimum at $u \in K$. Assume the contrary, that is, $F(mv) < F(u)$ for some $mv \in K$. Since F is exponentially convex, then

$$e^{F(u+t(mv-u))} < te^{F(mv)} + (1-t)e^{F(u)}, \quad \text{for } 0 < t < 1.$$

Thus,

$$e^{F(u+t(mv-u))} - e^{F(u)} < t[e^{F(mv)} - e^{F(u)}] < 0,$$

from which it follows that

$$e^{F(u+t(mv-u))} < e^{F(u)},$$

for arbitrary small $t > 0$, contradicting the local minimum.

Theorem 3 *If the function F on the m -convex set K is exponentially m -convex, then the level set $L_\alpha = \{u \in K : e^{F(u)} \leq \alpha, \alpha \in \mathbb{R}\}$ is a m -convex set.*

Proof Let $u, mv \in L_\alpha$. Then, $e^{F(u)} \leq \alpha$ and $e^{F(mv)} \leq \alpha$. Now, $\forall t \in (0, 1), w = u + t(mv - u) \in K$, since K is a m -convex set. Thus, by the exponentially m -convexity of F , we have

$$\begin{aligned} F e^{F(u+t(mv-u))} &\leq (1-t)e^{F(u)} + te^{F(mv)} \\ &\leq (1-t)\alpha + t\alpha = \alpha, \end{aligned}$$

from which it follows that $u + t(mv - u) \in L_\alpha$. Hence, L_α is a m -convex set.

Theorem 4 *The function F is exponentially m -convex, if and only if,*

$$epi(F) = \{(u, \alpha) : u \in K : e^{F(u)} \leq \alpha, \alpha \in \mathbb{R}\}$$

is a m -convex set.

Proof Assume that F is exponentially convex. Let $(u, \alpha), (mv, \beta) \in epi(F)$. Then, it follows that $e^{F(u)} \leq \alpha$ and $e^{F(mv)} \leq \beta$. Thus, $\forall t \in [0, 1], u, mv \in K$, we have

$$\begin{aligned} e^{F(u+t(mv-u))} &\leq (1-t)e^{F(u)} + te^{F(mv)} \\ &\leq (1-t)\alpha + t\beta, \end{aligned}$$

which implies that

$$(u + t(mv - u), (1-t)\alpha + t\beta) \in epi(F).$$

Thus, $epi(F)$ is a m -convex set. Conversely, let $epi(F)$ be a convex set. Let $u, mv \in K$. Then, $(u, e^{F(u)}) \in epi(F)$ and $(mv, e^{F(mv)}) \in epi(F)$. Since $epi(F)$ is a m -convex set, we must have

$$(u + t(mv - u), (1-t)e^{F(u)} + te^{F(mv)}) \in epi(F),$$

which implies that

$$e^{F(u+t(mv-u))} \leq (1-t)e^{F(u)} + te^{F(mv)}.$$

This shows that F is an exponentially m -convex function.

Theorem 5 *The function F is exponentially quasi m -convex, if and only if, the level set $L_\alpha = \{u \in K, \alpha \in R : e^{F(u)} \leq \alpha\}$ is a m -convex set.*

Proof Let $u, mv \in L_\alpha$. Then, $u, mv \in K$ and $\max(e^{F(u)}, e^{F(mv)}) \leq \alpha$. Now, for $t \in (0, 1)$, $w = u + t(mv - u) \in K$, we have to prove that $u + t(mv - u) \in L_\alpha$. By the exponentially quasi m -convexity of F , we have

$$e^{F(u+t(mv-u))} \leq \max(e^{F(u)}, e^{F(mv)}) \leq \alpha,$$

which implies that $u + t(mv - u) \in L_\alpha$, showing that the level set L_α is indeed a m -convex set.

Conversely, assume that L_α is a m -convex set. Then, for any $u, mv \in L_\alpha, t \in [0, 1]$, $u + t(mv - u) \in L_\alpha$. Let $u, mv \in L_\alpha$ for

$$\alpha = \max(e^{F(u)}, e^{F(mv)}) \quad \text{and} \quad e^{F(mv)} \leq e^{F(u)}.$$

Then, from the definition of the level set L_α , it follows that

$$e^{F(u+t(mv-u))} \leq \max(e^{F(u)}, e^{F(mv)}) \leq \alpha.$$

Thus, F is an exponentially quasi m -convex function. This completes the proof.

Theorem 6 *Let F be an exponentially m -convex function. Let $\mu = \inf_{u \in K} F(u)$. Then, the set $E = \{u \in K : e^{F(u)} = \mu\}$ is a m -convex set of K . If F is strictly exponentially m -convex, then E is a singleton.*

Proof Let $u, mv \in E$. For $0 < t < 1$, let $w = u + t(mv - u)$. Since F is an exponentially m -convex function, then

$$\begin{aligned} F(w) &= e^{F(u+t(mv-u))} \leq (1-t)e^{F(u)} + te^{F(mv)} \\ &= t\mu + (1-t)\mu = \mu, \end{aligned}$$

which implies that $w \in E$, and hence E is a m -convex set. For the second part, assume to the contrary that $F(u) = F(mv) = \mu$. Since K is a m -convex set, then for $0 < t < 1$, $u + t(mv - u) \in K$. Furthermore, since F is strictly exponentially m -convex,

$$\begin{aligned} e^{F(u+t(mv-u))} &< (1-t)e^{F(u)} + te^{F(mv)} \\ &= (1-t)\mu + t\mu = \mu. \end{aligned}$$

This contradicts the fact that $\mu = \inf_{u \in K} F(u)$, and hence the result follows.

Theorem 7 *If F is an exponentially m -convex function such that $e^{F(mv)} < e^{F(u)}$, $\forall u, mv \in K$, then F is a strictly exponentially quasi m -convex function.*

Proof By the exponentially m -convexity of the function F , $\forall u, mv \in K, m, t \in [0, 1]$, we have

$$e^{F(u+t(mv-u))} \leq (1-t)e^{F(u)} + te^{F(mv)} < e^F(u),$$

since $e^{F(mv)} < e^{F(u)}$, which shows that the function F is strictly exponentially quasi m -convex.

We now discuss some properties of the differentiable exponentially m -convex functions.

Theorem 8 *Let F be a differentiable function on the m -convex set K . Then, the function F is an exponentially m -convex function, if and only if,*

$$e^{F(mv)} - e^{F(u)} \geq \langle e^{F(u)} F'(u), mv - u \rangle, \quad \forall mv, u \in K. \tag{9}$$

Proof Let F be an exponentially m -convex function. Then,

$$e^{F(u+t(mv-u))} \leq (1-t)e^{F(u)} + te^{F(mv)}, \quad \forall u, mv \in K,$$

which can be written as

$$e^{F(mv)} - e^{F(u)} \geq \left\{ \frac{e^{F(u+t(mv-u))} - e^{F(u)}}{t} \right\}.$$

Taking the limit in the above inequality as $t \rightarrow 0$, we have

$$e^{F(mv)} - e^{F(u)} \geq \langle e^{F(u)} F'(u), mv - u \rangle,$$

which is (9), the required result.

Conversely, let (9) hold. Then,

$\forall u, mv \in K, t \in [0, 1], v_t = u + t(mv - u) \in K$, we have

$$\begin{aligned} e^{F(mv)} - e^{F(v_t)} &\geq \langle e^{F(v_t)} F'(v_t), mv - v_t \rangle \\ &= (1-t) \langle e^{F(v_t)} F'(v_t), mv - u \rangle. \end{aligned} \tag{10}$$

In a similar way, we have

$$\begin{aligned} e^{F(u)} - e^{F(v_t)} &\geq \langle e^{F(v_t)} F'(v_t), u - v_t \rangle \\ &= -t \langle e^{F(v_t)} F'(v_t), mv - u \rangle. \end{aligned} \tag{11}$$

Multiplying (10) by t and (11) by $(1 - t)$ and adding the resultant, we have

$$e^{F(u+t(mv-u))} \leq (1 - t)e^{F(u)} + te^{F(mv)},$$

showing that F is an exponentially m -convex function.

Remark 2 From (9), we have

$$e^{F(mv)-F(u)} - 1 \geq \langle F'(u), mv - u \rangle, \quad \forall mv, u \in K,$$

which can be written as

$$F(mv) - F(u) \geq \log\{1 + \langle F'(u), mv - u \rangle\} \quad \forall mv, u \in K. \tag{12}$$

Changing the role of u and mv in (12), we also have

$$F(u) - F(mv) \geq \log\{1 + \langle F'(mv), u - mv \rangle\} \quad \forall v, u \in K. \tag{13}$$

Adding (12) and (13), we have

$$\langle F'(u) - F'(mv), u - mv \rangle \geq (\langle F'(u)u - mv \rangle)(\langle F'(mv), u - mv \rangle),$$

which expresses the monotonicity of the differential $F'(\cdot)$ of the exponentially m -convex function.

Theorem 8 enables us to introduce the concept of the exponentially m -monotone operators, which appears to be new ones.

Definition 11 The differential $F'(\cdot)$ is said to be exponentially m -monotone, if

$$\langle e^{F(u)} F'(u) - e^{F(mv)} F'(mv), u - mv \rangle \geq 0, \quad \forall u, mv \in H.$$

Definition 12 The differential $F'(\cdot)$ is said to be exponentially pseudo m -monotone, if

$$\langle e^{F(u)} F'(u), mv - u \rangle \geq 0, \quad \Rightarrow \langle e^{F(mv)} F'(mv), mv - u \rangle \geq 0, \quad \forall u, mv \in H.$$

From these definitions, it follows that exponentially monotonicity implies exponentially pseudo m -monotonicity, but the converse is not true.

Theorem 9 Let F be differentiable on the m -convex set K . Then, (9) holds, if and only if, $F'(\cdot)$ satisfies

$$\langle e^{F(u)} F'(u) - e^{F(mv)} F'(mv), u - mv \rangle \geq 0, \quad \forall u, mv \in K. \tag{14}$$

Proof Let F be an exponentially convex function on the m -convex set K . Then, from Theorem 3.1, we have

$$e^{F(mv)} - e^{F(u)} \geq \langle e^{F(u)} F'(u), mv - u \rangle, \quad \forall u, mv \in K. \tag{15}$$

Changing the role of u and mv in (15), we have

$$e^{F(u)} - e^{F(mv)} \geq \langle e^{F(mv)} F'(mv), u - mv \rangle, \quad \forall u, mv \in K. \tag{16}$$

Adding (15) and (16), we have

$$\langle e^{F(u)} F'(u) - e^{F(mv)} F'(mv), u - mv \rangle \geq 0,$$

which shows that F' is exponentially monotone.

Conversely, from (14), we have

$$\langle e^{F(mv)} F'(mv), u - mv \rangle \leq \langle e^{F(u)} F'(u), u - mv \rangle. \tag{17}$$

Since K is a m -convex set, $\forall u, mv \in K, \quad t \in [0, 1] \quad v_t = u + t(mv - u) \in K$.

Taking $v = v_t$ in (17), we have

$$\begin{aligned} \langle e^{F(v_t)} F'(v_t), u - v_t \rangle &\leq \langle e^{F(u)} F'(u), u - v_t \rangle \\ &= -t \langle e^{F(u)} F'(u), mv - u \rangle, \end{aligned}$$

which implies that

$$\langle e^{F(v_t)} F'(v_t), mv - u \rangle \geq \langle e^{F(u)} F'(u), v - u \rangle. \tag{18}$$

Consider the auxiliary function

$$g(t) = e^{F(u+t(mv-u))},$$

from which, we have

$$g(1) = e^{F(mv)}, \quad g(0) = e^{F(u)}.$$

Then, from (18), we have

$$g'(t) = \langle e^{F(v_t)} F'(v_t), mv - u \rangle \geq \langle e^{F(u)} F'(u), mv - u \rangle. \tag{19}$$

Integrating (19) between 0 and 1, we have

$$g(1) - g(0) = \int_0^1 g'(t) dt \geq \langle e^{F(u)} F'(u), mv - u \rangle.$$

Thus, it follows that

$$e^{F(mv)} - e^{F(u)} \geq \langle e^{F(u)} F'(u), mv - u \rangle,$$

which is the required (9).

We now give a necessary condition for exponentially pseudo m -convex functions.

Theorem 10 *Let F' be exponentially pseudo m -monotone. Then, F is an exponentially pseudo m -convex function.*

Proof Let $F'(\cdot)$ be an exponentially pseudo m -monotone function. Then, $\forall u, mv \in K$,

$$\langle e^{F(u)} F'(u), mv - u \rangle \geq 0.$$

implies that

$$\langle e^{F(mv)} F'(mv), mv - u \rangle \geq 0. \tag{20}$$

Since K is a m -convex set, $\forall u, mv \in K, \quad t \in [0, 1], v_t = u + t(mv - u) \in K$.

Taking $v = v_t$ in (20), we have

$$\langle e^{F(v_t)} F'(v_t), mv - u \rangle \geq 0. \tag{21}$$

Consider the auxiliary function

$$g(t) = e^{F(u+t(mv-u))} = e^{F(v_t)}, \quad \forall u, mv \in K, \quad t \in [0, 1],$$

which is differentiable, since F is a differentiable function. Then, using (21), we have

$$g'(t) = \langle e^{F(v_t)} F'(v_t), mv - u \rangle \geq 0.$$

Integrating the above relation between 0 and 1, we have

$$g(1) - g(0) = \int_0^1 g'(t) dt \geq 0,$$

that is,

$$e^{F(mv)} - e^{F(u)} \geq 0,$$

showing that F is an exponentially pseudo m -convex function.

Definition 13 The function F is said to be sharply exponentially pseudo m -convex, if there exists a constant $\mu > 0$ such that

$$\begin{aligned} \langle e^{F(u)} F'(u), mv - u \rangle &\geq 0 \\ \Rightarrow \\ e^{F(mv)} &\geq e^{F(u+t(mv-u))}, \quad \forall u, mv \in K, \quad t \in [0, 1]. \end{aligned}$$

Theorem 11 Let F be a sharply exponentially pseudo m -convex function on K . Then,

$$\langle e^{F(mv)} F'(mv), mv - u \rangle \geq 0, \quad \forall u, mv \in K.$$

Proof Let F be a sharply exponentially pseudo m -convex function on K . Then,

$$e^{F(mv)} \geq e^{F(u+t(mv-u))}, \quad \forall u, mv \in K, \quad t \in [0, 1],$$

from which, we have

$$\begin{aligned} 0 &\leq \lim_{t \rightarrow 0} \left\{ \frac{e^{F(u+t(mv-u))} - e^{F(mv)}}{t} \right\} \\ &= \langle e^{F(mv)} F'(mv), mv - u \rangle, \end{aligned}$$

the required result.

Definition 14 A function F is said to be a pseudo m -convex function, if there exists a strictly positive bifunction $b(., .)$, such that

$$\begin{aligned} e^{F(mv)} &< e^{F(u)} \\ \Rightarrow \\ e^{F(u+t(mv-u))} &< e^{F(u)} + t(t-1)b(mv, u), \quad \forall u, mv \in K, \quad t \in [0, 1]. \end{aligned}$$

Theorem 12 If the function F is exponentially convex function such that $e^{F(v)} < e^{F(u)}$, then the function F is exponentially pseudo convex.

Proof Since $e^{F(v)} < e^{F(u)}$ and F is an exponentially convex function, then $\forall u, v \in K, \quad t \in [0, 1]$, we have

$$\begin{aligned} e^{F(u+(1-t)v)} &\leq e^{F(u)} + t(e^{F(mv)} - e^{F(u)}) \\ &< e^{F(u)} + t(1-t)(e^{F(mv)} - e^{F(u)}) \\ &= e^{F(u)} + t(t-1)(e^{F(u)} - e^{F(mv)}) \\ &< e^{F(u)} + t(t-1)b(u, mv), \end{aligned}$$

where $b(u, v) = e^{F(u)} - e^{F(mv)} > 0$, the required result. This shows that the function F is an exponentially m -convex function.

We now discuss the optimality condition for the differentiable exponentially convex functions, which is the main motivation of our next result.

Theorem 13 *Let F be a differentiable exponentially m -convex function. Then, $u \in K$ is the minimum of the function F , if and only if, $u \in K$ satisfies the inequality*

$$\langle e^{F(u)} F'(u), mv - u \rangle \geq 0, \quad \forall u, mv \in K. \tag{22}$$

Proof Let $u \in K$ be a minimum of the function F . Then,

$$F(u) \leq F(mv), \forall mv \in K,$$

from which, we have

$$e^{F(u)} \leq e^{F(mv)}, \forall mv \in K. \tag{23}$$

Since K is a m -convex set, then, $\forall u, mv \in K, \quad t \in [0, 1]$,

$$v_t = (1 - t)u + tmv \in K.$$

Taking $v = v_t$ in (23), we have

$$0 \leq \lim_{t \rightarrow 0} \left\{ \frac{e^{F(u+t(mv-u))} - e^{F(u)}}{t} \right\} = \langle e^{F(u)} F'(u), mv - u \rangle. \tag{24}$$

Since F is differentiable exponentially m -convex function, then

$$e^{F(u+t(mv-u))} \leq e^{F(u)} + t(e^{F(mv)} - e^{F(u)}), \quad u, mv \in K, \quad t \in [0, 1],$$

from which, using (24), we have

$$\begin{aligned} e^{F(mv)} - e^{F(u)} &\geq \lim_{t \rightarrow 0} \left\{ \frac{e^{F(u+t(mv-u))} - e^{F(u)}}{t} \right\} \\ &= \langle e^{F(u)} F'(u), mv - u \rangle \geq 0, \end{aligned}$$

from which, we have

$$e^{F(mv)} - e^{F(u)} \geq 0,$$

which implies that

$$F(u) \leq F(mv), \quad \forall mv \in K.$$

This shows that $u \in K$ is the minimum of the differentiable exponentially m -convex function, the required result.

Remark 3 The inequality of the type (22) is called the exponentially variational inequality and appears to be new one. For the applications, formulations, numerical methods, and other aspects of variational inequalities, see Noor [10, 11] and Noor et al.[16].

We now show that the difference of exponentially convex function and exponentially affine convex function is again an exponentially convex function.

Theorem 14 *Let f be an exponentially affine m -convex function. Then, F is an exponentially m -convex function, if and only if, $g = F - f$ is an exponentially m -convex function.*

Proof Let f be an exponentially affine convex function. Then,

$$e^{f((1-t)u+tmv)} = (1-t)e^{f(u)} + te^{f(mv)}, \quad \forall u, mv \in K, \quad t \in [0, 1]. \quad (25)$$

From the exponentially m -convexity of F , we have

$$e^{F((1-t)u+tmv)} \leq (1-t)e^{F(u)} + te^{F(mv)}, \quad \forall u, mv \in K, \quad t \in [0, 1]. \quad (26)$$

From (25) and (26), we have

$$e^{F((1-t)u+tmv)} - e^{f((1-t)u+tmv)} \leq (1-t)(e^{F(u)} - e^{f(u)}) + t(e^{F(mv)} - e^{f(mv)}), \quad (27)$$

from which it follows that

$$\begin{aligned} e^{g((1-t)u+tmv)} &= e^{F((1-t)u+tmv)} - e^{f((1-t)u+tmv)} \\ &\leq (1-t)(e^{F(u)} - e^{f(u)}) + t(e^{F(mv)} - e^{f(mv)}), \end{aligned}$$

which shows that $g = F - f$ is an exponentially m -convex function.

The inverse implication is obvious.

4 Conclusion

In this paper, we have introduced and studied a new class of convex functions, which is called the exponentially m -convex function. It has been shown that exponentially m -convex functions enjoy several properties which convex functions have. We have shown that the minimum of the differentiable exponentially m -convex functions can be characterized by a new class of variational inequalities, which is called the exponentially variational inequality. One can explore the applications of the exponentially variational inequalities. This may stimulate further research.

Acknowledgments The authors would like to thank the Rector, COMSATS University Islamabad, Islamabad, Pakistan, for providing excellent research and academic environments. Authors are grateful to Prof. Dr. Themistocles M. Rassias for his kind invitation and support.

References

1. N.I. Akhiezer, *The Classical Moment Problem and Some Related Questions in Analysis* (Oliver and Boyd, Edinburgh, 1965)
2. G. Alirezaei, R. Mazhar, On exponentially concave functions and their impact in information theory. *J. Inform. Theory Appl.* **9**(5), 265–274 (2018)
3. T. Antczak, On (p, r) -invex sets and functions. *J. Math. Anal. Appl.* **263**, 355–379 (2001)
4. M. Avriel, r -Convex functions. *Math. Program.* **2**, 309–323 (1972)
5. M.U. Awan, M.A. Noor, K.I. Noor, Hermite-Hadamard inequalities for exponentially convex functions. *Appl. Math. Inform. Sci.* **12**(2), 405–409 (2018)
6. S.N. Bernstein, Sur les fonctions absolument monotones. *Acta Math.* **52**, 1–66 (1929)
7. G. Cristescu, L. Lupsa, *Non Connected Convexities and Applications* (Kluwer Academic Publisher, Dordrecht, 2002)
8. S.S. Dragomir, I. Gomm, Some Hermite-Hadamard type inequalities for functions whose exponentials are convex. *Stud. Univ. Babeş-Bolyai Math.* **60**(4), 527–534 (2015)
9. C.F. Niculescu, L.E. Persson, *Convex Functions and Their Applications* (Springer, New York, 2018)
10. M.A. Noor, New approximation schemes for general variational inequalities. *J. Math. Anal. Appl.* **251**, 217–229 (2000)
11. M.A. Noor, Some developments in general variational inequalities. *Appl. Math. Comput.* **152**, 199–277 (2004)
12. M.A. Noor, K.I. Noor, Exponentially convex functions. *J. Orisa Math. Soc.* **39**(2019)
13. M.A. Noor, K.I. Noor, Strongly exponentially convex functions. *U.P.B. Bull. Sci. Appl. Math. Series A.* **81**(4), 75–84 (2019)
14. M.A. Noor, K.I. Noor, Strongly exponentially convex functions and their properties. *J. Adv. Math. Stud.* **9**(2) (2019)
15. M.A. Noor, K.I. Noor, On generalized strongly convex functions involving bifunction. *Appl. Math. Inform. Sci.* **13**(3), 411–416 (2019)
16. M.A. Noor, K.I. Noor, Th. M. Rassias, Some aspects of variational inequalities. *J. Comput. Appl. Math.* **47**, 285–312 (1993)
17. S. Pal, T.K. Wong, On exponentially concave functions and a new information geometry. *Annal. Prob.* **46**(2), 1070–1113 (2018)
18. Z. Pavic, A. Ardic, The most important inequalities of m -convex functions. *Turkish J. Math.* **41**, 625–635 (2017)
19. J. Pecaric, J. Jaksetic, On exponential convexity, Euler-Radau expansions and Stolarsky means. *Rad Hrvat. Matematike Znanosti* **17**(515), 81–94 (2013)
20. J. Pecaric, F. Proschan, Y.L. Tong, *Convex Functions, Partial Orderings and Statistical Applications* (Academic Press, New York, 1992)
21. J. Pecaric, C.E.M. Pearce, V. Simic, Stolarsky means and Hadamard's inequality. *J. Math. Anal. Appl.* **220**, 99–109 (1998)
22. S. Rashid, M.A. Noor, K.I. Noor, Fractional exponentially m -convex functions and inequalities. *Inter. J. Anal. Appl.* **17**(3), 464–478 (2019)
23. S. Rashid, M.A. Noor, K.I. Noor, A.O. Akdemir, Some new generalizations for exponentially s -convex functions and inequalities via fractional operators. *Fractal. Fract.* **3**(24), 1–16 (2019)
24. G.H. Toader, Some generalizations of the convexity, in *Proceedings of the Colloquium on Approximation and Optimization* (1984), pp. 329–338
25. Y.X. Zhao, S.Y. Wang, L. Coladas Uria, Characterizations of r -convex functions. *J. Optim. Theory Appl.* **145**, 186–195 (2010)

Natural vs. Artificial Topologies on a Relativistic Spacetime



Kyriakos Papadopoulos

Abstract Consider a set M equipped with a structure $*$. We call a natural topology T_* , on $(M, *)$, the topology induced by $*$. For example, a natural topology for a metric space (X, d) is a topology T_d induced by the metric d , and for a linearly ordered set $(X, <)$, a natural topology should be the topology $T_<$ that is induced by the order $<$. This fundamental property, for a topology to be called “natural,” has been largely ignored while studying topological properties of spacetime manifolds (M, g) , where g is the Lorentz “metric,” and the manifold topology T_M has been used as a natural topology, ignoring the spacetime “metric” g . In this survey, we review critically candidate topologies for a relativistic spacetime manifold, and we pose open questions and conjectures with the aim to establish a complete guide on the latest results in the field and give the foundations for future discussions. We discuss the criticism against the manifold topology, a criticism that was initiated by people like Zeeman, Göbel, Hawking-King-McCarthy and others, and we examine what should be meant by the term “natural topology” for a spacetime. Since the common criticism against spacetime topologies, other than the manifold topology, claims that there has not been established yet a physical theory to justify such topologies, we give examples of seemingly physical phenomena, under the manifold topology, which are actually purely effects depending on the choice of the topology; the Limit Curve Theorem, which is linked to singularity theorems in general relativity, and the Gao–Wald type of “time dilation” are such examples.

1 Motivation: The Topologization Problem

Almost six decades from the first papers on Einstein’s theory of relativity, and simultaneously with the appearance of the first results on spacetime singularities, a freshly new discussion was initiated on whether the manifold topology should be

K. Papadopoulos (✉)

Department of Mathematics, Kuwait University, Kuwait City, Kuwait

© Springer Nature Switzerland AG 2021

T. M. Rassias, P. M. Pardalos (eds.), *Nonlinear Analysis and Global Optimization*,

Springer Optimization and Its Applications 167,

https://doi.org/10.1007/978-3-030-61732-5_18

called a *natural* topology for a spacetime or not. A spacetime (M, g) , in general relativity, is a four-dimensional, time-oriented, connected, C^d manifold, which is equipped with a C^{d-1} Lorentz “metric” g ¹ (see, for example, [23]). Thus, the problem of assigning a spacetime (M, g) to a natural topology should take into account the Lorentz tensor field g . This idea lies on the principle that if one considers a set M equipped with a structure $*$, then a natural topology T_M (or T_*), on $(M, *)$, should be induced by $*$; otherwise, such a topology cannot be called a natural topology on $(M, *)$.²

A serious problem that appears when one uses the manifold topology as a natural topology T_M , for a spacetime (M, g) , is that T_M is a natural topology for the manifold M , as it is induced by the metric structure of the manifold, but it is not natural in (M, g) , where g is the Lorentz “metric.” As a consequence, the manifold topology does not incorporate the causal structure of the spacetime and, under this topology, the spacetime itself carries properties that might not be as natural as we once thought to be. In Section 2, we will review the obvious differences between T_M and appropriate candidates for a spacetime topology and how the properties of T_M are incompatible with the structure of light cone, a structure which corresponds to each point in the spacetime. In Section 3, we will mention issues related to the *singularity problem* in general relativity and how the choice of an appropriate natural topology might influence the way that we view singularities. We will extend the discussion to naked singularities and the world of wormholes, all in the frame of spacetime topology. In Section 4, we will see a mysterious duality between spacelike and timelike, with respect to two dual order relations in a spacetime, each of which induces a topology which is dual, in a particular sense, to the other. This section, as well as the previous one, will give a lot of space for questions that we will list in the concluding Section 5.

There is a general confusion of the meaning of the term “natural,” in topology, and this has led to a sequence of misunderstandings in the field of spacetime geometry. In a discussion like this one, a topology is not just a tool, but something vital for the description of a spacetime as a mathematical entity. It is evident that under appropriate topologies, a spacetime cannot admit singularities and several other effects, including, for example, the Gao–Wald “time-dilation,” which is related to a property called causal pseudo-convexity; such effects are a result of the exclusive use of the manifold metric topology instead of a topology which embodies the causal and conformal structure of spacetime. A finer topology, than the manifold one, might not be related in a straightforward way to the metric structure of the manifold (as it might not be metrizable), but it contains coarser topologies, such as the usual manifold metric topology, which can do this job.

¹The term metric, for the Lorentz tensor field, is an abuse of language, as was also pointed by Zeeman in [29], but it is so widely used that we will put it in quotes, in order to distinguish from the Riemannian metric.

²This problem, in the case of the orderability of a set is addressed in [11] and [12].

The Riemannian metric itself has proven to have a significance in theories like the Wick rotation (for a critical review on this topic, see, for example, Penrose [24]), but a topology that is induced by the Riemannian metric is far from being called natural in a spacetime. In this chapter, we will restrict our entire discussion to general relativity, and even if we are against the use of the term “natural” for the manifold topology, we should highlight that an appropriate Riemannian metric will still play a significant role in the construction of (really) natural spacetime topologies, different from the manifold one.

2 What Is (or Should be) the Role of Spacetime Topology?

In order to answer the question of the title in this section, we first need to list properties of the manifold topology T_M that make it an inappropriate choice for a natural topology for a spacetime M . Zeeman, in 1967 (see [29]), pointed that

1. The Minkowski space, (M, g) , has $M = \mathbb{R}^4$, and under the Euclidean topology $T_{\mathbb{R}^4}$, on \mathbb{R}^4 , it is locally homogeneous (in the sense that it looks, topologically, the same at any point). The Minkowski space is not just the set M though; it is the pair (M, g) and this is not a locally homogeneous space; at each point, there corresponds a light cone, which separates spacelike from timelike vectors.
2. The group of all homeomorphisms of (M, g) under the Euclidean topology $T_{\mathbb{R}^4}$ is vast and has no known physical meaning. An appropriate topology should associate the group of homeomorphisms with the Lorentz group and dilatations.

Göbel, in 1976, generalized the arguments of Zeeman for curved spacetimes, highlighting that the manifold topology (the analogue of the Euclidean topology in the case of the flat Minkowski space) is artificial both in a mathematical and in a physical sense. He added that experts were primarily concerned with Riemannian structures, where the manifold topology is indeed natural, and not with spaces with a pseudo-Riemannian metric (Lorentz metric is a particular example). It is rather interesting the comment that Göbel adds that it is not plausible to consider a spacetime as locally Euclidean and there is no justification why it should be: “There are no experiments known to justify a Euclidean topology along lightlike geodesics.”

So, Zeeman, as a solution to the problems that he pointed out, came up with a topology that mimics the Euclidean space \mathbb{R}^4 , in the sense that it induces the one-dimensional Euclidean topology on \mathbb{R} and the three-dimensional Euclidean topology on \mathbb{R}^3 . He named this topology “the Fine topology F on Minkowski space M ” and defined it to be the *finest* topology on M , which induces the one-dimensional Euclidean topology on every time axis and the three-dimensional Euclidean topology on every space axis. Zeeman’s intuition worked pretty successfully, since he proved that, under F , the group of homeomorphisms of M is the Lorentz group with translations and dilatations, a significant result, indeed.

Göbel (see [6]) extended Zeeman’s result to general relativity, by giving the definition of the analogue of F : let M be a spacetime manifold, T_M its manifold

topology, and let S be a collection of subsets of M . A set $A \subset M$ is open in $Z(S, T_M)$, a topology in the class $\mathfrak{Z} - \mathfrak{G}$ of Zeeman-Göbel, if $A \cap B$ is open in $(B, T_M|B)$, the subspace topology of the manifold topology (M, T_M) with respect to (B, T_M) , for all $B \in S$. The finest such topology, call it \mathcal{F} , is the general relativistic analogue of F . Under \mathcal{F} , and without any restrictions on the spacetime M , Göbel showed that the group of all homeomorphisms of M is the group of all homothetic transformations of M , leading to the fact that a homeomorphism, under \mathcal{F} , is an isometry.

Hawking, King and McCarthy (and in communication with Göbel) in [7] emphasized that the standard manifold topology merely characterizes continuity properties and proposed a topology that determines the causal, differential and conformal structures of spacetime but criticized Zeeman-Göbel topologies $\mathfrak{Z} - \mathfrak{G}$ of having the following disadvantages:

1. A three-dimensional section of simultaneity has no meaning in terms of physically plausible experiments.
2. While the isometry and conformal groups of M are significantly physical, it is not necessarily clear that this is so for the homothety group of M .
3. F is technically complicated; in particular, the fact that no point has a countable neighbourhood basis makes F hard to calculate with.

We believe that point number 3, of Hawking-King-McCarthy, is not so fruitful; one cannot expect to have a natural topology (as we defined the term “natural” in Section 1) and simultaneously “easy to use”; if the topology is difficult to handle with, this can be due to the complicated structure of the universe set in which the topology is defined.

The topology that Hawking-King-McCarthy proposed is widely known as the *Path topology* on a spacetime and is defined as follows. For each $x \in M$ and each open neighbourhood U of x , let $I(p, U)$ denote the set of points connected to p by a timelike path lying in U and by $K(p, U)$ the set $I(p, U) \cup \{x\}$. By choosing an arbitrary Riemannian metric h on M , let $B_\epsilon(x)$ denote an open ball centred at x with radius $\epsilon > 0$, with respect to h . The *Path topology* \mathcal{P} , on M , is defined to be the finest topology such that the induced topology on every timelike curve coincides with the topology induced from the manifold topology. Hawking et al. proved that the sets of the form $K(p, U) \cap B_\epsilon(x)$ form a basis for the topology \mathcal{P} , giving to \mathcal{P} properties, like \mathcal{P} has an explicit neighbourhood basis, \mathcal{P} is strictly finer than T_M and incomparable to \mathcal{F} and the \mathcal{P} -continuous paths are Feynman paths (for proofs of these statements, see [7]), and overall advantages, like \mathcal{P} determines both the causal, differential and conformal structure of M , making calculations linked to these structures purely topological.

Low has shown that the Limit Curve Theorem (LCM) does not hold under \mathcal{P} , and because of this result, he considered \mathcal{P} as a not fruitful topology (for details, see [9]). We have a bit of a disagreement on this conclusion, and we will discuss about it, in particular, in the next section.

A list of people³ have studied different topologies in the class $\mathfrak{J} - \mathfrak{G}$, using tools from general topology. There is a little concern about this study: even if it is interesting to know the topological properties of several Zeeman–Göbel topologies, there is a lack of unity in notation and a common motivation is absent, throughout the existing literature; there are scattered results on whether a separation axiom is satisfied or not, results with respect to connectedness, metrizable, etc., but there is a lack of a main question. The question, in our opinion, should be not to simply find alternative “better”⁴ topologies to the manifold topology T_M , but to justify which is the most *natural* topology for a spacetime manifold. There is an obvious qualitative difference between the two approaches.

As an example of this general problem, we mention the Fermat Real Line $\bullet\mathbb{R}$, which was defined by Giordano and Kunzinger⁵ as a possible alternative to Synthetic Differential Geometry, aiming to develop new foundations of smooth differential geometry for finite- and infinite-dimensional spaces. Two different topologies were introduced on this line, the so-called “Omega Topology” and the “Fermat Topology”; the first topology is generated by a complete metric and is linked to the differentiation of smooth functions on infinitesimals, and the latter one is generated by a complete pseudo-metric and is linked to the differentiation of non-standard smooth functions. Both topologies play a different role, but none of them is a natural topology for $\bullet\mathbb{R}$; a linearly ordered set should be assigned to its natural topology which is induced by the order. So, it is easy for a confusion about which properties are “natural” to appear; for example, continuity properties, under a topology different from the natural topology, might not hold within the natural topology. A simple example that illustrates this issue in spacetime geometry is given by the Zeno sequences, in [29].

In the sequence of papers, [1, 16, 17, 19] and [20], the authors aim to establish a common background for the topologization problem of a spacetime. This background is the Lorentz “metric” and the structure of the light cone, where one can define the chronological order \ll , the causal order \prec , the relation horismos \rightarrow and also the chorological order $<$; for the last one, see in particular [1], and for a complete list of relations R depending on the light cone, see [20]. One can use the following weak version of the *interval topology*, in order to get the induced topology from such a relation R on a spacetime M . For a set X , consider the sets $I^+(x) = \{y \in X : xRy\}$ and $I^-(x) = \{y \in X : yRx\}$, as well as the collections $\mathcal{S}^+ = \{X \setminus I^-(x) : x \in X\}$ and $\mathcal{S}^- = \{X \setminus I^+(x) : x \in X\}$. A basic-open set U in the weak interval topology T^{in} is defined as $U = A \cap B$, where $A \in \mathcal{S}^+$ and $B \in \mathcal{S}^-$; in other words, $\mathcal{S}^+ \cup \mathcal{S}^-$ forms a subbase for T^{in} . Such topologies were constructed in [16, 19] and [1], covering the cases of

³For example, Nada, Agarwal, Shrivastava, Dossena and Williams; for a complete list of names and articles, see [26].

⁴“Better” in a topological sense: that is, topologies easier to work with and rich in topological properties.

⁵For a short survey, see Section 5, from [13].

horismos, chronology, causality and chorology (which are lightlike, timelike, causal and spacelike relations, respectively). Such topologies belong to the class $\mathfrak{J} - \mathfrak{G}$, as we have shown in [19]. The seemingly real problem that for each point there exists, for each of these topologies, respectively, a local base of unbounded open sets, is solved, by considering the least topology that contains both the manifold topology and a topology T^{in} ; this topology is called the *join topology* or, as it was misnamed by Reed in [25], the “intersection topology” between two given topologies and is defined to be the topology with base $\{U_1 \cap U_2 : U_1 \in T_1 \text{ and } U_2 \in T_2\}$, where T_1 and T_2 are topologies on some set X . One can use De Morgan’s laws to show that a base for the join topology can also be given by $\{U_1 \cap U_2 : U_1 \in B_1 \text{ and } U_2 \in B_2\}$, where B_1 is a base for the topology T_1 and B_2 is a base for the topology T_2 . In [1], we have shown that the join topology between T_M and the weak interval topology which is induced by the reflexive chorological order \leq is actually the Path topology of Hawking–King–McCarthy which, in turn, belongs to the class $\mathfrak{J} - \mathfrak{G}$ and has, locally, an order structure. There is a kind of a dual such topology, studied in [16], which is the join topology between T_M and the reflexive chronological order; this topology, again, has a locally ordered structure.

We now have enough information to dig a bit deeper in the subject and talk about spacetime singularities.

3 Singularities, Naked Singularities and a Kind of Unexpected Gravitational Time Delay Effects

“Time stays long enough for anyone who will use it.”—Leonardo da Vinci

In the previous section, we discussed the role of spacetime topology, as a part of the structure of spacetime, and we stressed that, if one sees a spacetime as a mathematical entity, the spacetime topology should be natural. Since the structure of null cone cannot be recovered by the manifold topology,⁶ we have excluded the manifold topology as a natural candidate topology for a spacetime. There are more serious issues though, in this discussion, that should not be neglected. For example, the Path topology \mathcal{P} on a spacetime manifold M is finer than the manifold topology T_M , it belongs to the class $\mathfrak{J} - \mathfrak{G}$ and it has locally an order structure that connects it with the time cone, but the Limit Curve Theorem (LCT) does not hold under \mathcal{P} (see [9] and for a further discussion, [17]). It is evident that the singularity problem depends on the spacetime topology; one can support this, by looking, for example, the use of the LCT in basic singularity theorems (see [10, 28] as well as [22]). In particular, the LCT, under the manifold topology, states that if γ_n is a sequence of

⁶We refer, again, to [29] for a rigorous proof.

causal curves, x_n is a point on γ_n , for each n , and if x is a limit point of $\{x_n\}$, then there is an endless causal curve γ , passing through x , which is a limit curve of the sequence γ_n . The failure of this theorem to hold is very important, because it avoids basic contradiction arguments that are present in the proofs of (in our knowledge) all singularity theorems. The fact that the LCT holds under T_M does not make the manifold topology a natural topology though. The failure of LCT to hold under a more proper spacetime topology, like \mathcal{P} , for example, should ring a bell about the appearance of singularities in the basic singularity theorems: do these singularity theorems depend exclusively from the use of the manifold topology? Are they a purely topological effect that sieges to exist if one considers a more appropriate topology?

The above question has almost certainly a positive answer for classical singularity theorems like in [22]. This is not so obvious though, at least for the case of naked singularities, if one considers the questions raised by Kip S. Thorne in [27]; the laws of general relativity do not enforce chronology protection: it is easy to find solutions to the Einstein field equation that have closed timelike curves (CTCs—for example, Van Stockum’s spacetime, Gödel’s solution of the Einstein equation, etc.). Physicists have generally dismissed such solutions as unphysical ones, but Thorne sees nothing unphysical in them.⁷ Here, we will copy a very important paragraph in our opinion, from this mentioned paper: *It would be rather surprising to me, if Nature uses one protection mechanism in one situation (e.g. collapsing, spinning bodies), a different one in another situation (e.g. moving cosmic strings) and a third mechanism in a third situation (e.g. the interior of a spinning black hole). More likely, there is one universal mechanism that always does the job, if other mechanisms fail.* We feel that such a “universal mechanism” is the topology of the spacetime. For example, exactly as the Path topology \mathcal{P} prevents a spacetime from satisfying the classical singularity theorems (due to the failure of LCT), in a similar way, Low has proved that a spacetime is *nakedly singular*, if the space of causal curves is non-Hausdorff (Proposition 3.1, [8]) as well as the following two propositions, which bring the discussion about singularities into a purely topological context:

Proposition 1 *For a strongly causal spacetime M , the following are equivalent:*

1. M has no geodesically accessible singularities.
2. M is causally pseudo-convex.
3. The space of causal geodesics C , of M , is Hausdorff.

Proposition 2 *A strongly causal spacetime M is globally hyperbolic, iff its space of smooth endless causal curves is Hausdorff.*

This is really a place that one has to dig a bit deeper; since the Einstein’s field equation permits solutions that bring us in front of CTCs, one has to place the problem of “rejecting specific solutions as unphysical” to topology; we are tempted to conjecture that, if there is a final and definite answer about which is the natural

⁷For more details, and Thorne’s arguments, read Section 3, from [27].

topology for a spacetime, then if under such a topology there is no (interior topological) mechanism to avoid CTCs, then one should not have the right to reject such solutions with CTCs as unphysical. If, on the other hand, under *the* natural topology of a spacetime, classical singularities fail to hold, then one has the right to claim that such theorems have no physical meaning.

Here, we feel also commenting about the “in fashion” technique to increase the spacetime dimensions, in order to “make the zeros disappear” (for a discussion, see [21]). As an example, in [2] and [3], the authors have built a model of a five-dimensional space, whose conformal infinity is our four-dimensional spacetime, its “ambient boundary.” The aim of this model was to create a topological environment where basic singularity theorems would not hold any longer (see, in particular, [3] and [4] as well as [14]). The authors finally concluded that the topology on the ambient boundary should be the Fine Zeeman topology F ; we have corrected this erratum in [14], as the F refers to special relativity while Göbel’s general relativistic analogue \mathcal{F} would be a more appropriate topology to use in a curved spacetime. We have also mentioned that the argument that the “lack” of “Euclidean-open-balls” does not necessarily imply the lack of singularities is incorrect. First of all, in a curved spacetime, an open ball will be defined via a Riemannian metric and not through the natural Euclidean metric. Second, since the topologies in the class $\mathfrak{Z} - \mathfrak{G}$ are finer than the manifold topology T_M , it is obvious that every open set in T_M will also be open in a topology T in $\mathfrak{Z} - \mathfrak{G}$; such errata, which are not rare in models in spacetime geometry, show why we need to take methods of general topology more seriously.⁸ The authors of [2] and [3] though have had an interesting idea to sort of “force” the ambient boundary, in their model, to be equipped with a topology in Zeeman–Göbel class, so that the LCT does not hold (that would work with the Path topology \mathcal{P} , for example, as we have already mentioned). And here comes the critical question: why is there a need then to increase the spacetime dimensions, while such a topology would “hide the infinities” already in four dimensions?

To bring this discussion a bit further, in Proposition 1, there is a connection between pseudo-convexity and geodesically accessible singularities.

Definition 1 A spacetime M is *causally pseudo-convex* if, for any compact set K in M , there exists another compact set K' in M , such that any causal geodesic segment with endpoints in K lies in K' .

A step further from our discussion on singularities will be a discussion on some kind of “time dilation” phenomena, in general relativity, which were noticed by Sijie Gao and Robert M. Wald in [5]. We focus our attention in the theorem below.

Theorem 1 (Gao–Wald) *Let (M, g_{ab}) be a null geodesically complete spacetime, satisfying the null energy condition (NEC) and the null generic condition (NGC). Then, given any compact region $K \subset M$, there exists another compact region K'*

⁸For a critical survey on this discussion, we refer to [18].

containing K , such that if $q, p \notin K'$ and $q \in J^+(p) - I^+(p)$, then any causal curve γ connecting p to q cannot intersect the region K' .

Gao and Wald claim that their theorem contains some suggestion of a general “time delay” phenomena in general relativity, but since K' could be far larger than K , it is difficult to make a strong argument for this kind of interpretation of the theorem. In [15], we have interpreted Gao–Wald theorem in terms of sliced spaces, and we have shown that K' can be chosen as a “small enough” causal diamond containing K . There is a more general issue here though: for the proof of Gao–Wald theorem, the role of the manifold topology T_M is vital. Based on simple topological arguments (see [15]), we see that if one used, for example, the Path topology \mathcal{P} , or any topology in the class $\mathfrak{J} - \mathfrak{O}$, the Gao–Wald theorem will fail to hold, and so some of the corollaries that follow like, for example, the one (Corollary 1 from [5]), which states that there is an absence of particle horizons, in a class of cosmological models, will fail as well.

We believe that the evidence that classical spacetime singularities depending on LCM, naked singularities depending on causal pseudo-convexity and “time-dilation” effects of the type of Gao–Wald, are all topological effects is strong, and thus such results are more topological in their nature and “less physical.”

4 A Duality Between Timelike–Spacelike Events: Between “Chronos” and “Choros”

In article [1], we have studied a duality between two order relations, in Minkowski spacetime \mathcal{M} : the chronological order \ll and the “chorological”⁹ order $<$, as well as their induced topologies. In order to define these orders, we need to have a closer look to the light cone of an event x first.

For an event $x \in \mathcal{M}$, we define the following sets:

1. $C^T(x) = \{y : y = x \text{ or } Q(y - x) < 0\}$ the *time cone* of x ,
2. $C^L(x) = \{y : Q(y - x) = 0\}$ the *light cone* of x ,
3. $C^S(x) = \{y : y = x \text{ or } Q(y - x) > 0\}$, the *space cone*¹⁰ of x ,
4. $C^{LT}(x) = C^T(x) \cup C^L(x)$ the union of the time and light cones of x , also known as the *causal cone* of x , and
5. $C^{LS}(x) = C^S(x) \cup C^L(x)$ the union of the space and light cones of x .

In [20], we present all possible relations (to our knowledge), in \mathcal{M} , that are related to the Lorentz “metric” and their induced topologies. Here, we will highlight the following to ones: $x \ll y$ iff $y \in C^T_+(x)$ (*chronology*) and for non-causally

⁹Choros stands for space, in Greek, like chronos stands for time.

¹⁰Here, the word “cone” is used in a generalized sense, i.e. it is a cone on $I \times \mathbb{S}^{n-2}$ in Minkowski space \mathcal{M}^n .

related events $x, y \in M$, $x < y$ iff $y \in C_+^S(x)$, where we have defined $C_+^S(x)$ for some fixed choice of $m \in M$ (*chorology*). For a precise and analytical mathematical description of the partition of the space cone $C^S(x)$ into two spaces, $C_+^S(x)$ and $C_-^S(x)$, we refer to [1]. Here, we will comment on the significance of this duality, without focusing on its technical details. In particular, Zeeman, in [29], stated three alternative topologies to his Fine topology F . Several authors, all listed in [26], have worked on these topologies, and in particular, in [19] and [1], we have shown that these topologies are join topologies of the Euclidean topology \mathbb{R}^4 and a particular weak interval topology; the topology that has a local base of open sets of the form $B_\epsilon(x) \cap C^T(x)$, of bounded time cones (of a radius $\epsilon > 0$) by Euclidean balls, is the join of the topology on \mathbb{R}^4 and the weak interval topology generated by $<$, while the topology that has a local base of open sets of the form $B_\epsilon(x) \cap C^S(x)$, of bounded space cones, is the join of the topology on \mathbb{R}^4 and a weak interval topology generated by \ll . In a few words, we have two topologies in $\mathfrak{J} - \mathfrak{G}$ (or, to be more precise, in \mathfrak{J}) such that, the one is generated by open sets that are bounded time cones and the other by space cones and, respectively, the one has locally an order structure by a spacelike (chorological) order while the other (which is generated by bounded space cones) by a timelike (chronological) order.

We conjecture that this duality exists in curved spacetimes, as well, but one will need to find an alternative route to define a partition of tilted space cones, to that one that we followed in [1], and create a spacelike orientation dual to timelike orientation. We believe that there is strong evidence that this problem is consistent; wherever there is (relativistic) spacetime, there are events, and wherever there are events, there are light cones¹¹ and there can be relations depending on the light cone, such as chronology \ll , causality $<$ and horismos \rightarrow . Since the space cone is defined in Minkowski space \mathcal{M} as the complement of the causal cone, one has to define general relativistic analogues of the half-planes $P_+(x)$ and $P_-(x)$ that we defined in [1]. A general relativistic analogue of $<$ will certainly be of a high interest, as one would be able to talk about a duality between timelike and spacelike, in the frame of general relativity, something that might give insights about the passage from locality to nonlocality.

5 Questions

The preceding four sections raise more questions than to those that are supposed to answer.

¹¹Indeed, there are solutions of the Einstein's field equation in general relativity, which imply an extreme tilt of the light cones that lead, for example, to CTCs: independently of whether there exists a chronology protection mechanism in a more general frame, something that was conjectured by Hawking, or if such solutions are once accepted (see [27]), we should underline that our discussion lies within the scope of general relativity and not where the theory collapses within a singularity.

1. As we mentioned in Section 3, the LCT holds under T_M and not under \mathcal{P} . In fact, there is a wider range of topologies within $\mathfrak{Z} - \mathfrak{G}$ where LCT fails to hold, while there are other topologies where LCT holds.¹² Roughly speaking, we have topologies that incorporate the causal structure of a spacetime, and the classical singularity theorems cannot be formed, while—on the other hand—these singularity theorems are formed when using other topologies, like T_M , for example, which do not incorporate the causal structure of the spacetime but are linked with the metric of the manifold structure. One could probably view this phenomenon from the perspective of Google Earth: depending from the choice available in the package, one could view satellite photos of the Earth in significant detail while, with the use of a different choice, one could make a road system appear, intervening with the satellite picture or, with another choice, one could simply view the civil map of a city with the anaglyph disappearing completely.

It might be that different topologies reveal a different perspective of spacetime, but is there a topology that is actually the smallest one from all these spacetime topologies that contains all the information that each one of them contains?

2. Given the topologies in the class $\mathfrak{Z} - \mathfrak{G}$, the general relativistic analogue \mathcal{F} , of the Fine topology F , is incomparable with several of them, including \mathcal{P} ; it might be that the condition for a topology to belong to the Zeeman–Göbel class might exclude topologies that have a significance and might be appropriate candidates to be called natural topologies. There are such topologies that are mentioned in [8], such as the topologies \mathcal{T}^0 and \mathcal{T}^1 , which by themselves belong to a class that contains finer topologies than each of them, respectively, which are defined on the space of smooth endless causal curves, in a very natural way, indeed; a further study of these topologies is needed, as they give the topological conditions for a spacetime to be globally hyperbolic (Proposition 4.3, from [8]) and connect global hyperbolicity to metrizability (Proposition 4.4).
3. Given a general relativistic analogue to the partition of the space cone that we studied in [1] (which is, still, an open question), it would be interesting to know if the spacelike geodesics form a submanifold, study their topology, as well as their convergence. Given a +–ve spacelike orientation, dual to the timelike orientation, is there a duality in results regarding the space of timelike or causal geodesics with the spacelike ones? A similar question holds for the space of endless spacelike curves (always under the frame of [1]) and a possible duality to results concerning the space of causal endless curves. Before attempting any study related to this general question, one should not forget that acausal is a global property, while spacelike is a local one.
4. An idea, which was first communicated with the mathematician Santanu Acharjee, is to consider a spacetime as a bitopological space choosing, for example, the manifold topology and another appropriate spacetime topology (for example,

¹²See [17] for an introductory discussion on this particular problem.

in the class of $\mathfrak{J} - \mathfrak{G}$) to serve the definition of bitopological space. It would be interesting to examine if such a topology incorporates the causal, differential and conformal structure of a spacetime and if it is useful to handle with.

5. Kip Thorn's comments, in [27], on rotating contracting bodies and CTCs are linked to the Einstein's field equation and are seemingly independent from the topology of a spacetime. In the Low's work, in [8], it is clear that the naked singularities are a topological effect. How could one connect these two seemingly different results?
6. Having stated the previous question, on particular solutions to the Einstein's field equation leading to CTCs, it is tempting to pose the following related question. In a spacetime manifold, is there a metrizable topology finer than the manifold and coarser than the Fine one?

There is some criticism about diminishing returns: why one should continue a study on the topology of a spacetime, if we have not concluded to something general and fruitful yet. We dare to write that such a question is not fruitful, because the topological problems that were mentioned in this chapter, including the singularity problems that are topological in nature, are too crucial to be ignored.

Acknowledgments The author wishes to thank Rolf Suabedissen, from Oxford, for being kind to reply to our topological questions, even if they were elementary; the author is grateful for his valuable time and for the collegiality. Infinite thanks to Andreas Boukas for sharing thoughts on quantum gravity, some of which are incorporated in the last section.

References

1. W. Al-Qallaf, K. Papadopoulos, On a duality between time and space cones. *Kuwait J. Sci.* **47**(2), 1–5 (2020)
2. I. Antoniadis, S. Cotsakis, Ambient cosmology and spacetime singularities. *Eur. Phys. J.C.* **75**(35), 1–12 (2015)
3. I. Antoniadis, S. Cotsakis, Topology of the ambient boundary and the convergence of causal curves. *Mod. Phys. Lett. A* **30**(30), 1550161 (2015)
4. I. Antoniadis, S. Cotsakis, K. Papadopoulos, The causal order on the ambient boundary. *Mod. Phys. Lett. A* **31**, 20 (2016)
5. S. Gao, R.M. Wald, Theorems on gravitational time delay and related issues. *Classical Quantum Gravity* **17**(24) (2000)
6. R. Göbel, Zeeman topologies on space-times of general relativity theory. *Comm. Math. Phys.* **46**, 289–307 (1976)
7. S.W. Hawking, A.R. King, P.J. McCarthy, A new topology for curved space-time which incorporates the causal, differential, and conformal structures. *J. Math. Phys.* **17**(2), 174–181 (1976)
8. R.J. Low, Spaces of causal paths and naked singularities. *Classical Quantum Gravity* **7**(6), 943–954 (1990)
9. R.J. Low, Spaces of paths and the path topology. *J. Math. Phys.* **57**, 092503 (2016)
10. E. Minguzzi, Limit curve theorems in Lorentzian geometry. *J. Math. Phys.* **49**, 092501 (2008)

11. K. Papadopoulos, On the orderability problem and the interval topology, in *Topics in Mathematical Analysis and Applications*. The Optimization and Its Applications Springer Series, ed. by T. Rassias, L. Toth (Springer, Berlin, 2014)
12. K. Papadopoulos, On properties of nests: some answers and questions. *Quest. Ans. Gen. Topol.* **33**(2) (2015)
13. K. Papadopoulos, Nests, and their role in the orderability problem, in *Mathematical Analysis, Approximation Theory and their Applications*, ed. by Th. M. Rassias, V. Gupta (Springer, Berlin, 2016), pp. 517–533
14. K. Papadopoulos, On the possibility of singularities on the ambient boundary. *Int. J. Geom. Meth. Mod. Phys.* **14**(10) (2017)
15. K. Papadopoulos, On null geodesically complete spacetimes under NEC and NGC; is the Gao-Wald “time dilation” a topological effect? (2019). arxiv:1904.12123
16. K. Papadopoulos, B.K. Papadopoulos, On two topologies that were suggested by Zeeman. *Math. Meth. Appl. Sci.* **41**(17), 7742–7747 (2018) <https://doi.org/10.1002/mma.5238>
17. K. Papadopoulos, B.K. Papadopoulos, Spacetime singularities vs. topologies of Zeeman-Göbel class. *Gravit. Cosmol.* **25**(2), 116–121 (2019)
18. K. Papadopoulos, F. Scardigli, Spacetimes as topological spaces, and the need to take methods of general topology more seriously, in *Current Trends in Mathematical Analysis and Its Interdisciplinary Applications*, ed. by H. Dutta, L.D.R. Kocinac, H.M. Srivastava (Birkhäuser-Springer, Berlin, 2019)
19. K. Papadopoulos, S. Acharjee, B.K. Papadopoulos, The order on the light cone and its induced topology. *Int. J. Geom. Meth. Mod. Phys.* **15**(05), 1850069 (2018)
20. K. Papadopoulos, N. Kurt, B.K. Papadopoulos, On the causal and topological structure of the 2-dimensional Minkowski space. *Universe* **5**(3), 70 (2019)
21. K. Papadopoulos, N. Kurt, B.K. Papadopoulos, Are four dimensions enough? A note on ambient cosmology. *Int. J. Geom. Meth. Mod. Phys.* **16**(06), 1950090 (2019)
22. R. Penrose, Gravitational collapse and space-time singularities. *Phys. Rev. Lett.* **14**, 57 (1965)
23. R. Penrose, Techniques of differential topology in relativity, in *CBMS-NSF Regional Conference Series in Applied Mathematics* (1972)
24. R. Penrose, *The Road to Reality: A Complete Guide to the Laws of the Universe* (Vintage Books, New York, 2007)
25. G.M. Reed, The intersection topology w.r.t. the real line and the countable ordinals. *Trans. Am. Math. Society* **297**(2), 509–520 (1986)
26. R. Saraykar, S. Janardhan, Zeeman-like topologies in special and general theory of relativity. *J. Mod. Phys.* **7**(7) (2018)
27. K.S. Thorne, Closed timelike curves, in *Proceedings of the 13th International Conference on General Relativity and Gravitation*, ed. by C. Kozameh (Institute of Physics, Bristol, 1993)
28. J.-G. Yun, A note on the Lorentzian limit curve theorem. *Commun. Korean Math. Soc.* **28**(3), 571–580 (2013)
29. E.C. Zeeman, The topology of Minkowski space. *Topology* **6**, 161–170 (1967)

On the Approximation of Monotone Variational Inequalities in L^p Spaces with Probability Measure



Mauro Passacantando and Fabio Raciti

Abstract In this paper we propose an approximation procedure for a class of monotone variational inequalities in probabilistic Lebesgue spaces. The implementation of the functional approximation in L^p , with $p > 2$, leads to a finite dimensional variational inequality whose structure is different from the one obtained in the case $p = 2$, already treated in the literature. The proposed computational scheme is applied to the random traffic equilibrium problem with polynomial cost functions.

1 Introduction

In many equilibrium problems arising in applied sciences, the data are often not known with precision and this uncertainty can be modeled by using some probability distributions. In this paper we are interested in the variational inequality approach to equilibrium problems which has been very fruitful in the last decades. Motivated by the need to cope with uncertain data, many authors have developed various approaches to the theory of random variational inequalities (the term stochastic variational inequalities is also used by numerous authors). Our contribution falls in the so-called L^p approach to random variational inequalities introduced in [6, 7] and subsequently developed in a series of papers [4, 5, 8–10, 12]. A comparison of the rigorous L^p approach with a sample-path approach has been carried out in [11]. In this last paper, the authors also proposed a regularization method to deal with the case where the operator is monotone but not strictly monotone and applied their results to the traffic equilibrium problem with linear cost functions, which is modeled by a variational inequality in L^2 . In this case, the regularization

M. Passacantando
Dipartimento di Informatica, Università di Pisa, Pisa, Italy
e-mail: mauro.passacantando@unipi.it

F. Raciti (✉)
Dipartimento di Matematica e Informatica, Università di Catania, Catania, Italy
e-mail: fraciti@dmi.unict.it

term is the identity operator, i.e., the Riesz isometry, and after a discretization procedure the original infinite dimensional variational inequality is transformed in a large number of *independent* finite dimensional variational inequalities. To the best of our knowledge, the above mentioned abstract regularization procedure has not yet been applied to random variational inequalities in L^p , with $p > 2$. In this paper, we show that when $p > 2$ the structure of the regularizing duality operator does not allow to split the L^p variational inequality into a large number of finite dimensional variational inequalities. Instead, it can be approximated by a single variational inequality whose operator $F : \mathbb{R}^L \rightarrow \mathbb{R}^L$ has a special structure such that all the summands in F , excepted the regularization term, depend on a number of variables which is much smaller than L . As an application of our results, we investigate the random traffic equilibrium problem with polynomial cost functions.

The paper is organized as follows. First, we give an overview of the L^p approach for random variational inequalities in Section 2.1. Then, in Section 2 we describe a functional approximation scheme combined with a regularization procedure to find approximated solutions of a random monotone variational inequality, while its implementation in L^p spaces, with $p > 2$, is analyzed in detail in Section 2.3. In Section 3 we apply the results illustrated in Section 2 to the random traffic network equilibrium problem with polynomial cost functions. The deterministic version of the problem and its variational inequality formulation are recalled in Section 3.1. Section 3.2 is devoted to the stochastic version of the problem, where both the traffic demand and the travel cost functions may include random perturbations, and a stochastic variational inequality formulation is given. Finally, the regularization and approximation procedures described in Section 2 have been applied to some instances of the random traffic network equilibrium problem in order to show the impact of different probability distributions of the random data on the average cost at equilibrium.

2 Regularization of Random Variational Inequalities

This section is devoted to the regularization and approximation procedures for random monotone variational inequalities. In particular, Section 2.1 is an overview of the L^p approach for random variational inequalities; Section 2.2 describes a functional approximation scheme combined with a regularization procedure to find approximated solutions of a random monotone variational inequality, while in Section 2.3 we discuss in detail the implementation of the regularization and approximation procedures in L^p spaces with $p > 2$.

2.1 Random Variational Inequalities in Probabilistic Lebesgue Spaces

Let (Ω, \mathcal{A}, P) be a probability space, $A, B : \mathbb{R}^k \rightarrow \mathbb{R}^k$ two given mappings and $b, c \in \mathbb{R}^k$ two given vectors. Moreover, let R and S be two real-valued random variables defined on Ω , D a random vector in \mathbb{R}^m and $G \in \mathbb{R}^{m \times k}$ a given matrix. For any $\omega \in \Omega$ we define a random set

$$M(\omega) := \{x \in \mathbb{R}^k : Gx \leq D(\omega)\}.$$

Consider the following random variational inequality: for almost every $\omega \in \Omega$, find $\hat{x} := \hat{x}(\omega) \in M(\omega)$ such that

$$(S(\omega)A(\hat{x}) + B(\hat{x}))^\top(z - \hat{x}) \geq (R(\omega)c + b)^\top(z - \hat{x}), \quad \forall z \in M(\omega). \tag{1}$$

To facilitate the foregoing discussion, we set

$$T(\omega, x) := S(\omega)A(x) + B(x).$$

We assume that A, B , and S are such that the map $T : \Omega \times \mathbb{R}^k \mapsto \mathbb{R}^k$ is a Carathéodory function, that is, for each fixed $x \in \mathbb{R}^k$ the function $T(\cdot, x)$ is measurable with respect to the σ -algebra \mathcal{A} , whereas for almost every $\omega \in \Omega$ the function $T(\omega, \cdot)$ is continuous. We also assume that $T(\omega, \cdot)$ is monotone for every $\omega \in \Omega$, i.e.,

$$(T(\omega, x) - T(\omega, y))^\top(x - y) \geq 0, \quad \forall x, y \in \mathbb{R}^k, \forall \omega \in \Omega. \tag{2}$$

If (1) is uniquely solvable, then conditions can be given to ensure that the solution belongs to an L^p space for some $p \geq 2$. This allows us to compute statistical quantities such as mean values and variances of the solution. Since we are only interested in solutions with finite first- and second-order moments, another approach is to consider an integral variational inequality instead of the parametric variational inequality (1).

Thus, for a fixed $p \geq 2$, consider the Banach space $L^p(\Omega, P, \mathbb{R}^k)$ of random vectors V from Ω to \mathbb{R}^k such that the expectation (p -moment) is given by

$$E^P(\|V\|^p) = \int_{\Omega} \|V(\omega)\|^p dP(\omega) < \infty.$$

For subsequent developments, we assume the following growth condition:

$$\|T(\omega, z)\| \leq \alpha(\omega) + \beta(\omega)\|z\|^{p-1}, \quad \forall z \in \mathbb{R}^k, \quad \text{for some } p \geq 2, \tag{3}$$

where $\alpha \in L^q(\Omega, P)$ and $\beta \in L^\infty(\Omega, P)$ where $p^{-1} + q^{-1} = 1$. Due to the above growth condition, the Nemytskii operator \hat{T} associated with T acts from $L^p(\Omega, P, \mathbb{R}^k)$ to $L^q(\Omega, P, \mathbb{R}^k)$, and is defined by

$$\hat{T}(V)(\omega) := T(\omega, V(\omega)), \quad \omega \in \Omega. \tag{4}$$

It will be useful to notice that if $T(\omega, \cdot)$ is monotone for each ω , then \hat{T} is monotone from $L^p(\Omega, P, \mathbb{R}^k)$ to $L^q(\Omega, P, \mathbb{R}^k)$, i.e.,

$$\int [T(\omega, V(\omega)) - T(\omega, U(\omega))]^\top (V(\omega) - U(\omega)) dP(\omega) \geq 0$$

holds for all $U, V \in L^p(\Omega, P, \mathbb{R}^k)$. Assuming $D \in L^p_m(\Omega) := L^p(\Omega, P, \mathbb{R}^m)$, we introduce the following nonempty, closed and convex subset of $L^p_k(\Omega)$:

$$M^P := \{V \in L^p_k(\Omega) : G V(\omega) \leq D(\omega), \quad P - a.s.\}.$$

Let $S(\omega) \in L^\infty$, $0 < \underline{s} < S(\omega) < \bar{s}$, and $R(\omega) \in L^q$. Equipped with these notations, we consider the following L^p formulation of (1): find $\hat{U} \in M^P$ such that for every $V \in M^P$ we have

$$\begin{aligned} \int_\Omega (S(\omega) A[\hat{U}(\omega)] + B[\hat{U}(\omega)])^\top (V(\omega) - \hat{U}(\omega)) dP(\omega) \\ \geq \int_\Omega (b + R(\omega) c)^\top (V(\omega) - \hat{U}(\omega)) dP(\omega). \end{aligned} \tag{5}$$

If problems (1) and (5) admit a unique solution, then they are equivalent provided that the solution of (1) belongs to L^p .

To get rid of the abstract sample space Ω , we consider the joint distribution \mathbb{P} of the random vector (R, S, D) and work with the special probability space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathbb{P})$, where $d := 2 + m$ and \mathcal{B} is the Borel σ -algebra on \mathbb{R}^d . For simplicity, we assume that R, S , and D are independent random vectors and we set

$$r = R(\omega), \quad s = S(\omega), \quad t = D(\omega), \quad y = (r, s, t).$$

For each $y \in \mathbb{R}^d$, we define the set

$$M(y) := \{x \in \mathbb{R}^k : Gx \leq t\}.$$

The pointwise formulation of the variational inequality reads: find \hat{x} such that $\hat{x}(y) \in M(y)$, \mathbb{P} -a.s., and for \mathbb{P} -almost every $y \in \mathbb{R}^d$ and for every $x \in M(y)$, we have

$$(s A[\hat{x}(y)] + B[\hat{x}(y)])^\top (x - \hat{x}(y)) \geq (rc + b)^\top (x - \hat{x}(y)). \tag{6}$$

In order to obtain the integral formulation of (6), consider the space $L^p(\mathbb{R}^d, \mathbb{P}, \mathbb{R}^k)$ and introduce the closed and convex set

$$M_{\mathbb{P}} := \{v \in L^p(\mathbb{R}^d, \mathbb{P}, \mathbb{R}^k) : Gv(r, s, t) \leq t, \mathbb{P} - a.s.\}.$$

Without any loss of generality, we assume that $R \in L^q(\Omega, P)$ and $D \in L^p(\Omega, P, \mathbb{R}^m)$ are nonnegative (otherwise we can use the standard decomposition in the positive part and the negative part). Moreover, we assume that the support (i.e., the set of possible outcomes) of $S \in L^\infty(\Omega, P)$ is the interval $[\underline{s}, \bar{s}] \subset (0, \infty)$.

With these ingredients, we consider the variational inequality problem of finding $\hat{u} \in M_{\mathbb{P}}$ such that for every $v \in M_{\mathbb{P}}$ we have

$$\begin{aligned} & \int_0^\infty \int_{\underline{s}}^{\bar{s}} \int_{\mathbb{R}_+^m} (s A[\hat{u}(y)] + B[\hat{u}(y)])^\top (v(y) - \hat{u}(y)) d\mathbb{P}(y) \\ & \geq \int_0^\infty \int_{\underline{s}}^{\bar{s}} \int_{\mathbb{R}_+^m} (b + r c)^\top (v(y) - \hat{u}(y)) d\mathbb{P}(y). \end{aligned} \tag{7}$$

We conclude this section by recalling the following general result that ensures the solvability of an infinite dimensional variational inequality like (5) or (7) (see [13] for a recent survey on existence results for variational inequalities).

Theorem 1 *Let E be a reflexive Banach space and let E^* denote its topological dual space. We denote the duality pairing between E and E^* by $\langle \cdot, \cdot \rangle_{E, E^*}$. Let K be a nonempty, closed, and convex subset of E , and $A : K \rightarrow E^*$ be monotone and continuous on finite dimensional subspaces of K . Consider the variational inequality problem of finding $u \in K$ such that*

$$\langle Au, v - u \rangle_{E, E^*} \geq 0, \quad \forall v \in K.$$

Then, a necessary and sufficient condition for the above problem to be solvable is the existence of $\delta > 0$ such that at least a solution of the variational inequality:

$$\text{find } u_\delta \in K_\delta \text{ such that } \langle Au_\delta, v - u_\delta \rangle_{E, E^*} \geq 0, \quad \forall v \in K_\delta$$

satisfies $\|u_\delta\| < \delta$, where $K_\delta = \{v \in K : \|v\| \leq \delta\}$.

In the next section, we show how the set $M_{\mathbb{P}}$ can be approximated by a sequence $\{M_{\mathbb{P}}^n\}$ of finite dimensional sets, and the functions r and s can be approximated by the sequences $\{\rho_n\}$ and $\{\sigma_n\}$ of step functions, with $\rho_n \rightarrow \rho$ in L^p and $\sigma_n \rightarrow \sigma$ in L^∞ , respectively, where $\rho(r, s, t) = r$ and $\sigma(r, s, t) = s$. When the solution of (7) is unique, we can compute a sequence of step functions \hat{u}_n which converges strongly to \hat{u} under suitable hypotheses.

2.2 A Functional Approximation Scheme for the Random Variational Inequality

We start with a discretization of the space $X := L^p(\mathbb{R}^d, \mathbb{P}, \mathbb{R}^k)$. We introduce a sequence $\{\pi_n\}$ of partitions of the support

$$\mathcal{Y} := [0, \infty[\times [\underline{s}, \bar{s}] \times \mathbb{R}_+^m$$

of the probability measure \mathbb{P} induced by the random elements $R, S,$ and D . For this, we set

$$\pi_n = (\pi_n^R, \pi_n^S, \pi_n^D),$$

where

$$\begin{aligned} \pi_n^R &:= (r_n^0, \dots, r_n^{N_n^R}), \quad \pi_n^S := (s_n^0, \dots, s_n^{N_n^S}), \quad \pi_n^{D_i} := (t_{n,i}^0, \dots, t_{n,i}^{N_n^{D_i}}), \\ 0 &= r_n^0 < r_n^1 < \dots < r_n^{N_n^R} = n, \\ \underline{s} &= s_n^0 < s_n^1 < \dots < s_n^{N_n^S} = \bar{s}, \\ 0 &= t_{n,i}^0 < t_{n,i}^1 < \dots < t_{n,i}^{N_n^{D_i}} = n \quad (i = 1, \dots, m), \\ |\pi_n^R| &:= \max\{r_n^j - r_n^{j-1} : j = 1, \dots, N_n^R\} \rightarrow 0 \quad (n \rightarrow \infty), \\ |\pi_n^S| &:= \max\{s_n^k - s_n^{k-1} : k = 1, \dots, N_n^S\} \rightarrow 0 \quad (n \rightarrow \infty), \\ |\pi_n^{D_i}| &:= \max\{t_{n,i}^{h_i} - t_{n,i}^{h_i-1} : h_i = 1, \dots, N_n^{D_i}\} \rightarrow 0 \quad (i = 1, \dots, m; n \rightarrow \infty). \end{aligned}$$

These partitions give rise to an exhausting sequence $\{\mathcal{Y}_n\}$ of subsets of \mathcal{Y} , where each \mathcal{Y}_n is given by the finite disjoint union of the intervals:

$$I_{jkh}^n := [r_n^{j-1}, r_n^j[\times [s_n^{k-1}, s_n^k[\times I_h^n,$$

where we use the multi-index $h = (h_1, \dots, h_m)$ and

$$I_h^n := \prod_{i=1}^m [t_{n,i}^{h_i-1}, t_{n,i}^{h_i}[.$$

For each $n \in \mathbb{N}$, we consider the space of the \mathbb{R}^l -valued step functions on \mathcal{Y}_n , extended by 0 outside of \mathcal{Y}_n :

$$X_n^l := \left\{ v_n : v_n(r, s, t) = \sum_j \sum_k \sum_h v_{jkh}^n 1_{I_{jkh}^n}(r, s, t), \quad v_{jkh}^n \in \mathbb{R}^l \right\},$$

where 1_I denotes the $\{0, 1\}$ -valued characteristic function of a subset I . To approximate an arbitrary function $w \in L^p(\mathbb{R}^d, \mathbb{P}, \mathbb{R})$, we employ the mean value truncation operator μ_0^n associated with the partition π_n given by

$$\mu_0^n w := \sum_{j=1}^{N_n^R} \sum_{k=1}^{N_n^S} \sum_h (\mu_{jkh}^n w) 1_{I_{jkh}^n}, \tag{8}$$

where

$$\mu_{jkh}^n w := \begin{cases} \frac{1}{\mathbb{P}(I_{jkh}^n)} \int_{I_{jkh}^n} w(y) d\mathbb{P}(y), & \text{if } \mathbb{P}(I_{jkh}^n) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Analogously, for a L^p vector function $v = (v_1, \dots, v_l)$, we define

$$\mu_0^n v := (\mu_0^n v_1, \dots, \mu_0^n v_l),$$

for which one can prove that $\mu_0^n v$ converges to v in $L^p(\mathbb{R}^d, \mathbb{P}, \mathbb{R}^l)$.

To construct approximations for the set

$$M_{\mathbb{P}} = \left\{ v \in L^p(\mathbb{R}^d, \mathbb{P}, \mathbb{R}^k) : Gv(r, s, t) \leq t, \quad \mathbb{P} - \text{a.s.} \right\},$$

we introduce the orthogonal projector $q : (r, s, t) \in \mathbb{R}^d \mapsto t \in \mathbb{R}^m$ and define, for each elementary cell I_{jkh}^n , the quantities

$$\bar{q}_{jkh}^n = (\mu_{jkh}^n q) \in \mathbb{R}^m \quad \text{and} \quad (\mu_0^n q) = \sum_{jkh} \bar{q}_{jkh}^n 1_{I_{jkh}^n} \in X_n^m.$$

This leads to the following sequence of polyhedra

$$M_{\mathbb{P}}^n := \{v \in X_n^k : Gv_{jkh}^n \leq \bar{q}_{jkh}^n, \quad \forall j, k, h\}.$$

Since our objective is to approximate the random variables R and S , we introduce

$$\rho_n = \sum_{j=1}^{N_n^R} r_n^{j-1} 1_{[r_n^{j-1}, r_n^j]} \in X_n^1 \quad \text{and} \quad \sigma_n = \sum_{k=1}^{N_n^S} s_n^{k-1} 1_{[s_n^{k-1}, s_n^k]} \in X_n^1.$$

Notice that

$$\sigma_n(r, s, t) \rightarrow \sigma(r, s, t)=s \text{ in } L^\infty(\mathbb{R}^d, \mathbb{P}), \quad \rho_n(r, s, t) \rightarrow \rho(r, s, t)=r \text{ in } L^p(\mathbb{R}^d, \mathbb{P}).$$

Combining the above ingredients, for any $n \in \mathbb{N}$ we consider the following discretized variational inequality: find $\hat{u}_n := \hat{u}_n(y) \in M_{\mathbb{P}}^n$ such that, for every $v_n \in M_{\mathbb{P}}^n$, we have

$$\begin{aligned} & \int_0^\infty \int_{\underline{s}}^{\bar{s}} \int_{\mathbb{R}^d} [\sigma_n(y) A(\hat{u}_n) + B(\hat{u}_n)]^\top [v_n - \hat{u}_n] d\mathbb{P}(y) \\ & \geq \int_0^\infty \int_{\underline{s}}^{\bar{s}} \int_{\mathbb{R}^d} [b + \rho_n(y) c]^\top [v_n - \hat{u}_n] d\mathbb{P}(y). \end{aligned} \tag{9}$$

We also assume that the probability measures $P_R, P_S,$ and P_{D_i} have the probability densities φ_R, φ_S and φ_{D_i} , with $i = 1, \dots, m$, respectively. Therefore, for $i = 1, \dots, m$, we have

$$dP_R(r) = \varphi_R(r) dr, \quad dP_S(s) = \varphi_S(s) ds, \quad dP_{D_i}(t_i) = \varphi_{D_i}(t_i) dt_i.$$

In absence of strict monotonicity, the solution of (5) and (7) is not unique. In this case (which often occurs in our application) the previous approximation procedure must be coupled with a *regularization* scheme as follows. We choose a sequence $\{\varepsilon_n\}$ of regularization parameters and choose the regularization map to be the duality map $J : L^p(\mathbb{R}^d, \mathbb{P}, \mathbb{R}^k) \rightarrow L^q(\mathbb{R}^d, \mathbb{P}, \mathbb{R}^k)$. We assume that $\varepsilon_n > 0$ for every $n \in \mathbb{N}$ and that $\varepsilon_n \downarrow 0$ as $n \rightarrow \infty$.

We can then consider, for any $n \in \mathbb{N}$, the following regularized stochastic variational inequality: find $w_n = w_n^{\varepsilon_n}(y) \in M_{\mathbb{P}}^n$ such that, for every $v_n \in M_{\mathbb{P}}^n$, we have

$$\begin{aligned} & \int_0^\infty \int_{\underline{s}}^{\bar{s}} \int_{\mathbb{R}_+^m} (\sigma_n(y) A[w_n(y)] + B[w_n(y)] + \varepsilon_n J(w_n(y)))^\top (v_n(y) - w_n(y)) d\mathbb{P}(y) \\ & \geq \int_0^\infty \int_{\underline{s}}^{\bar{s}} \int_{\mathbb{R}_+^m} (b + \rho_n(y) c)^\top (v_n(y) - w_n(y)) d\mathbb{P}(y). \end{aligned} \tag{10}$$

As usual, the solution w_n will be referred to as the regularized solution. Weak and strong convergence of w_n to the minimal-norm solution of (7) can be proved under suitable hypotheses (see below). We also recall (see, e.g., [1]) that in L^p we have

$$J(u) = \|u\|_{L^p}^{2-p} |u|^{p-2} u.$$

We recall the following convergence result (see [8]).

Theorem 2 *Assume that the growth condition (3) holds and $T(\omega, \cdot)$ is strongly monotone, uniformly with respect to $\omega \in \Omega$, that is there exists $\tau > 0$ such that*

$$(T(\omega, x) - T(\omega, y))^\top(x - y) \geq \tau \|x - y\|^2 \quad \forall x, y, \text{ a.e. } \omega \in \Omega.$$

Then the sequence $\{\hat{u}_n\}$, where \hat{u}_n is the unique solution of (9), converges strongly in $L^p(\mathbb{R}^d, \mathbb{P}, \mathbb{R}^k)$ to the unique solution \hat{u} of (7).

The following results (see [11]) highlight some of the features of the regularized solutions.

Theorem 3 *The following statements hold.*

1. *For every $n \in \mathbb{N}$, the regularized stochastic variational inequality (10) has the unique solution w_n .*
2. *Any weak limit of the sequence of regularized solutions $\{w_n\}$ is a solution of (7).*
3. *The sequence of regularized solutions $\{w_n\}$ is bounded provided that the following coercivity condition holds: there exists a bounded sequence $\{\delta_n\}$, with $\delta_n \in M_{\mathbb{P}}^n$, such that*

$$\frac{\int_0^\infty \int_{\underline{s}}^{\bar{s}} \int_{\mathbb{R}_+^m} [\sigma_n(y) A(u_n(y)) + B(u_n(y))]^\top(u_n(y) - \delta_n(y)) d\mathbb{P}(y)}{\|u_n\|} \rightarrow \infty$$

as $\|u_n\| \rightarrow \infty$.

To obtain strong convergence, we need to use the concept of fast Mosco convergence [14], as given by the following definition.

Definition 1 Let X be a normed space, let $\{K_n\}$ be a sequence of closed and convex subsets of X and let $K \subset X$ be closed and convex. Let $\{\varepsilon_n\}$ be a sequence of positive real numbers such that $\varepsilon_n \rightarrow 0$. The sequence $\{K_n\}$ is said to converge to K in the fast Mosco sense, related to ε_n , if

1. For each $x \in K$, $\exists \{x_n\} \in K_n$ such that $\varepsilon_n^{-1} \|x_n - x\| \rightarrow 0$;
2. For $\{x_m\} \subset X$, if $\{x_m\}$ weakly converges to $x \in K$, then $\exists \{z_m\} \in K$ such that $\varepsilon_m^{-1} (z_m - x_m)$ weakly converges to 0.

Theorem 4 *Assume that $M_{\mathbb{P}}^n$ converges to $M_{\mathbb{P}}$ in the fast Mosco sense related to ε_n . Moreover, assume that $\varepsilon_n^{-1} \|\sigma_n - \sigma\| \rightarrow 0$ and $\varepsilon_n^{-1} \|\rho_n - \rho\| \rightarrow 0$ as $n \rightarrow \infty$. Then the sequence of regularized solutions $\{w_n\}$ converges strongly to the minimal-norm solution of the stochastic variational inequality (7), provided that $\{w_n\}$ is bounded.*

2.3 Implementation

In this section, we derive an equivalent form of the regularized stochastic variational inequality (10) suitable for being solved on a computer. We first rewrite (10) for the

reader convenience: given any $n \in \mathbb{N}$, find $w_n = w_n^{\varepsilon_n}(y) \in M_{\mathbb{P}}^n$ such that, for every $v_n \in M_{\mathbb{P}}^n$, we have

$$\begin{aligned} \int_0^\infty \int_{\underline{s}}^{\bar{s}} \int_{\mathbb{R}_+^m} (\sigma_n(y) A[w_n] + B[w_n] + \varepsilon_n J(w_n))^\top (v_n - w_n) d\mathbb{P}(y) \\ \geq \int_0^\infty \int_{\underline{s}}^{\bar{s}} \int_{\mathbb{R}_+^m} (b + \rho_n(y) c)^\top (v_n - w_n) d\mathbb{P}(y). \end{aligned}$$

The solution of (10) is a step function which is determined by its constant (vector) values in each elementary cell I_{jlh}^n . Since for each partition of the support of \mathbb{P} we have

$$[0, \infty[\times [\underline{s}, \bar{s}[\times \mathbb{R}_+^m = \bigcup_{j,l,h} I_{jlh}^n,$$

we can write (10) as

$$\begin{aligned} \sum_j \sum_l \sum_h \int_{I_{jlh}^n} (\sigma_n(y) A[w_n] + B[w_n] + \varepsilon_n J(w_n))^\top (v_n - w_n) d\mathbb{P}(y) \\ \geq \sum_j \sum_l \sum_h \int_{I_{jlh}^n} (b + \rho_n(y) c)^\top (v_n - w_n) d\mathbb{P}(y). \end{aligned} \tag{11}$$

Bearing in mind that the components of $A[w]$ and $B[w]$ are multivariate polynomials in w_2, \dots, w_n , and that v_{jlh}^n denotes the constant vector value of $v_n(y)$ in the cell I_{jlh}^n , the value of $A[w]$ in I_{jlh}^n can be written as $A v_{jlh}^n$ and, analogously, the value of $B[w]$ in I_{jlh}^n can be written as $B v_{jlh}^n$.

For the subsequent development it is useful to notice that for a step function $w \in X_n^k$, we have

$$\|w\|_{L^p} = \left[\sum_j \sum_l \sum_h \left(\sqrt{(w_{1jlh})^2 + \dots + (w_{kjlh})^2} \right)^p \mathbb{P}(I_{jlh}^n) \right]^{1/p}.$$

Let us denote with L_n the total number of the cells I_{jlh}^n induced by the partition π_n and group all the values w_{jlh}^n , for any j, l, h , in a vector which, with abuse of notation, we denote $(w_1^n, \dots, w_{L_n}^n) \in \mathbb{R}^{k \times L_n}$, i.e., we use the same symbol for both a step function of X_n^k and its associated vector of $\mathbb{R}^{k \times L_n}$ which describes its constant values on each cell. Moreover, we make the position

$$\|w\|_{L^p}^{2-p} = f(w_1^n, \dots, w_{L_n}^n).$$

A way of ordering the elements w_{jlh}^n into a vector $(w_\alpha^n)_\alpha \in \mathbb{R}^{k \times L_n}$ will be specified later and is fundamental for the implementation of our approximation procedure. We can thus associate to the set of step functions $M_{\mathbb{P}}^n$, the set

$$M^n = \{v^n \in \mathbb{R}^{k \times L_n} : v_{jlh}^n \in M_{jlh}^n, \quad \forall j, l, h\},$$

where

$$M_{jlh}^n = \{v_{jlh}^n \in \mathbb{R}^k : Gv_{jlh}^n \leq \bar{q}_{jlh}^n\}, \quad \forall j, l, h.$$

Equipped with these notations, (11) can be equivalently written as

$$\begin{aligned} & \sum_j \sum_l \sum_h s_n^{l-1} A[w_{jlh}^n]^\top (v_{jlh}^n - w_{jlh}^n) \mathbb{P}(I_{jlh}^n) + \sum_j \sum_l \sum_h B[w_{jlh}^n]^\top (v_{jlh}^n - w_{jlh}^n) \mathbb{P}(I_{jlh}^n) \\ & + \varepsilon_n \sum_j \sum_l \sum_h f(w_1^n, \dots, w_{L_n}^n) |w_{jlh}^n|^{p-2} (w_{jlh}^n)^\top (v_{jlh}^n - w_{jlh}^n) \mathbb{P}(I_{jlh}^n) \\ & \geq \sum_j \sum_l \sum_h (b^\top + r_n^{j-1} c^\top) (v_{jlh}^n - w_{jlh}^n) \mathbb{P}(I_{jlh}^n). \end{aligned} \tag{12}$$

In (12) we can choose $v_{jlh}^n = w_{jlh}^n$ for all the cells excepted one, so as to simplify the factor $\mathbb{P}(I_{jlh}^n)$. However, the resulting inequality cannot be interpreted as a variational inequality on a single cell, because the term f involves the variables of all the cells. We can then sum again the resulting expression over the indices j, l, h and obtain

$$\begin{aligned} & \sum_j \sum_l \sum_h s_n^{l-1} A[w_{jlh}^n]^\top (v_{jlh}^n - v_{jlh}^n) + \sum_j \sum_l \sum_h B[w_{jlh}^n]^\top (v_{jlh}^n - v_{jlh}^n) \\ & + \varepsilon_n \sum_j \sum_l \sum_h f(w_1^n, \dots, w_{L_n}^n) |w_{jlh}^n|^{p-2} (w_{jlh}^n)^\top (v_{jlh}^n - v_{jlh}^n) \\ & \geq \sum_j \sum_l \sum_h (b^\top + r_n^{j-1} c^\top) (v_{jlh}^n - v_{jlh}^n). \end{aligned} \tag{13}$$

Let us notice that if $p = 2$ the variational inequality above can be split into a large number of independent variational inequalities in \mathbb{R}^k , one for each elementary cell I_{jlh} (see, e.g., [11]). This decomposition is not possible for $p > 2$ but, in this case, the last expression represents a variational inequality in $\mathbb{R}^{k \times L_n}$ with a special structure. In order to specify the structure of the operator of (13), as well as the constant term in the right hand side, so as to obtain a computational scheme that can be implemented in a straightforward manner, we need to specify a way in which the two (scalar) indices j, l and the multi-index h are mapped into a single index α . Thus, remember that:

$$j = 1, \dots, N_n^R, \quad l = 1, \dots, N_n^S, \quad h_i = 1, \dots, N_n^{D_i}, \quad i = 1, \dots, m,$$

and define

$$\alpha = 1 + (j - 1) + (l - 1)N_n^R + (h_1 - 1)N_n^R N_n^S + \dots + (h_m - 1)N_n^R N_n^S \prod_{i=1}^{m-1} N_n^{D_i}. \tag{14}$$

On the other hand, from any given value of $\alpha \in \{1, 2, \dots, L_n\}$, we can derive the corresponding indices j, l, h . This can be done in various ways and here we describe a sequential algorithm. We recall that $\lfloor a/b \rfloor$ denotes the result of the integer division of a divided by b , while $a \bmod b$ denotes the remainder. Define $\alpha_1 = \alpha - 1$ and compute

$$\begin{cases} j = (\alpha_1 \bmod N_n^R) + 1, & \alpha_2 = \lfloor \alpha_1 / N_n^R \rfloor, \\ l = (\alpha_2 \bmod N_n^S) + 1, & \alpha_3 = \lfloor \alpha_2 / N_n^S \rfloor, \\ h_1 = (\alpha_3 \bmod N_n^{D_1}) + 1, & \alpha_4 = \lfloor \alpha_3 / N_n^{D_1} \rfloor, \\ \vdots & \vdots \\ h_m = (\alpha_{m+2} \bmod N_n^m) + 1. \end{cases}$$

If we denote

$$T_l^n = s_n^{l-1} A + B \quad \text{and} \quad e_j^n = b + r_n^{j-1} c, \tag{15}$$

then (13) can be written as

$$\begin{aligned} \sum_{\alpha} [T_{\alpha}^n(w_{\alpha}^n)]^{\top} (v_{\alpha}^n - w_{\alpha}^n) + \varepsilon_n \sum_{\alpha} f(w_n) |w_{\alpha}^n|^{p-2} (w_{\alpha}^n)^{\top} (v_{\alpha}^n - w_{\alpha}^n) \\ \geq \sum_{\alpha} (e_{\alpha}^n)^{\top} (v_{\alpha}^n - w_{\alpha}^n). \end{aligned} \tag{16}$$

Notice that the expressions for T_{α}^n and e_{α}^n can be derived from (15) by using the inversion of formula (14) given above. Finally, we remark that any of the numerous algorithms for finite dimensional variational inequalities can be exploited for solving (16).

3 Application to the Traffic Network Equilibrium Problem with Random Data

In this section we apply the results shown in Section 2 to the traffic network equilibrium problem with random data. First, we recall the deterministic version of the problem and its variational inequality formulation (Section 3.1). Section 3.2 deals with the problem where both the traffic demand and the travel cost functions include random perturbations and a stochastic variational inequality formulation is given. Moreover, we prove a convergence result for the average cost at equilibrium by exploiting the approximation and regularization procedure described in Section 2.2. Finally, Section 3.3 is devoted to some numerical experiments showing the impact of different probability distributions of the random data on the average cost at equilibrium.

3.1 An Outline of the Traffic Network Equilibrium Problem

We now recall the basic definitions and the variational inequality formulation of a network equilibrium flow (see, e.g., [3, 17]). For a comprehensive treatment of all the mathematical aspects of the traffic network equilibrium problem, we refer the interested reader to the classical book of Patriksson [16]. A traffic network consists of a triple $G = (N, A, W)$, where $N = \{N_1, \dots, N_p\}$ is the set of nodes, $A = \{a_1, \dots, a_n\}$ represents the set of direct arcs (also called links) connecting pairs of nodes, and $W = \{W_1, \dots, W_m\} \subseteq N \times N$ is the set of the origin-destination (O-D) pairs. The flow on the link a_i is denoted by f_i and we group all the link flows in a vector $f = (f_1, \dots, f_n)$. A path (or route) is defined as a set of consecutive links and we assume that each O-D pair W_j is connected by r_j paths whose set is denoted by P_j . All the paths in the network are grouped into a vector (R_1, \dots, R_k) . The link structure of the paths can be described by using the link-path incidence matrix $\Delta = (\delta_{ir}), i = 1, \dots, n, r = 1, \dots, k$, with entries $\delta_{ir} = 1$, if $a_i \in R_r$, and 0 otherwise. To each path R_r it is associated a flow F_r . The path flows are grouped into a vector (F_1, \dots, F_k) which is called the network path flow (or simply, the network flow if it is clear that we refer to paths). The flow f_i on the link a_i is equal to the sum of the flows on the paths containing a_i , so that $f = \Delta F$. The unit cost of traveling through a_i is a real function $c_i(f) \geq 0$ of the flows on the network, so that $c(f) = (c_1(f), \dots, c_n(f))$ denotes the link cost vector on the network. The meaning of the cost is usually that of travel time and, in the simplest case, the generic component c_i only depends on f_i . A very popular link cost function was introduced by the Bureau of Public Roads [2] and explicitly take into account the link capacities. More precisely, the travel cost for link a_i is given by

$$c_i(f_i) = t_i^0 \left[1 + \gamma \left(\frac{f_i}{u_i} \right)^\beta \right], \tag{17}$$

where u_i describes the capacity of link a_i , t_i^0 is the free flow travel time on link a_i , while β and γ are positive parameters. Analogously, one can define a cost on the paths as $C(F) = (C_1(F), \dots, C_k(F))$. Usually, $C_r(F)$ is just the sum of the costs on the links which build that path:

$$C_r(F) = \sum_{i=1}^n \delta_{ir} c_i(f),$$

or in compact form $C(F) = \Delta^\top c(\Delta F)$. For each pair W_j , there is a given traffic demand $D_j > 0$, so that $D = (D_1, \dots, D_m)$ is the demand vector of the network. Feasible path flows are nonnegative and satisfy the demands, i.e., belong to the set

$$K = \{F \in \mathbb{R}^k : F \geq 0 \text{ and } \Phi F = D\}, \tag{18}$$

where Φ is the pair-path incidence matrix whose entries, for $j = 1, \dots, m$, $r = 1, \dots, k$, are

$$\varphi_{jr} = \begin{cases} 1, & \text{if the path } R_r \text{ connects the pair } W_j, \\ 0, & \text{elsewhere.} \end{cases}$$

The notion of a user traffic equilibrium is given by the following definition.

Definition 2 A network flow $H \in K$ is a *Wardrop equilibrium* if, for each O-D pair W_j and for each pair of paths R_r, R_s which connect W_j , the following implication holds:

$$C_r(H) > C_s(H) \implies H_r = 0;$$

that is, if traveling along the path R_r takes more time than traveling along R_s , then the flow along R_r vanishes.

Remark 1 Among the various paths which connect a given O-D pair W_j some will carry a positive flow and others zero flow. It follows from the previous definition that, for a given O-D pair, the travel cost is the same for all nonzero flow paths, otherwise users would choose a path with a lower cost. Hence, H is a Wardrop equilibrium if for each O-D pair W_j there exists $\lambda_j \in \mathbb{R}$ such that

$$C_r(H) \begin{cases} = \lambda_j, & \text{if } H_r > 0, \\ \geq \lambda_j, & \text{if } H_r = 0. \end{cases} \tag{19}$$

Hence, λ_j denotes the equilibrium cost shared by all the used paths connecting W_j . The variational inequality formulation of the Wardrop equilibrium is given by the following result (see, e.g., [3]).

Theorem 5 *A network flow $H \in K$ is a Wardrop equilibrium iff it satisfies the variational inequality*

$$C(H)^\top(F - H) \geq 0, \quad \forall F \in K. \tag{20}$$

Sometimes it is useful to decompose the scalar product in (20) according to the various O-D pairs W_j :

$$\sum_{j=1}^m \sum_{r \in P_j} C_r(H) (F_r - H_r) \geq 0, \quad \forall F \in K.$$

For the subsequent development the monotonicity properties of the cost operators will be exploited. We recall them in this section for the reader convenience.

Definition 3 A map $T : \mathbb{R}^k \rightarrow \mathbb{R}^k$ is said *monotone* if

$$(T(x) - T(y))^\top(x - y) \geq 0, \quad \forall x, y \in \mathbb{R}^k,$$

and *strictly monotone* if the equality holds only for $x = y$. T is said *strongly monotone* if there exists $\tau > 0$ such that

$$(T(x) - T(y))^\top(x - y) \geq \tau \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^k.$$

The strict monotonicity assumption of the link cost functions is commonly used because it models the congestion effect. However, this does not necessarily imply the strict monotonicity of the path cost functions, as the following lemma shows.

Lemma 1

1. *If c is monotone, then C is monotone.*
2. *If c is strictly monotone and Δ has full column rank, then C is strictly monotone.*
3. *If c is strongly monotone and Δ has full column rank, then C is strongly monotone.*

Proof

1. If $F^1, F^2 \in K$, then

$$\begin{aligned} [F^1 - F^2]^\top [C(F^1) - C(F^2)] &= [F^1 - F^2]^\top \Delta^\top [c(\Delta F^1) - c(\Delta F^2)] \\ &= [\Delta F^1 - \Delta F^2]^\top [c(\Delta F^1) - c(\Delta F^2)] \\ &\geq 0. \end{aligned}$$

2. If $F^1 \neq F^2$, then $\Delta F^1 \neq \Delta F^2$ since Δ has full column rank, hence

$$\begin{aligned} [F^1 - F^2]^\top [C(F^1) - C(F^2)] &= [F^1 - F^2]^\top \Delta^\top [c(\Delta F^1) - c(\Delta F^2)] \\ &= [\Delta F^1 - \Delta F^2]^\top [c(\Delta F^1) - c(\Delta F^2)] \\ &> 0. \end{aligned}$$

3. Let $F^1, F^2 \in K$. The strong monotonicity of c implies that there exists $\tau > 0$ such that

$$\begin{aligned} [F^1 - F^2]^\top [C(F^1) - C(F^2)] &= [F^1 - F^2]^\top \Delta^\top [c(\Delta F^1) - c(\Delta F^2)] \\ &= [\Delta F^1 - \Delta F^2]^\top [c(\Delta F^1) - c(\Delta F^2)] \\ &\geq \tau \|\Delta F^1 - \Delta F^2\|^2 \\ &= \tau (F^1 - F^2)^\top \Delta^\top \Delta (F^1 - F^2) \\ &\geq \tau \lambda_{\min}(\Delta^\top \Delta) \|F^1 - F^2\|^2, \end{aligned}$$

where $\lambda_{\min}(\Delta^\top \Delta)$, which denotes the minimum eigenvalue of $\Delta^\top \Delta$, is positive since Δ has full column rank. \square

3.2 The Stochastic VI Formulation of the Traffic Network Equilibrium Problem

We now consider the traffic network equilibrium problem where both the demand and the costs are affected by random perturbations.

Let Ω be a sample space and P be a probability measure on Ω , and consider the following feasible set which takes into consideration random fluctuations of the demand:

$$K(\omega) = \{F \in \mathbb{R}^k : F \geq 0, \Phi F = D(\omega)\}, \quad \omega \in \Omega.$$

Moreover, let $C : \Omega \times \mathbb{R}^k \mapsto \mathbb{R}^k$ be the random cost function. We can thus introduce ω as a random parameter in (20) and consider the problem of finding a vector $H(\omega) \in K(\omega)$ such that, P -a.s.:

$$C(\omega, H(\omega))^\top (F - H(\omega)) \geq 0, \quad \forall F \in K(\omega). \quad (21)$$

Definition 4 A random vector $H \in K(\omega)$ is a *random Wardrop equilibrium* if for P -almost every $\omega \in \Omega$, for each O-D pair W_j and for each pair of paths R_r, R_s which connect W_j , the following implication holds:

$$C_r(\omega, (H(\omega))) > C_s(\omega, (H(\omega))) \implies H_r(\omega) = 0. \tag{22}$$

Consider then the set

$$K_P = \{F \in L^p(\Omega, P, \mathbb{R}^k) : F_r(\omega) \geq 0, P. - \text{a.s.}, \forall r = 1, \dots, k, \\ \Phi F(\omega) = D(\omega), P. - \text{a.s.}\},$$

which is convex, closed, and bounded, hence weakly compact. Furthermore, assume that the cost function C satisfies the growth condition:

$$\|C(\omega, z)\| \leq \alpha(\omega) + \beta(\omega)\|z\|^{p-1}, \quad \forall z \in \mathbb{R}^k, P. - \text{a.s.}, \tag{23}$$

for some $\alpha \in L^q(\Omega, P)$, $\beta \in L^\infty(\Omega, P)$, $p^{-1} + q^{-1} = 1$.

Remark 2 We note that polynomial cost functions are often used to model the network congestion, e.g., the BPR cost functions (17), hence condition (23) is naturally satisfied. In particular, with linear costs the functional setting is the Hilbert space L^2 .

The Carathéodory function C gives rise to a Nemytskii map $\hat{C} : L^p(\Omega, P, \mathbb{R}^k) \rightarrow L^q(\Omega, P, \mathbb{R}^k)$ defined through the usual position

$$\hat{C}(F)(\omega) = C(\omega, F(\omega)), \tag{24}$$

and, with abuse of a notation, instead of \hat{C} , the same symbol C is often used for both the Carathéodory function and the Nemytskii map that it induces. We thus consider the following integral variational inequality: find $H \in K_P$ such that

$$\int_{\Omega} C(\omega, H(\omega))^\top (F - H(\omega)) dP(\omega) \geq 0, \quad \forall F \in K_P. \tag{25}$$

A solution of (25) satisfies the random Wardrop conditions in the sense shown by the following lemma (see [12] for the proof).

Lemma 2 *If $H \in K_P$ is a solution of (25), then H is a random Wardrop equilibrium.*

As a consequence of the previous lemma, we get that there exists a vector function $\lambda \in L^p(\Omega, P, \mathbb{R}^m)$ such that

$$C_l(\omega, H(\omega)) = \lambda_j(\omega) \tag{26}$$

for any O-D pair W_j and any path R_l connecting W_j , with $H_l(\omega) > 0$, P -almost surely.

In order to better address the modeling and computational aspects, we specify how the deterministic and the random variables appear in the operator structure.

More precisely, we assume that the operator is the sum of a purely deterministic term and of a random term, where randomness acts as a modulation. With the specifying of the constant term in the operator explicitly, we have

$$C(\omega, H(\omega)) = S(\omega)A[H(\omega)] + B[H(\omega)] - b - R(\omega)c, \tag{27}$$

where $S \in L^\infty(\Omega, P)$, $R \in L^q(\Omega)$, $A, B : L^p(\Omega, P, \mathbb{R}^k) \rightarrow L^q(\Omega, P, \mathbb{R}^k)$, $b, c \in \mathbb{R}^k$. The integral variational inequality (25) now reads

$$\begin{aligned} & \int_{\Omega} (S(\omega)(A[H(\omega)])^\top + (B[H(\omega)])^\top)(F - H(\omega))dP(\omega) \\ & \geq \int_{\Omega} (b^\top + R(\omega)c^\top)(F - H(\omega))dP(\omega), \quad \forall F \in K_P. \end{aligned} \tag{28}$$

The *average cost at equilibrium* is defined as

$$E_P[\lambda] = \int_{\Omega} \lambda(\omega)dP(\omega), \tag{29}$$

where $\lambda = \lambda(\omega) = (\lambda_1(\omega), \dots, \lambda_m(\omega))$ is defined as in (26).

Remark 3 Let us note that the integral in (29) is different from zero under the natural assumption that in each path R_r there is a link where the cost is bounded from below by a positive number (uniformly in $\omega \in \Omega$). This hypothesis is fulfilled in real networks because the cost is positive for positive flows, but also the cost at zero flow (called the free flow time) is positive, because it represents the travel time without congestion.

As already explained in the previous section, the random vector $t = D(\omega)$ and the two random variables $r = R(\omega)$ and $s = S(\omega)$ generate a probability \mathbb{P} in the image space \mathbb{R}^{2+m} of (r, s, t) from the probability P on the abstract sample space Ω . Hence, we can express the earlier defined quantities in terms of the image space variables, thus obtaining functions which can be approximated through a discretization procedure. The integration now runs over the image space variables, but to keep notation simple we just write \int instead of $\int_0^\infty \int_{\underline{s}}^{\bar{s}} \int_{\mathbb{R}_+^m}$. The transformed expression reads as follows:

$$E_{\mathbb{P}}[\lambda] = \int \lambda(r, s, t)d\mathbb{P}(r, s, t), \tag{30}$$

Let us recall that the solution $H = H(r, s, t)$ of the stochastic variational inequality which describes the network equilibrium can be approximated using the procedure explained in Section 2.2 by a sequence $\{H^n\}$ of step functions such that $H^n \rightarrow H$ in L^p , as $n \rightarrow \infty$. In the next result we give converging approximations for the mean values defined previously.

Theorem 6 For any $n \in \mathbb{N}$, we denote

$$C^n[\rho_n, \sigma_n, H^n(r, s, t)] = \sigma_n A[H^n(r, s, t)] + B[H(r, s, t)] - b - \rho_n c$$

and

$$\lambda^n(r, s, t) = (\lambda_1^n(r, s, t), \dots, \lambda_m^n(r, s, t)),$$

where $\lambda_i^n(r, s, t) = C_i^n[\rho_n, \sigma_n, H^n(r, s, t)]$ for all paths R_l connecting W_j , for which $H_l^n(r, s, t) > 0$, \mathbb{P} -a.s.. Let $\rho(r, s, t) = r$, $\sigma(r, s, t) = s$. If $\rho_n \rightarrow \rho$ strongly in L^q , $\sigma_n \rightarrow \sigma$ strongly in L^∞ , and $H^n \rightarrow H$ strongly in L^p , then the sequence

$$\{E_{\mathbb{P}}[\lambda^n]\}_n = \left\{ \int \lambda^n(r, s, t) d\mathbb{P}(r, s, t) \right\}_n$$

converges to $E_{\mathbb{P}}[\lambda]$, as $n \rightarrow \infty$. Moreover, $\text{Var}(\lambda^n) \rightarrow \text{Var}(\lambda)$.

Proof Since $H^n \rightarrow H$ strongly in L^p , it follows that $A[H^n] \rightarrow A[H]$ and $B[H^n] \rightarrow B[H]$, strongly in $L^q = L^{\frac{p}{p-1}}$ because of the continuity of the Nemytskii operators A and B . Moreover, $\rho_n \rightarrow \rho$ strongly in L^q and $\sigma_n \rightarrow \sigma$ strongly in L^∞ . As a consequence,

$$\sigma_n A[H^n] + B[H^n] - b - \rho_n c \rightarrow \sigma A[H] + B[H] - b - \rho c$$

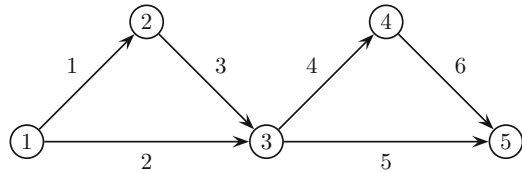
strongly in L^q , and also strongly in L^1 because \mathbb{P} is a probability measure. Hence, for each $i = 1, \dots, k$, we get $C_i^n[\rho_n, \sigma_n, H^n] \rightarrow C_i[r, s, H]$ strongly in L^1 . Moreover, since $p > 2$ strong convergence in L^p also implies convergence of variances and, by the definitions of λ and λ^n , the thesis is proved. \square

3.3 Numerical Experiments

We now report some numerical tests obtained by implementing the approximation and regularization procedures described in the previous sections. We consider a stochastic framework where both the traffic demands and the cost functions are affected by random perturbations. In particular, we assume that the random Wardrop equilibria depend on random vectors $r = R(\omega)$ and $t = D(\omega)$. The numerical computation of random Wardrop equilibria has been performed by implementing in Matlab 2018a the approximation and regularization procedures described in Section 2.2 combined with the algorithm designed in [15] for deterministic Wardrop equilibria.

Example 1 We consider the network consisting of 5 nodes and 6 links shown in Figure 1. We assume that (1,5) is the only O-D pair and the traffic demand is

Fig. 1 Test network of Example 1



$D = 100 + \delta$, where δ is a random variable which varies in the interval $[-10, 10]$ with either uniform distribution or truncated normal distribution with mean 0 and standard deviation 2.5.

The deterministic link cost functions are of the BPR form (17) defined as follows:

$$\begin{aligned}
 c_1 &= 0.5 [1 + 0.15 (f_1/5)^4], & c_2 &= 1 + 0.15 (f_2/10)^4, \\
 c_3 &= 0.5 [1 + 0.15 (f_3/5)^4], & c_4 &= 0.5 [1 + 0.15 (f_4/5)^4], \\
 c_5 &= 1 + 0.15 (f_5/10)^4, & c_6 &= 0.5 [1 + 0.15 (f_6/10)^4].
 \end{aligned}$$

The O-D pair is connected by four paths. We assume that the path cost operator is defined as in (27), where $S = 0$, $B(H) - b$ represents the deterministic path costs corresponding to the above link cost functions, while $c = -(1, \dots, 1)$ and $r = R(\omega)$ is a random variable which varies in the interval $[0, 200]$ with either uniform distribution or truncated normal distribution with mean 100 and standard deviation 25.

Notice that in this case the link-path incidence matrix Δ has not full column rank and the path cost operator is monotone but not strongly monotone. Moreover, since the deterministic part of the path cost operator is polynomial with degree 4, the operator C satisfies the growth condition (23) with $p = 5$. Therefore, the approximated regularized variational inequality (13) cannot be decomposed into a large number of small size variational inequalities.

Both the intervals $[-10, 10]$ and $[0, 200]$ have been partitioned into N_I subintervals in the approximation procedure and the regularization parameter ε has been chosen equal to $1/(N_I)^6$.

Table 1 shows the convergence of the mean values and standard deviations of the cost at equilibrium λ for increasing values of N_I , assuming that the random variables δ and r vary with uniform distribution. Similarly, Table 2 shows the mean values and standard deviations of λ , when δ and r vary with truncated normal distribution.

Example 2 We now consider the grid network shown in Figure 2 consisting of 36 nodes and 60 links. We assume that there are three O-D pairs: (1,18), (13,30), (19,36) with traffic demands equal to $D = d + \delta(1, 1, 1)$, where $d = (150, 100, 200)$ and δ is a random variable which varies in the interval $[-50, 50]$ with either uniform distribution or truncated normal distribution with mean 0 and standard deviation 10.

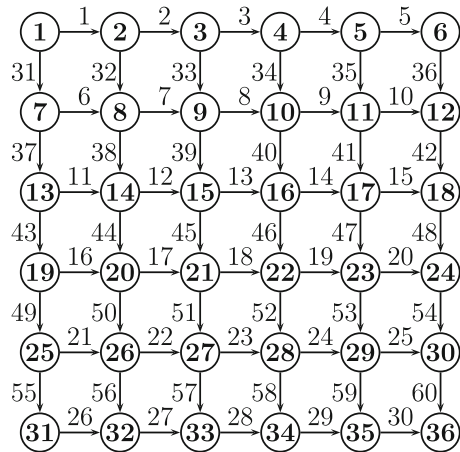
Table 1 Mean values and standard deviation of the cost at equilibrium when the random variables vary with uniform distribution

N_I	Cost at equilibrium	
	Mean value	Std deviation
5	545.825	121.69
10	546.146	123.68
15	546.205	124.04
20	546.226	124.17
25	546.236	124.23
30	546.241	124.26

Table 2 Mean values and standard deviation of the cost at equilibrium when the random variables vary with truncated normal distribution

N_I	Cost at equilibrium	
	Mean value	Std deviation
5	537.218	48.54
10	537.469	52.12
15	537.524	52.87
20	537.544	53.15
25	537.553	53.27
30	537.559	53.34

Fig. 2 Test network of Example 2



The deterministic link cost functions are of the BPR form (17) with $\gamma = 0.15$ and $\beta = 4$ for all the links, while $t_i^0 = 1$ and $u_i = 50$ for any $i = 1, \dots, 30$, and $t_i^0 = 5$ and $u_i = 100$ for any $i = 31, \dots, 60$.

We assume that the path cost operator is defined as in (27), where $S = 0$, $B(H) - b$ represents the deterministic path costs corresponding to the above link cost functions, while $c = -(1, \dots, 1)$ and $r = R(\omega)$ is a random variable which varies in the interval $[0, 20]$ with either uniform distribution or truncated normal distribution with mean 10 and standard deviation 2.

Notice that the link-path incidence matrix Δ has not full column rank since the total number of paths is greater than the number of links. Hence, the path cost

Table 3 Mean values and standard deviations of the costs at equilibrium when the random variables vary with uniform distribution

N_I	Costs at equilibrium					
	Mean values			Std deviations		
	(1,18)	(13,30)	(19,36)	(1,18)	(13,30)	(19,36)
5	19.860	22.011	22.735	3.366	4.869	5.080
10	19.940	22.044	22.833	3.456	4.942	5.229
15	19.970	22.073	22.867	3.476	4.997	5.251
20	19.974	22.078	22.871	3.481	5.012	5.256
25	19.976	22.080	22.873	3.483	5.018	5.259

Table 4 Mean values and standard deviations of the costs at equilibrium when the random variables vary with truncated normal distribution

N_I	Costs at equilibrium					
	Mean values			Std deviations		
	(1,18)	(13,30)	(19,36)	(1,18)	(13,30)	(19,36)
5	19.089	20.854	21.574	0.973	1.335	1.513
10	19.112	20.901	21.626	1.082	1.534	1.653
15	19.121	20.907	21.633	1.106	1.563	1.692
20	19.134	20.911	21.639	1.112	1.573	1.705
25	19.140	20.913	21.644	1.114	1.589	1.708

operator is monotone but not strongly monotone. Moreover, similarly to Example 1, the cost operator satisfies the growth condition (23) with $p = 5$.

Both the intervals $[-50, 50]$ and $[0, 20]$ have been partitioned into N_I subintervals in the approximation procedure and the regularization parameter ε has been chosen equal to $1/(N_I)^6$.

Tables 3 and 4 show the convergence of the mean values and standard deviations of the costs at equilibrium λ of the three O-D pairs, assuming that the random variables δ and r vary with uniform distribution or with truncated normal distribution, respectively.

Acknowledgments The authors are members of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA—National Group for Mathematical Analysis, Probability and their Applications) of the Istituto Nazionale di Alta Matematica (INdAM—National Institute of Higher Mathematics). The work of Fabio Raciti has been partially supported by University of Catania (“Piano della Ricerca 2016/2018 Linea di intervento 2”).

References

1. Y.I. Alber, Metric and generalized projection operators in Banach spaces: properties and applications, in *Theory and Applications of Nonlinear Operators of Accretive and Monotone Type*. Lecture Notes in Pure and Applied Mathematics, vol. 178 (Dekker, New York, 1996), pp. 15–50
2. Bureau of Public Roads, *Traffic Assignment Manual* (U.S. Department of Commerce, Urban Planning Division, Washington, 1964)
3. S. Dafermos, Traffic equilibrium and variational inequalities. *Transp. Sci.* **14**, 42–54 (1980)

4. P. Daniele, S. Giuffr , Random variational inequalities and the random traffic equilibrium problem. *J. Optim. Theory Appl.* **167**, 363–381 (2015)
5. F. Faraci, B. Jadamba, F. Raciti, On stochastic variational inequalities with mean value constraints. *J. Optim. Theory Appl.* **171**, 675–693 (2016)
6. J. Gwinner, F. Raciti, Random equilibrium problems on networks. *Math. Comput. Model.* **43**, 880–891 (2006)
7. J. Gwinner, F. Raciti, On a class of random variational inequalities on random sets. *Num. Funct. Anal. Optim.* **27**, 619–636 (2006)
8. J. Gwinner, F. Raciti, Some equilibrium problems under uncertainty and random variational inequalities. *Ann. Oper. Res.* **200**, 299–319 (2012)
9. 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)
10. B. Jadamba, F. Raciti, On the modelling of some environmental games with uncertain data. *J. Optim. Theory Appl.* **167**, 959–968 (2015)
11. B. Jadamba, A.A. Khan, F. Raciti, Regularization of stochastic variational inequalities and a comparison of an L^p and a sample-path approach. *Nonlinear Anal. Theory Methods Appl.* **94**, 65–83 (2014)
12. B. Jadamba, M. Pappalardo, F. Raciti, Efficiency and vulnerability analysis for congested networks with random data. *J. Optim. Theory Appl.* **177**, 563–583 (2018)
13. A. Maugeri, F. Raciti, On existence theorems for monotone and nonmonotone variational inequalities. *J. Convex Anal.* **16**, 899–911 (2009)
14. U. Mosco, Converge of convex sets and of solutions of variational inequalities. *Adv. Math.* **3**, 510–585 (1969)
15. B. Panucci, M. Pappalardo, M. Passacantando, A path-based double projection method for solving the asymmetric traffic network equilibrium problem. *Optim. Lett.* **1**, 171–185 (2007)
16. M. Patriksson, *The Traffic Assignment Problem* (VSP BV, Utrecht, 1994)
17. M.J. Smith, The existence, uniqueness and stability of traffic equilibria. *Transp. Res. B Methodol.* **13**, 295–304 (1979)

Operator Factorization and Solution of Second-Order Nonlinear Difference Equations with Variable Coefficients and Multipoint Constraints



E. Providas

Abstract A method for constructing solutions to boundary value problems for a class of second-order nonlinear difference equations with variable coefficients together with multipoint conditions is presented. The technique is based on the decomposition of the nonlinear difference equation into linear components of the same or lower order and the factorization of the associated second-order linear difference operators. The efficiency of the procedure is demonstrated by considering several examples.

1 Introduction

Nonlinear difference equations arise often in mathematical modeling of phenomena and processes in natural sciences, economics, and social sciences, see, for example, [4–6, 8, 16] and the references therein.

In general, there exist no universal methods for solving nonlinear difference equations, and, moreover, most of them cannot be solved explicitly. However, there are some types of nonlinear difference equations which can be solved in closed form by transforming them to linear difference equations. The majority of these are first-order equations with constant or variable coefficients and higher-order equations with constant coefficients. Difference equations of order $m \geq 2$ with variable coefficients, even for linear equations, are in general difficult to solve exactly [8].

Some explicit formulae for second-order and m th-order linear difference equations with arbitrarily varying coefficients have been reported, respectively, in [14, 22] and [3, 13, 15]. Algorithms for solving linear recurrence equations with polynomial coefficients are surveyed in [21]. The factorization method has also been used to solve both differential equations [10, 23, 26] and second-order difference equations [7, 12]. Multipoint boundary value problems for difference equations

E. Providas (✉)
University of Thessaly, Larissa, Greece
e-mail: providas@uth.gr

have been studied intensively over the last three decades. For example, in the pioneering article [9], existence of solutions and iterative schemes for obtaining approximate solutions to m th-order linear difference equations subject to multipoint conditions are discussed. Nonlocal boundary value problems for discrete systems of general first-order equations have been studied in [1, 24]. Solutions for second-order difference problems with nonlocal boundary conditions, following different approaches, have been addressed in [2, 19, 20, 25]. Existence and non-existence of positive solutions for second-order nonlinear difference equations subject to multipoint boundary conditions have been investigated in [11]. Finally, the papers [17, 18] are devoted to factorization method for solving multipoint problems for second-order linear difference equations with polynomial coefficients.

This paper is concerned with the exact solution of a class of linear and nonlinear second-order difference equations with variable coefficients coupled with multipoint constraints by the operator factorization method. Specifically, in Section 2, we present the operator factorization method for solving the second-order linear difference equation

$$y(n + 2) + p(n)y(n + 1) + q(n)y(n) = f(n), \quad n \in \mathbb{N}, \tag{1}$$

subject to multipoint constraints

$$\mu_{i1}y(1) + \mu_{i2}y(2) + \dots + \mu_{il}y(l) = \beta_i, \quad i = 1, 2, \quad l \geq 2, \tag{2}$$

where $\mathbb{N} = \{1, 2, 3, \dots\}$, the coefficients $p(n)$ and $q(n)$ and the nonhomogeneous term $f(n)$ are given functions (sequences) of n , $y(n)$ is an unknown function, and $\mu_{ij}, \beta_i \in \mathbb{R}, i = 1, 2, j = 1, \dots, l$. In Section 3, we deal with the construction of explicit solutions to second-order nonlinear difference equations of the kinds

$$F\left(\frac{y(n + 2)}{y(n)}, n\right) = F(x(n), n) = x^2(n) + a(n)x(n) + b(n) = 0, \quad n \in \mathbb{N}, \tag{3}$$

and

$$y(n + 2)y(n + 1) + [q(n)y(n + 1) - t(n)y(n + 2)]y(n) - q(n)t(n)y^2(n) = 0, \tag{4}$$

where the nonlinear function F is a second-degree polynomial function of $x(n) = y(n + 2)/y(n)$ and the coefficients $a(n), b(n), q(n)$ and $t(n)$ are given functions (sequences) of n , along with the multipoint conditions (2). Equations (3) and (4) can be decomposed into linear second-order difference equations, which under certain conditions can be solved by the abovementioned operator factorization method.

2 Factorization Method

Let $\mathbb{N} = \{1, 2, 3, \dots\}$ denote the set of positive non-zero integers, $y(n) : \mathbb{N} \rightarrow \mathbb{R}$ be a discrete function (sequence), and S be the space of all functions $y(n)$ defined on \mathbb{N} .

Let $L_1, L_2 : S \rightarrow S$ be two first-order linear difference operators defined by

$$L_1 y(n) = [E - r(n)] y(n), \quad D(L_1) = S, \tag{5}$$

$$L_2 y(n) = [E - s(n)] y(n), \quad D(L_2) = S, \tag{6}$$

respectively, where the independent variable $n \in \mathbb{N}$, $E y(n) = y(n + 1)$ denotes the shift operator and the coefficients $r(n), s(n) \neq 0 \in S$. Consider the composition,

$$\begin{aligned} L_1 L_2 y(n) &= L_1 (L_2 y(n)) \\ &= [E - r(n)] ([E - s(n)] y(n)) \\ &= [E^2 - (r(n) + s(n + 1)) E + r(n)s(n)] y(n), \end{aligned} \tag{7}$$

where $E^j y(n) = y(n + j)$, $j = 1, 2$. Hence, we can state the following proposition.

Proposition 1 *Let the second-order linear difference operator $L : S \rightarrow S$ be defined by*

$$L y(n) = [E^2 + p(n)E + q(n)] y(n), \quad D(L) = S, \tag{8}$$

where $n \in \mathbb{N}$, the coefficients $p(n), q(n) \in S$, $q(n) \neq 0$, and $y(n) \in S$. If there exist functions $r(n), s(n) \in S$, which satisfy the nonlinear equations

$$r(n) + s(n + 1) = -p(n), \tag{9}$$

$$r(n)s(n) = q(n), \tag{10}$$

then the operator L can be factorized into two first-order linear difference operators $L_1, L_2 : S \rightarrow S$,

$$L_1 y(n) = [E - r(n)] y(n), \quad L_2 y(n) = [E - s(n)] y(n), \tag{11}$$

with $D(L_1) = S$ and $D(L_2) = S$, respectively, such that

$$L y(n) = L_1 L_2 y(n). \tag{12}$$

Let the standard second-order discrete initial value problem

$$L y(n) = f(n), \quad y(1) = \beta_1, \quad y(2) = \beta_2, \tag{13}$$

where $\beta_1, \beta_2 \in \mathbb{R}$, $f(n) \in S$ is a forcing function, and $y(n) \in S$ is the sought function describing the response of the system modeled by (13). If relations (9) and (10) are satisfied, then by Proposition 1, we have

$$L_1 L_2 y(n) = f(n), \quad y(1) = \beta_1, \quad y(2) = \beta_2. \tag{14}$$

By setting $L_2 y(n) = z(n)$, problem (13) can be factorized into the following two first-order initial value problems,

$$L_1 z(n) = f(n), \quad z(1) = \beta_2 - s(1)\beta_1, \tag{15}$$

$$L_2 y(n) = z(n), \quad y(1) = \beta_1. \tag{16}$$

Problem (15) can be solved analytically with respect to $z(n)$ by using the standard techniques for first-order initial value problems [8]. Substituting $z(n)$ into (16) and solving in like manner, we can obtain the solution of the initial value problem (16) in closed form, which is the solution of the second-order initial value problem (13). To formalize this procedure, we prove the next lemma.

Lemma 1 *Let L be the second-order linear difference operator in (8) and \widehat{L} be its restriction on*

$$D(\widehat{L}) = \{y(n) : y(n) \in D(L), y(1) = \beta_1, y(2) = \beta_2\}, \tag{17}$$

where $\beta_1, \beta_2 \in \mathbb{R}$. If conditions (9) and (10) are satisfied, then

(i) *The operator \widehat{L} can be factorized as*

$$\widehat{L}y(n) = \widehat{L}_1 \widehat{L}_2 y(n), \tag{18}$$

where \widehat{L}_1 and \widehat{L}_2 are restrictions of the first-order linear difference operators L_1 and L_2 , defined in (5) and (6), on

$$D(\widehat{L}_1) = \{z(n) : z(n) \in D(L_1), z(1) = \beta_2 - s(1)\beta_1\}, \tag{19}$$

$$D(\widehat{L}_2) = \{y(n) : y(n) \in D(L_2), y(1) = \beta_1\}, \tag{20}$$

respectively.

(ii) *The unique solution of the initial value problem*

$$\widehat{L}y(n) = f(n), \tag{21}$$

where $f(n) \in S$ is a known function, can be obtained in closed form by

$$y(n) = \widehat{L}^{-1} f(n) = \widehat{L}_2^{-1} \widehat{L}_1^{-1} f(n), \tag{22}$$

where

$$z(n) = \widehat{L}_1^{-1} f(n) = \left[\prod_{i=1}^{n-1} r(i) \right] z(1) + \sum_{j=1}^{n-1} \left[\prod_{i=j+1}^{n-1} r(i) \right] f(j), \tag{23}$$

and

$$y(n) = \widehat{L}_2^{-1} z(n) = \left[\prod_{i=1}^{n-1} s(i) \right] y(1) + \sum_{j=1}^{n-1} \left[\prod_{i=j+1}^{n-1} s(i) \right] z(j). \tag{24}$$

Proof

(i) By Proposition 1, we have $Ly(n) = L_1L_2y(n)$. Since the operator \widehat{L} is a restriction of L and the operators \widehat{L}_1 and \widehat{L}_2 are restrictions of the operators L_1 and L_2 , respectively, it suffices to show that $D(\widehat{L}) = D(\widehat{L}_1\widehat{L}_2)$. By using (19) and (20), we obtain

$$\begin{aligned} D(\widehat{L}_1\widehat{L}_2) &= \{y(n) : y(n) \in D(\widehat{L}_2), \widehat{L}_2y(n) \in D(\widehat{L}_1)\} \\ &= \{y(n) : y(n) \in D(L_2), y(1) = \beta_1, y(n+1) - s(n)y(n) \in D(\widehat{L}_1)\} \\ &= \{y(n) : y(n) \in S, y(1) = \beta_1, y(2) - s(1)y(1) = \beta_2 - s(1)\beta_1\} \\ &= \{y(n) : y(n) \in S, y(1) = \beta_1, y(2) = \beta_2\}. \end{aligned} \tag{25}$$

Thus, if $y(n) \in D(\widehat{L}_1\widehat{L}_2)$, then $y(n) \in D(\widehat{L})$, which implies $D(\widehat{L}_1\widehat{L}_2) \subset D(\widehat{L})$. Let now $y(n) \in D(\widehat{L})$, then $y(n) \in D(L) = S, y(1) = \beta_1, y(2) = \beta_2$. It follows that $\widehat{L}_2y(n) = y(n+1) - s(n)y(n) \in D(\widehat{L}_1)$ since $y(n), y(n+1), s(n) \in S$ and $y(2) - s(1)y(1) = \beta_2 - s(1)\beta_1$, and then from (25), it is concluded that $y(n) \in D(\widehat{L}_1\widehat{L}_2)$, which means $D(\widehat{L}) \subset D(\widehat{L}_1\widehat{L}_2)$. Hence, $D(\widehat{L}) = D(\widehat{L}_1\widehat{L}_2)$.

(ii) Setting $\widehat{L}_2y(n) = z(n)$, we get the initial value problem $\widehat{L}_1z(n) = f(n)$. The latter possesses exactly one solution, namely $z(n) = \widehat{L}_1^{-1} f(n)$ as in (23), see, for example, in [8]. By substituting $z(n)$ into the former and solving in an analogous way, we obtain (24), which is the solution (22) of the second-order initial value problem (21). □

Consider now the second-order linear difference equation,

$$Ly(n) = y(n+2) + p(n)y(n+1) + q(n)y(n) = f(n), \quad n \in \mathbb{N}, \tag{26}$$

subject to multipoint boundary conditions

$$\mu_{i1}y(1) + \mu_{i2}y(2) + \dots + \mu_{il}y(l) = \beta_i, \quad i = 1, 2, \quad l \geq 2, \tag{27}$$

where $\mu_{ij}, \beta_i \in \mathbb{R}, i = 1, 2, j = 1, \dots, l$. Set up the following system of l algebraic equations,

$$\begin{aligned} q(n)y(n) + p(n)y(n + 1) + y(n + 2) &= f(n), \quad n = 1, 2, \dots, l - 2, \\ \mu_{i1}y(1) + \mu_{i2}y(2) + \dots + \mu_{il}y(l) &= \beta_i, \quad i = 1, 2, \quad l \geq 2. \end{aligned} \tag{28}$$

By using the last equation to eliminate the appearance of $y(l)$ from the other $l - 1$ equations and then using the $(l - 1)$ th equation to eliminate $y(l - 1)$ from each of the other $l - 2$ equations and repeating the same process, we finally get

$$y(i) = \hat{\beta}_i, \quad i = 1, 2, \tag{29}$$

where $\hat{\beta}_1, \hat{\beta}_2 \in \mathbb{R}$. Thus, the multipoint boundary value problem (26), (27) can be reshaped into initial value problem

$$Ly(n) = f(n), \quad y(1) = \hat{\beta}_1, \quad y(2) = \hat{\beta}_2, \tag{30}$$

which can be solved by means of Lemma 1, if relations (9) and (10) are satisfied. We prove the next theorem.

Theorem 1 *Let L be the second-order linear difference operator in (8) and \hat{P} be its restriction on*

$$\begin{aligned} D(\hat{P}) &= \{y(n) : y(n) \in D(L), \\ &\mu_{i1}y(1) + \mu_{i2}y(2) + \dots + \mu_{il}y(l) = \beta_i, \quad i = 1, 2, \quad l \geq 2\}, \end{aligned} \tag{31}$$

where $\mu_{ij}, \beta_i \in \mathbb{R}, i = 1, 2, j = 1, \dots, l$. Assume that conditions (9) and (10) are satisfied, and let

$$\det \mathbf{W} = \det \begin{bmatrix} q(1) & p(1) & 1 & 0 & \dots & 0 \\ 0 & q(2) & p(2) & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & q(l-2) & p(l-2) & 1 \\ \mu_{11} & \mu_{12} & \dots & \mu_{1,l-2} & \mu_{1,l-1} & \mu_{1l} \\ \mu_{21} & \mu_{22} & \dots & \mu_{2,l-2} & \mu_{2,l-1} & \mu_{2l} \end{bmatrix} \neq 0, \tag{32}$$

and $\hat{\mathbf{W}} = \mathbf{W}^{-1}$. Then,

(i) *The operator \hat{P} can be factorized as follows:*

$$\hat{P}y(n) = \hat{P}_1 \hat{P}_2 y(n), \tag{33}$$

where \hat{P}_1 and \hat{P}_2 are restrictions of the two first-order linear difference operators L_1 and L_2 , defined in (5) and (6), on

$$D(\widehat{P}_1) = \{z(n) : z(n) \in S, z(1) = \widehat{\beta}_2 - s(1)\widehat{\beta}_1\}, \tag{34}$$

$$D(\widehat{P}_2) = \{y(n) : y(n) \in S, y(1) = \widehat{\beta}_1\}, \tag{35}$$

respectively, where

$$\widehat{\beta}_i = \sum_{j=1}^{l-2} \widehat{w}_{ij} f(j) + \sum_{j=l-1}^l \widehat{w}_{ij} \beta_{j-l+2}, \quad i = 1, 2. \tag{36}$$

(ii) The unique solution of the multipoint boundary value problem

$$\widehat{P}y(n) = f(n), \tag{37}$$

where $f(n) \in S$ is a given function, can be obtained in closed form by

$$y(n) = \widehat{P}^{-1} f(n) = \widehat{P}_2^{-1} \widehat{P}_1^{-1} f(n), \tag{38}$$

where

$$z(n) = \widehat{P}_1^{-1} f(n) = \left[\prod_{i=1}^{n-1} r(i) \right] \left(\widehat{\beta}_2 - s(1)\widehat{\beta}_1 \right) + \sum_{j=1}^{n-1} \left[\prod_{i=j+1}^{n-1} r(i) \right] f(j), \tag{39}$$

and

$$y(n) = \widehat{P}_2^{-1} z(n) = \left[\prod_{i=1}^{n-1} s(i) \right] \widehat{\beta}_1 + \sum_{j=1}^{n-1} \left[\prod_{i=j+1}^{n-1} s(i) \right] z(j). \tag{40}$$

Proof

(i) We write Equation (28) in the compact matrix form

$$\mathbf{W}y = \mathbf{f}, \tag{41}$$

where the $l \times l$ matrix

$$\mathbf{W} = \begin{bmatrix} q(1) & p(1) & 1 & 0 & \cdots & 0 \\ 0 & q(2) & p(2) & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & q(l-2) & p(l-2) & 1 \\ \mu_{11} & \mu_{12} & \cdots & \mu_{1,l-2} & \mu_{1,l-1} & \mu_{1l} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2,l-2} & \mu_{2,l-1} & \mu_{2l} \end{bmatrix},$$

and the l -vectors

$$\mathbf{y} = \begin{pmatrix} y(1) \\ y(2) \\ \vdots \\ y(l-2) \\ y(l-1) \\ y(l) \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} f(1) \\ f(2) \\ \vdots \\ f(l-2) \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

If $\det \mathbf{W} \neq 0$, then the system of Equation (41) can be solved uniquely with respect to \mathbf{y} , namely

$$\mathbf{y} = \mathbf{W}^{-1}\mathbf{f} = \hat{\mathbf{W}}\mathbf{f} = \begin{bmatrix} \hat{w}_{11} & \cdots & \hat{w}_{1l} \\ \vdots & \ddots & \vdots \\ \hat{w}_{l1} & \cdots & \hat{w}_{ll} \end{bmatrix} \mathbf{f},$$

from where, we get

$$y(i) = \sum_{j=1}^{l-2} \hat{w}_{ij} f(j) + \sum_{j=l-1}^l \hat{w}_{ij} \beta_{j-l+2}, \quad i = 1, 2,$$

or

$$y(i) = \hat{\beta}_i, \quad i = 1, 2, \tag{42}$$

where

$$\hat{\beta}_i = \sum_{j=1}^{l-2} \hat{w}_{ij} f(j) + \sum_{j=l-1}^l \hat{w}_{ij} \beta_{j-l+2}, \quad i = 1, 2. \tag{43}$$

Hence, by means of (42) and (43), the multipoint boundary value problem (37) degenerates to initial value problem

$$\begin{aligned} \widehat{P}y(n) &= Ly(n) = y(n+2) + p(n)y(n+1) + q(n)y(n) = f(n), \\ D(\widehat{P}) &= \{y(n) : y(n) \in S, \quad y(1) = \hat{\beta}_1, \quad y(2) = \hat{\beta}_2\}. \end{aligned} \tag{44}$$

By assumption, Equations (9) and (10) hold true, and therefore by Lemma 1, the operator \widehat{P} can be factorized as in (33)–(35).

(ii) Application of Lemma 1 yields the solution formulae (38)–(40). □

3 Second-Order Nonlinear Difference Equations

In this section, we elaborate on a technique for solving two types of multipoint nonlinear problems for second-order difference equations, which can be written as products of linear second-order difference equations, by using the operator factorization method presented in the previous section.

3.1 Type I

First, consider the second-order nonlinear difference equation

$$y^2(n + 2) + a(n)y(n + 2)y(n) + b(n)y^2(n) = 0, \quad n \in \mathbb{N}, \tag{45}$$

subject to two general multipoint constraints

$$\mu_{i1}y(1) + \mu_{i2}y(2) + \dots + \mu_{il}y(l) = \beta_i, \quad i = 1, 2, \tag{46}$$

where $\mu_{ij}, \beta_i \in \mathbb{R}, i = 1, 2, j = 1, \dots, l$, and $l \geq 2$.

The difference equation (45) can be put in the form

$$F\left(\frac{y(n + 2)}{y(n)}, n\right) = F(x(n), n) = x^2(n) + a(n)x(n) + b(n) = 0, \quad n \in \mathbb{N},$$

where $x(n) = y(n + 2)/y(n)$ and the nonlinear function F is a second degree polynomial of $x(n)$. Then,

$$[x(n) + q^-(n)][x(n) + q^+(n)] = 0,$$

where $q^-(n), q^+(n) \in S$ and $a(n) = q^-(n) + q^+(n), b(n) = q^-(n)q^+(n)$, and therefore, Equation (45) can be written as

$$[y(n + 2) + q^-(n)y(n)][y(n + 2) + q^+(n)y(n)] = 0. \tag{47}$$

It follows that, either

$$y(n + 2) + q^-(n)y(n) = 0, \quad n \in \mathbb{N} \tag{48}$$

or

$$y(n + 2) + q^+(n)y(n) = 0, \quad n \in \mathbb{N}. \tag{49}$$

Thus, the solutions of the multipoint nonlinear problem (45), (46) may be obtained by solving the two multipoint linear difference problems (48), (46) and (49), (46).

Additionally, if relations (9) and (10) are satisfied, namely

$$r^-(n) + s^-(n + 1) = 0, \quad r^-(n)s^-(n) = q^-(n), \tag{50}$$

meaning there exists a function $s^-(n)$ such that $q^-(n) = -s^-(n + 1)s^-(n)$, and

$$\det \mathbf{W}^- = \det \begin{bmatrix} q^-(1) & 0 & 1 & 0 & \cdots & 0 \\ 0 & q^-(2) & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & q^-(l-2) & 0 & 1 \\ \mu_{11} & \mu_{12} & \cdots & \mu_{1,l-2} & \mu_{1,l-1} & \mu_{1l} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2,l-2} & \mu_{2,l-1} & \mu_{2l} \end{bmatrix} \neq 0, \tag{51}$$

then the problem (48), (46) can be solved by using Theorem 1.

Similarly, if the equations

$$r^+(n) + s^+(n + 1) = 0, \quad r^+(n)s^+(n) = q^+(n), \tag{52}$$

are fulfilled, i.e. there exists a function $s^+(n)$ such that $q^+(n) = -s^+(n + 1)s^+(n)$, and

$$\det \mathbf{W}^+ = \det \begin{bmatrix} q^+(1) & 0 & 1 & 0 & \cdots & 0 \\ 0 & q^+(2) & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & q^+(l-2) & 0 & 1 \\ \mu_{11} & \mu_{12} & \cdots & \mu_{1,l-2} & \mu_{1,l-1} & \mu_{1l} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2,l-2} & \mu_{2,l-1} & \mu_{2l} \end{bmatrix} \neq 0, \tag{53}$$

then the solution to problem (49), (46) may be obtained also by Theorem 1.

To illustrate the use of this technique, we contemplate the following example.

Example 1 Find the solutions of the multipoint nonlinear problem

$$\begin{aligned} y^2(n + 2) - 2(n + 1)^2y(n + 2)y(n) + n(n + 1)^2(n + 2)y^2(n) &= 0, \quad n \geq 1, \\ y(1) = \frac{1}{1000}, \quad 3y(4) - 10y(5) &= -\frac{3}{20}. \end{aligned} \tag{54}$$

Observe that the difference equation in (54) is of the kind (45), and therefore it can be written as in (47) with

$$q^-(n) = -n(n + 1), \quad q^+(n) = -(n + 1)(n + 2).$$

Thus, we get the following two multipoint linear problems:

$$\widehat{P}^- y(n) = y(n + 2) - n(n + 1)y(n) = 0, \quad n \geq 1,$$

$$D(\widehat{P}^-) = \left\{ y(n) \in S : y(1) = \frac{1}{1000}, \quad 3y(4) - 10y(5) = -\frac{3}{20} \right\}, \quad (55)$$

and

$$\widehat{P}^+ y(n) = y(n + 2) - (n + 1)(n + 2)y(n) = 0, \quad n \geq 1,$$

$$D(\widehat{P}^+) = \left\{ y(n) \in S : y(1) = \frac{1}{1000}, \quad 3y(4) - 10y(5) = -\frac{3}{20} \right\}. \quad (56)$$

Take first problem (55). Equation (50) is satisfied for

$$r^-(n) = -s^-(n + 1), \quad s^-(n) = n,$$

while by setting up the system

$$\begin{bmatrix} -2 & 0 & 1 & 0 & 0 \\ 0 & -6 & 0 & 1 & 0 \\ 0 & 0 & -12 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & -10 \end{bmatrix} \begin{pmatrix} y(1) \\ y(2) \\ y(3) \\ y(4) \\ y(5) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \frac{1}{1000} \\ -\frac{3}{20} \end{pmatrix},$$

we get $\det \mathbf{W}^- = 18 \neq 0$, i.e. the criterion (51) is also fulfilled. Hence, Theorem 1 is applicable. Inversion of the system yields

$$y(1) = \hat{\beta}_1 = \frac{1}{1000}, \quad y(2) = \hat{\beta}_2 = \frac{1}{200}.$$

Therefore, problem (55) can be factorized as follows:

$$\widehat{P}^- y(n) = \widehat{P}_1^- \widehat{P}_2^- y(n) = 0, \quad n \geq 1,$$

where

$$\widehat{P}_1^- z(n) = z(n + 1) + (n + 1)z(n) = 0, \quad n \geq 1,$$

$$D(\widehat{P}_1^-) = \left\{ z(n) \in S : z(1) = \hat{\beta}_2 - s^-(1)\hat{\beta}_1 = \frac{1}{250} \right\}, \quad (57)$$

and

$$\widehat{P}_2^- y(n) = y(n + 1) - ny(n) = z(n), \quad n \geq 1,$$

$$D(\widehat{P}_2^-) = \left\{ y(n) \in S : y(1) = \hat{\beta}_1 = \frac{1}{1000} \right\}. \quad (58)$$

By means of (39) and (40), we have

$$z(n) = (-1)^{n-1} \frac{n!}{250}, \quad n \geq 1,$$

and

$$y(n) = \frac{[3 + 2(-1)^n](n - 1)!}{1000}, \quad n \geq 1,$$

which is one of the solutions of the given multipoint nonlinear problem (54).

Take now problem (56), which can be solved in a similar manner to acquire a second solution of problem (54). Actually, the relation (52) is satisfied for

$$r^+(n) = -s^+(n + 1), \quad s^+(n) = n + 1,$$

and (53) holds true, since the system

$$\begin{bmatrix} -6 & 0 & 1 & 0 & 0 \\ 0 & -12 & 0 & 1 & 0 \\ 0 & 0 & -20 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & -10 \end{bmatrix} \begin{pmatrix} y(1) \\ y(2) \\ y(3) \\ y(4) \\ y(5) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \frac{1}{1000} \\ -\frac{3}{20} \end{pmatrix}$$

is nonsingular, $\det \mathbf{W}^+ = 36 \neq 0$. Therefore, Theorem 1 applies. Accordingly,

$$y(1) = \hat{\beta}_1 = \frac{1}{1000}, \quad y(2) = \hat{\beta}_2 = \frac{7}{240},$$

and problem (56) can be factorized as follows:

$$\hat{P}^+ y(n) = \hat{P}_1^+ \hat{P}_2^+ y(n) = 0, \quad n \geq 1,$$

where

$$\begin{aligned} \hat{P}_1^+ z(n) &= z(n + 1) + (n + 2)z(n) = 0, \quad n \geq 1, \\ D(\hat{P}_1^+) &= \left\{ z(n) \in S : z(1) = \hat{\beta}_2 - s(1)\hat{\beta}_1 = \frac{163}{6000} \right\}, \end{aligned} \tag{59}$$

and

$$\begin{aligned} \hat{P}_2^+ y(n) &= y(n + 1) - (n + 1)y(n) = z(n), \quad n \geq 1, \\ D(\hat{P}_2^+) &= \left\{ y(n) \in S : y(1) = \hat{\beta}_1 = \frac{1}{1000} \right\}. \end{aligned} \tag{60}$$

By using (39) and (40), we obtain

$$z(n) = (-1)^{n-1} \frac{163(n+1)!}{12,000}, \quad n \geq 1,$$

and

$$y(n) = \frac{[187 + 163(-1)^n]n!}{24,000}, \quad n \geq 1,$$

which is the second solution of the given multipoint nonlinear problem (54).

3.2 Type II

Consider the second-order nonlinear difference equation

$$y(n+2)y(n+1) + [q(n)y(n+1) - t(n)y(n+2)]y(n) - q(n)t(n)y^2(n) = 0, \tag{61}$$

subject to multipoint constraints

$$\mu_{i1}y(1) + \mu_{i2}y(2) + \dots + \mu_{il}y(l) = \beta_i, \quad i = 1, 2. \tag{62}$$

Equation (61) can be written as

$$[y(n+2) + q(n)y(n)][y(n+1) - t(n)y(n)] = 0,$$

and hence, either

$$y(n+2) + q(n)y(n) = 0 \tag{63}$$

or

$$y(n+1) - t(n)y(n) = 0. \tag{64}$$

Thus, the solutions of the nonlinear problem (61), (62) can be obtained by solving the second-order linear problem (63), (62) and the first-order linear problem (64), (62).

By solving the second-order linear problem (63) and (62), we may employ Theorem 1 if the relations (9) and (10) are satisfied, i.e.

$$r(n) + s(n+1) = 0, \quad r(n)s(n) = q(n), \tag{65}$$

and

$$\det \mathbf{W} = \det \begin{bmatrix} q(1) & 0 & 1 & 0 & \cdots & 0 \\ 0 & q(2) & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & q(l-2) & 0 & 1 \\ \mu_{11} & \mu_{12} & \cdots & \mu_{1,l-2} & \mu_{1,l-1} & \mu_{1l} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2,l-2} & \mu_{2,l-1} & \mu_{2l} \end{bmatrix} \neq 0. \tag{66}$$

The first-order linear problem (64) and (62) may be solved by standard means by transferring the multipoint constraints into a compatible initial condition.

To show the implementation of this procedure, we consider the following example problem.

Example 2 Let the nonlinear problem

$$y(n+2)y(n+1) - \left[\frac{n}{n+2}y(n+1) + 2^n y(n+2) \right] y(n) + \frac{n2^n}{n+2}y^2(n) = 0, \\ y(1) = y(4), \quad y(6) - y(3) = \frac{1}{3}. \tag{67}$$

The second-order nonlinear difference equation in (67) is of the type (61) with

$$q(n) = -\frac{n}{n+2}, \quad t(n) = 2^n,$$

and it can be decomposed as

$$\left[y(n+2) - \frac{n}{n+2}y(n) \right] [y(n+1) - 2^n y(n)] = 0.$$

Thus, we have the following two multipoint linear problems:

$$\widehat{P}y(n) = y(n+2) - \frac{n}{n+2}y(n) = 0, \quad n \geq 1, \\ D(\widehat{P}) = \left\{ y(n) \in S : y(1) = y(4), \quad y(6) - y(3) = \frac{1}{3} \right\}, \tag{68}$$

and

$$\widehat{T}y(n) = y(n+1) - 2^n y(n) = 0, \quad n \geq 1, \\ D(\widehat{T}) = \left\{ y(n) \in S : y(1) = y(4), \quad y(6) - y(3) = \frac{1}{3} \right\}. \tag{69}$$

The second-order linear problem (68) is factorable; actually, relation (65) is satisfied for

$$r(n) = -s(n + 1), \quad s(n) = \frac{n}{n + 1},$$

and the system

$$\begin{bmatrix} -\frac{1}{3} & 0 & 1 & 0 & 0 & 0 \\ 0 & -\frac{1}{2} & 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{3}{5} & 0 & 1 & 0 \\ 0 & 0 & 0 & -\frac{2}{3} & 0 & 1 \\ 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} y(1) \\ y(2) \\ y(3) \\ y(4) \\ y(5) \\ y(6) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \frac{1}{3} \end{pmatrix}$$

is nonsingular, $\det \mathbf{W} = 1/6 \neq 0$. Therefore, Theorem 1 applies. Accordingly,

$$y(1) = \hat{\beta}_1 = 1, \quad y(2) = \hat{\beta}_2 = 2,$$

and problem (68) can be factorized as follows:

$$\widehat{P}y(n) = \widehat{P}_1 \widehat{P}_2 y(n) = 0, \quad n \geq 1,$$

where

$$\begin{aligned} \widehat{P}_1 z(n) &= z(n + 1) + \frac{n + 1}{n + 2} z(n) = 0, \quad n \geq 1, \\ D(\widehat{P}_1) &= \left\{ z(n) \in S : z(1) = \hat{\beta}_2 - s(1)\hat{\beta}_1 = \frac{3}{2} \right\}, \end{aligned} \tag{70}$$

and

$$\begin{aligned} \widehat{P}_2 y(n) &= y(n + 1) - \frac{n}{n + 1} y(n) = z(n), \quad n \geq 1, \\ D(\widehat{P}_2) &= \left\{ y(n) \in S : y(1) = \hat{\beta}_1 = 1 \right\}. \end{aligned} \tag{71}$$

By using (39) and (40), we get

$$z(n) = (-1)^{n-1} \frac{3}{n + 1}, \quad n \geq 1,$$

and

$$y(n) = \frac{5 + 3(-1)^n}{2n}, \quad n \geq 1,$$

which is a solution of the given nonlinear problem (67).

For the first-order linear problem (69), we construct the system $\mathbf{W}y = \mathbf{f}$, viz.

$$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 & 0 \\ 0 & -4 & 1 & 0 & 0 & 0 \\ 0 & 0 & -8 & 1 & 0 & 0 \\ 0 & 0 & 0 & -16 & 1 & 0 \\ 0 & 0 & 0 & 0 & -32 & 1 \\ 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} y(1) \\ y(2) \\ y(3) \\ y(4) \\ y(5) \\ y(6) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \frac{1}{3} \end{pmatrix}.$$

The rank of the augmented matrix $[\mathbf{W} \ \mathbf{f}]$ is greater than that of matrix \mathbf{W} , and hence no solution exists to $\mathbf{W}y = \mathbf{f}$. Therefore, the first-order linear problem (69) has no solution.

References

1. R.P. Agarwal, D. O'Regan, Boundary value problems for general discrete systems on infinite intervals. *Comput. Math. Appl.* **33**, 85–99 (1997). [https://doi.org/10.1016/S0898-1221\(97\)00044-8](https://doi.org/10.1016/S0898-1221(97)00044-8)
2. M. Benchohra, S. Hamami, J. Henderson, S.K. Ntouyas, A. Ouahab, Differentiation and differences for solutions of nonlocal boundary value problems for second order difference equations. *Int. J. Diff. Eq.* **2**, 37–47 (2007)
3. F.G. Boese, On ordinary difference equations with variable coefficients. *J. Math. Anal. Appl.* **273**, 378–408 (2002). [https://doi.org/10.1016/S0022-247X\(02\)00244-5](https://doi.org/10.1016/S0022-247X(02)00244-5)
4. K.A. Chrysafis, B.K. Papadopoulos G. Papaschinopoulos, On the fuzzy difference equations of finance. *Fuzzy Sets Syst.* **159**, 3259–3270 (2008). <https://doi.org/10.1016/j.fss.2008.06.007>
5. J.M. Cushing, Difference equations as models of evolutionary population dynamics. *J. Biol. Dyn.* **13**, 103–127 (2019). <https://doi.org/10.1080/17513758.2019.1574034>
6. E.Y. Deeba, A. De Korvin, Analysis by fuzzy difference equations of a model of CO₂ level in the blood. *Appl. Math. Lett.* **12**, 33–40, (1999). [https://doi.org/10.1016/S0893-9659\(98\)00168-2](https://doi.org/10.1016/S0893-9659(98)00168-2)
7. A. Dobrogowska, M.N. Hounkonnou, Factorization method and general second order linear difference equation, in *Differential and Difference Equations with Applications. ICDDDEA 2017*, ed. by S. Pinelas, T. Caraballo, P. Kloeden, J. Graef. Springer Proceedings in Mathematics & Statistics, vol 230 (Springer, Cham, 2018), pp. 67–77. https://doi.org/10.1007/978-3-319-75647-9_6
8. S. Elaydi, *An Introduction to Difference Equations* (Springer, Berlin, 2005)
9. P.W. Eloe, Difference equations and multipoint boundary value problems. *Proc. Amer. Math. Soc.* **86**, 253–259 (1982)
10. E. García, L. Littlejohn, J.L. López, E.P. Sinusúa, Factorization of second-order linear differential equations and Liouville–Neumann expansions. *Math. Comput. Modell.* **57**, 1514–1530 (2013). <https://doi.org/10.1016/j.mcm.2012.12.012>
11. J. Henderson, R. Luca, On a multi-point discrete boundary value problem. *J. Differ. Equ. Appl.* **19**, 690–699 (2013). <https://doi.org/10.1080/10236198.2012.678839>
12. K. Janglajew, K. Valeev, The factorization of the difference operator. *Comput. Math. Appl.* **42**, 729–733 (2001). [https://doi.org/10.1016/S0898-1221\(01\)00192-4](https://doi.org/10.1016/S0898-1221(01)00192-4)

13. R.K. Kittappa, A representation of the solution of the n th order linear difference equation with variable coefficients. *Linear Algebra Appl.* **193**, 211–222 (1993). [https://doi.org/10.1016/0024-3795\(93\)90278-V](https://doi.org/10.1016/0024-3795(93)90278-V)
14. R.K. Mallik, On the solution of a second order linear homogeneous difference equation with variable coefficients. *J. Math. Anal. Appl.* **215**, 32–47 (1997). <https://doi.org/10.1006/jmaa.1997.5601>
15. R.K. Mallik, Solutions of linear difference equations with variable coefficients. *J. Math. Anal. Appl.* **222**, 79–91 (1998). <https://doi.org/10.1006/jmaa.1997.5903>
16. R.E. Mickens, *Difference Equations, Theory, Applications and Advanced Topics* (CRC Press, Boca Raton, 2015)
17. I.N. Parasidis, P. Hahamis, Factorization method for solving multipoint problems for second order difference equations with polynomial coefficients, in *Discrete Mathematics and Applications*, ed. by A. Raigorodskii, M.Th. Rassias. Springer Optimization and Its Applications (Springer, Cham, 2020, to appear). <https://www.springer.com/gp/book/9783030558567>
18. I.N. Parasidis, E. Providas, Factorization method for the second order linear nonlocal difference equations, in *International Conference Polynomial Computer Algebra '2018*, ed. by N.N. Vassiliev (Euler International Mathematical Institute, St. Petersburg, Russia, 2018), pp. 85–89
19. I.N. Parasidis, E. Providas, Closed-form Solutions for some classes of loaded difference equations with initial and nonlocal multipoint conditions, in *Modern Discrete Mathematics and Analysis*, ed. by N. Daras, T. Rassias. Springer Optimization and Its Applications, vol. 131 (Springer, Cham, 2018), pp. 363–387. https://doi.org/10.1007/978-3-319-74325-7_19
20. I.N. Parasidis, E. Providas, An exact solution method for a class of nonlinear loaded difference equations with multipoint boundary conditions. *J. Differ. Equ. Appl.* **24**, 1649–1663 (2018). <https://doi.org/10.1080/10236198.2018.1515928>
21. M. Petkovšek, H. Zakrajšek, Solving linear recurrence equations with polynomial coefficients, in *Computer Algebra in Quantum Field Theory*, ed. by C. Schneider, J. Blümlein. Texts & Monographs in Symbolic Computation (A Series of the Research Institute for Symbolic Computation, Johannes Kepler University, Linz, Austria) (Springer, Vienna, 2013), pp. 259–284. https://doi.org/10.1007/978-3-7091-1616-6_11
22. J. Popenda, One expression for the solutions of second order difference equations. *Proc. Amer. Math. Soc.* **100**, 87–93, (1987). <https://doi.org/10.1090/S0002-9939-1987-0883406-X>
23. W. Robin, Operator factorization and the solution of second-order linear ordinary differential equations. *Int. J. Math. Educ. Sci. Technol.* **38**, 189–211 (2007). <https://doi.org/10.1080/00207390601002815>
24. J. Rodríguez, K.K. Abernathy, Nonlocal boundary value problems for discrete systems. *J. Math. Anal. Appl.* **385**, 49–59 (2012). <https://doi.org/10.1016/j.jmaa.2011.06.028>
25. S. Roman, A. Štikonas, Green's function for discrete second-order problems with nonlocal boundary conditions. *Bound. Value Probl.* **2011**, 1–23 (2011). <https://doi.org/10.1155/2011/767024>
26. F. Schwarz, Decomposition of ordinary differential equations. *Bull. Math. Sci.* **7**, 575–613 (2017). <https://doi.org/10.1007/s13373-017-0110-0>

An Invitation to the Study of a Uniqueness Problem



Biagio Ricceri

Abstract In this very short chapter, we provide a strong motivation for the study of the following problem: given a real normed space E , a closed, convex, unbounded set $X \subseteq E$, and a function $f : X \rightarrow X$, find suitable conditions under which, for each $y \in X$, the function

$$x \rightarrow \|x - f(x)\| - \|y - f(x)\|$$

has at most one global minimum in X .

The aim of this very short chapter is merely to stimulate the study of the following uniqueness problem related to an unconventional way of finding fixed points based on a minimax approach.

Problem 1 Let E be a real normed space, $X \subseteq E$ a closed, convex, and unbounded set, and $f : X \rightarrow X$ a given function. Find suitable conditions under which, for each $y \in X$, the function

$$x \rightarrow \|x - f(x)\| - \|y - f(x)\|$$

has at most one global minimum in X .

A real-valued function g on a topological space S is said to be inf-compact (resp., sup-compact) if, for each $r \in \mathbf{R}$, the set $\{x \in S : g(x) \leq r\}$ (resp., $\{x \in S : g(x) \geq r\}$) is compact.

Let A be a subset of a normed space E . A function $f : A \rightarrow E$ is said to be sequentially weakly strongly continuous if, for every $x \in A$ and for every sequence $\{x_n\}$ in A converging weakly to x , the sequence $\{f(x_n)\}$ converges strongly to $f(x)$.

B. Ricceri (✉)

Department of Mathematics and Computer Science, University of Catania, Catania, Italy
e-mail: ricceri@dmf.unict.it

The motivation for studying Problem 1 is provided by Theorem 2 below which is a consequence of the following general result:

Theorem 1 *Let X be a non-empty convex set in a real vector space and let $J : X \times X \rightarrow \mathbf{R}$ be a function such that $J(x, x) = 0$ for all $x \in X$, $J(\cdot, y)$ has at most one global minimum in X for all $y \in X$ and $J(x, \cdot)$ is concave in X for all $x \in X$. Furthermore, assume that there are two topologies τ_1, τ_2 in X such that the following conditions are satisfied:*

- (a) $J(\cdot, y)$ is τ_1 -lower semicontinuous and τ_1 -inf-compact for all $y \in X$;
- (b) $J(x, \cdot)$ is τ_2 -upper semicontinuous for all $x \in X$ and, for some $x_0 \in X$, $J(x_0, \cdot)$ is τ_2 -sup-compact.

Then, there exists a point $x^ \in X$ that is, at the same time, the only global minimum of $J(\cdot, x^*)$ and a global maximum of $J(x^*, \cdot)$. In particular, x^* is a fixed point of any function $f : X \rightarrow X$ satisfying*

$$J(f(x), x) \leq \sup_{y \in X} J(x, y) \tag{1}$$

for all $x \in X$.

Proof Let us apply Theorem 1.2 of [1], considering X with the topology τ_1 . Then, that result ensures that

$$\sup_{y \in X} \inf_{x \in X} J(x, y) = \inf_{x \in X} \sup_{y \in X} J(x, y) . \tag{2}$$

In view of (a), the function $x \rightarrow \sup_{y \in X} J(x, y)$ has a global minimum in X , say x^* . Moreover, due to (b), the function $y \rightarrow \inf_{x \in X} J(x, y)$ has a global maximum in X , say y^* . Therefore, by (2), we have

$$J(x^*, y) \leq J(x^*, y^*) < J(x, y^*) \tag{3}$$

for all $x, y \in X$, with $x \neq x^*$. From (3), it follows that $x^* = y^*$. Indeed, if $x^* \neq y^*$, we would have

$$J(x^*, x^*) < J(y^*, y^*) ,$$

against the assumption that J is zero on the diagonal. So, we have

$$J(x^*, y) \leq 0 < J(x, x^*) \tag{4}$$

for all $x, y \in X$, with $x \neq x^*$. Hence, x^* is the only global minimum of $J(\cdot, x^*)$ and, at the same time, a global maximum of $J(x^*, \cdot)$. Now, let $f : X \rightarrow X$ be any function satisfying (1). We claim that $x^* = f(x^*)$. Indeed, if $x^* \neq f(x^*)$, by (4), we would have

$$\sup_{y \in X} J(x^*, y) < J(f(x^*), x^*)$$

against (1). △

As we said, an application of Theorem 1 gives the following result, which is the motivation for studying Problem 1:

Theorem 2 *Let E be a real reflexive Banach space, let $X \subseteq E$ be a closed, convex, and unbounded set, and let $f : X \rightarrow X$ be a sequentially weakly strongly continuous function such that*

$$\limsup_{\|x\| \rightarrow +\infty} \frac{\|f(x)\|}{\|x\|} < \frac{1}{2}. \tag{5}$$

Assume also that, for each $y \in X$, the function

$$x \rightarrow \|x - f(x)\| - \|y - f(x)\|$$

has at most one global minimum in X .

Then, f has a unique fixed point x^* , which satisfies

$$\|x^* - f(x)\| < \|x - f(x)\|$$

for all $x \in X \setminus \{x^*\}$.

Proof Consider the function $J : X \times X \rightarrow \mathbf{R}$ defined by

$$J(x, y) = \|x - f(x)\| - \|y - f(x)\|$$

for all $x, y \in X$. Fix $y \in X$. For each $x \in X \setminus \{0\}$, we have

$$J(x, y) \geq \|x\| - 2\|f(x)\| - \|y\| = \|x\| \left(1 - 2 \frac{\|f(x)\|}{\|x\|} \right) - \|y\|,$$

and so, in view of (5), it follows that

$$\lim_{\|x\| \rightarrow +\infty} J(x, y) = +\infty. \tag{6}$$

Further, let $x \in X$ and let $\{x_n\}$ be a sequence in X converging weakly to x . Since, by assumption, $\{f(x_n)\}$ converges strongly to $f(x)$, $\{x_n - f(x_n)\}$ converges weakly to $x - f(x)$, and so

$$\|x - f(x)\| \leq \liminf_{n \rightarrow \infty} \|x_n - f(x_n)\|.$$

As a consequence, we have

$$J(x, y) \leq \liminf_{n \rightarrow \infty} \|x_n - f(x_n)\| - \lim_{n \rightarrow \infty} \|f(x_n) - y\| = \liminf_{n \rightarrow \infty} J(x_n, y). \quad (7)$$

At this point, taking (6), (7), and the reflexivity of E into account, we see that all assumptions of Theorem 1 are satisfied, provided both τ_1 and τ_2 are the relative weak topology in X . Finally, notice that

$$J(f(x), x) = \|f(f(x)) - f(x)\| - \|f(f(x)) - x\| \leq \|x - f(x)\| = \sup_{y \in X} J(x, y)$$

for all $x \in X$. Hence, f has a unique fixed point x^* , which satisfies

$$\|x^* - f(x)\| < \|x - f(x)\|$$

for all $x \in X \setminus \{x^*\}$, as claimed. △

Reference

1. B. Ricceri, On a minimax theorem: an improvement, a new proof and an overview of its applications. *Minimax Theory Appl.* **2**, 99–152 (2017)

Schrödinger Equations in Nonlinear Optics



Martin Schechter

Abstract Using global optimization, we are able to find nontrivial solutions of the nonlinear steady-state Schrödinger equation arising in optics for wide ranges of the parameters. Our results hold in arbitrary dimensions.

1 Introduction

The study of light waves propagating in a photorefractive crystal leads to the following equation over a periodic domain $\Omega \subset \mathbb{R}^2$:

$$\Delta u = \frac{Pu}{1 + V(x) + |u|^2} + \lambda u, \quad (1)$$

where P and λ are parameters, and $V(x)$ is a nonnegative function periodic in $\overline{\Omega}$ (cf. [18]). The solution u is to be periodic in Ω with the same periods as those of Ω . This equation has the trivial solution $u = 0$. It was studied in [18, 19], where it was shown that there is a continuous energy or wavenumber spectrum that allows the existence of steady-state solutions. In particular, they showed that

1. If $P > 0$, there is a constant $\delta > 0$ such that Equation (1) has a nontrivial solution provided $0 < -\lambda < \delta$.
2. If $P < 0$ and $0 < \lambda < -P/(1 + V_0)$, then Equation (1) has a nontrivial solution.
3. If $P < 0$ and $\lambda \geq -P/(1 + v_0)$, then (1) has only the trivial solution.

Here,

$$V_0 = \max_{x \in \Omega} V(x), \quad v_0 = \min_{x \in \Omega} V(x).$$

M. Schechter (✉)

Department of Mathematics, University of California, Irvine, CA, USA
e-mail: mschecht@math.uci.edu

© Springer Nature Switzerland AG 2021

T. M. Rassias, P. M. Pardalos (eds.), *Nonlinear Analysis and Global Optimization*,
Springer Optimization and Its Applications 167,
https://doi.org/10.1007/978-3-030-61732-5_22

449

Wave propagation in nonlinear periodic lattices has been studied by many researchers (cf., e.g., [1–8, 13, 15–18] and their bibliographies.)

In [13], we studied (1) in the following way. Let $a(x) = 1/(1 + V(x))$, and assume that $a(x)$ is positive and bounded: $0 < m_0 \leq a(x) \leq m_1 < \infty$. Then, (1) becomes

$$\Delta u = \frac{Pau}{1 + a|u|^2} + \lambda u. \tag{2}$$

In stating our results, we made use of the following considerations. Let Ω be a bounded periodic domain in \mathbb{R}^n , $n \geq 1$. Consider the operator $-\Delta$ on functions in $L^2(\Omega)$ having the same periods as Ω . The spectrum of $-\Delta$ consists of isolated eigenvalues of finite multiplicity:

$$0 = \lambda_0 < \lambda_1 < \dots < \lambda_\ell < \dots,$$

with eigenfunctions in $L^\infty(\Omega)$. Let λ_ℓ , $\ell \geq 0$, be one of these eigenvalues, and define

$$N = \bigoplus_{\lambda \leq \lambda_\ell} E(\lambda), \quad M = N^\perp,$$

where $E(\lambda)$ is the eigenspace corresponding to λ .

We proved

Theorem 1

1. If $P > 0$, $\lambda < 0$, and there is an $\ell \geq 0$ such that $\lambda_\ell + m_1 P \leq |\lambda| < \lambda_{\ell+1} + m_0 P$, $|\lambda| > \lambda_{\ell+1}$, then (1) has a nontrivial solution.
2. If $P < 0$, $\lambda < 0$, and there is an $\ell \geq 0$ such that $\lambda_\ell + m_0 P < |\lambda| \leq \lambda_{\ell+1} + m_1 P$, $|\lambda| < \lambda_\ell$, then (1) has a nontrivial solution.
3. If $P > 0$, $\lambda < 0$, and $0 < |\lambda| < m_0 P$, then (1) has a nontrivial solution.
4. If $P < 0$, $\lambda > 0$, and $0 < \lambda < m_0 |P|$, then (1) has a nontrivial solution.

In the present paper, we wish to cover some remaining situations not mentioned in [13, 18, 19] as well as extend their results to higher dimensions. We shall show that there are many intervals of the parameters in which nontrivial solutions exist. Our results are true in any dimension.

We now extend the results of [13] in the following way. We remove the assumptions on $a(x)$ and assume only that $V(x)$ is a positive function in $L^1(\Omega)$. We prove

Theorem 2 *If $P < 0$, $\lambda > 0$, and*

$$\lambda|\Omega| + P \int_{\Omega} \frac{1}{1 + V} dx < 0,$$

then (1) has a nontrivial solution.

Theorem 3 *If $P > 0, \lambda < 0,$*

$$\lambda + P \frac{1}{1 + V} \geq 0 \text{ a.e.,}$$

and

$$\lambda|\Omega| + P \int_{\Omega} \frac{1}{1 + V} dx > 0,$$

then (1) has a nontrivial solution.

Theorems 2 and 3 will be proved using the lemmas of the next section.

2 Some Lemmas

In proving our results, we shall make use of the following lemmas (cf., e.g., [9, 12, 14]). For the definition of linking, cf. [9].

Lemma 1 *If $G(u) \in C^1(E, \mathbb{R})$ and*

$$b_0 = \inf_E G > -\infty, \tag{3}$$

then there is a sequence satisfying

$$G(u_k) \rightarrow b_0, \quad (1 + \|u_k\|_E)G'(u_k) \rightarrow 0. \tag{4}$$

Lemma 2 *The sets $\|u\|_E = R > 0$ and $\{e_1, e_2\}$ link each other provided $\|e_1\|_E < R$ and $\|e_2\|_E > R.$*

Lemma 3 *If A links $B,$ and $G(u) \in C^1(E, \mathbb{R})$ satisfies*

$$a_0 = \sup_A G \leq b_0 = \inf_B G, \tag{5}$$

then there is a sequence $\{u_k\}$ such that

$$G(u_k) \rightarrow c \geq b_0, \quad (1 + \|u_k\|_E)\|G'(u_k)\| \rightarrow 0. \tag{6}$$

We let E be the subspace of $H^{1,2}(\Omega)$ consisting of those functions having the same periodicity as Ω with norm given by

$$\|w\|_E^2 = \|\nabla w\|^2 + \|w\|^2.$$

Define

$$I_V(u) = \frac{1}{P} \|\nabla u\|^2 + \frac{\lambda}{P} \|u\|^2 + \int_{\Omega} \ln\{1 + V(x) + |u|^2\} dx. \tag{7}$$

Then,

$$(I'_V(u), v)/2 = \frac{1}{P} (\nabla u, \nabla v) + \frac{\lambda}{P} (u, v) + \int_{\Omega} \frac{u}{1 + V + u^2} v dx. \tag{8}$$

We have

Lemma 4 *If $G(u) = I_V(u)$ is given by (7), then every sequence satisfying (6) has a subsequence converging in E . Consequently, there is a $u \in E$ such that $I_V(u) = c$ and $I'_V(u) = 0$.*

Proof The sequence satisfies

$$I_V(u_k) = \frac{1}{P} \|\nabla u_k\|^2 + \frac{\lambda}{P} \|u_k\|^2 + \int_{\Omega} \ln\{1 + V + |u_k|^2\} dx \rightarrow c, \tag{9}$$

$$(I'_V(u_k), v)/2 = \frac{1}{P} (\nabla u_k, \nabla v) + \frac{\lambda}{P} (u_k, v) + \int_{\Omega} \frac{u_k}{1 + V + u_k^2} v dx \rightarrow 0, \tag{10}$$

and

$$(I'_V(u_k), u_k)/2 = \frac{1}{P} (\nabla u_k, \nabla u_k) + \frac{\lambda}{P} (u_k, u_k) + \int_{\Omega} \frac{u_k^2}{1 + V + u_k^2} dx \rightarrow 0. \tag{11}$$

Thus,

$$\int_{\Omega} H(x, u_k) dx \rightarrow c, \tag{12}$$

where

$$H(x, t) = \ln(1 + V + t^2) - \frac{t^2}{1 + V + t^2}. \tag{13}$$

Let $\rho_k = \|u_k\|_E$. Assume first that $\rho_k \rightarrow \infty$. Let $\tilde{u}_k = u_k/\rho_k$. Then, $\|\tilde{u}_k\|_E = 1$. Hence, there is a renamed subsequence such that $\tilde{u}_k \rightharpoonup \tilde{u}$ in E , and $\tilde{u}_k \rightarrow \tilde{u}$ in $L^2(\Omega)$ and a.e. By (11),

$$P(I'_V(u_k), u_k)/2 = \|\nabla u_k\|^2 + \lambda \|u_k\|^2 + P \int_{\Omega} \frac{u_k^2}{1 + V + u_k^2} dx \rightarrow 0. \tag{14}$$

Hence,

$$\begin{aligned}
 1 &= \|\tilde{u}_k\|_E^2 \leq |P(I'_V(u_k), u_k)/2\rho_k^2| \\
 &\quad + |1 - \lambda| \cdot \|\tilde{u}_k\|^2 + |P| \int_{\Omega} \frac{\tilde{u}_k^2}{1 + V + u_k^2} dx \\
 &\leq [|1 - \lambda| + |P|] \cdot \|\tilde{u}_k\|^2.
 \end{aligned}$$

In the limit, we have

$$1 \leq [|1 - \lambda| + |P|] \cdot \|\tilde{u}\|^2.$$

This shows that $\tilde{u} \neq 0$. Let Ω_0 be the subset of Ω where $\tilde{u}(x) \neq 0$. Then, $|\Omega_0| \neq 0$ and $|u_k(x)| \rightarrow \infty$ when $x \in \Omega_0$. Consequently,

$$0 \leq H(x, u_k) \rightarrow \infty, \quad x \in \Omega_0.$$

Thus,

$$\begin{aligned}
 \int_{\Omega} H(x, u_k) dx &= \int_{\Omega_0} H(x, u_k) dx + \int_{\Omega \setminus \Omega_0} H(x, u_k) dx \\
 &\geq \int_{\Omega_0} H(x, u_k) dx \rightarrow \infty.
 \end{aligned}$$

This contradicts (12). Thus, the sequence satisfying (6) is bounded in E . Hence, there is a renamed subsequence such that $u_k \rightharpoonup u$ in E , and $u_k \rightarrow u$ in $L^2(\Omega)$ and a.e. Taking the limit in (10), we obtain

$$(I'_V(u), v)/2 = \frac{1}{P}(\nabla u, \nabla v) + \frac{\lambda}{P}(u, v) + \int_{\Omega} \frac{uv}{1 + V + u^2} dx = 0, \quad v \in E. \quad (15)$$

Thus, u satisfies $I'_V(u) = 0$. Since $u \in E$, it satisfies

$$(I'_V(u), u)/2 = \frac{1}{P}(\nabla u, \nabla u) + \frac{\lambda}{P}(u, u) + \int_{\Omega} \frac{u^2}{1 + V + u^2} dx = 0. \quad (16)$$

Also, from the limit in (11), we have

$$\begin{aligned}
 \lim \frac{1}{P} \|\nabla u_k\|^2 &= \lim(I'_V(u_k), u_k)/2 \\
 &\quad - \lim\left[\frac{\lambda}{P} \|u_k\|^2 + \int_{\Omega} \frac{u_k^2}{1 + V + u_k^2} dx\right]
 \end{aligned}$$

$$\begin{aligned}
 &= - \left[\frac{\lambda}{P} \|u\|^2 + \int_{\Omega} \frac{u^2}{1 + V + u^2} dx \right] \\
 &= \frac{1}{P} \|\nabla u\|^2.
 \end{aligned}$$

Consequently, $\nabla u_k \rightarrow \nabla u$ in $L^2(\Omega)$. This shows that $I_V u_k \rightarrow I_V(u)$. Hence, $I_V(u) = c$.

Lemma 5 *If $I'_V(u) = 0$, then u is a solution of (1).*

Proof From (15), we see that

$$|(\nabla u, \nabla v)| \leq C \|v\|, \quad v \in E.$$

From the fact that the functions and Ω are periodic with the same period, it follows that $u \in H^{2,2}(\Omega)$ and satisfies (1) (cf., e.g., [10]).

Lemma 6

$$\int_{\Omega} \ln(1 + V + u^2) dx / \|u\|_H^2 \rightarrow 0, \quad \|u\|_H \rightarrow \infty. \tag{17}$$

Proof Suppose $v_k \in H$ is a sequence such that $\rho_k = \|v_k\|_H \rightarrow \infty$. Let $\tilde{v}_k = v_k / \rho_k$. Then, $\|\tilde{v}_k\|_H = 1$. Hence, there is a renamed subsequence such that $\tilde{v}_k \rightharpoonup \tilde{v}$ in H , and $\tilde{v}_k \rightarrow \tilde{v}$ in $L^2(\Omega)$ and a.e. Now,

$$\frac{\ln(1 + v_k^2 + V)}{\rho_k^2} = \frac{\ln(1 + v_k^2 + V)}{v_k^2 + V} [\tilde{v}_k^2 + (V/\rho_k^2)] \rightarrow 0 \text{ a.e.,}$$

and it is dominated a.e. by $\tilde{v}_k^2 + (V/\rho_k^2) \rightarrow \tilde{v}^2$ in $L^1(\Omega)$. Thus,

$$\int_{\Omega} \frac{\ln(1 + v_k^2 + V)}{\rho_k^2} dx \rightarrow 0.$$

Lemma 7 *If*

$$I_V(u) = \|u\|_H^2 - \int_{\Omega} \ln(1 + V + u^2) dx,$$

then

$$I_V(v) \rightarrow \infty \text{ as } \|v\|_H \rightarrow \infty. \tag{18}$$

Proof We have

$$I_V(u)/\|u\|_H^2 = 1 - \int_{\Omega} \ln(1 + V + u^2)dx/\|u\|_H^2 \rightarrow 1, \quad \|u\|_H \rightarrow \infty$$

by Lemma 6. This gives (18).

Lemma 8 *If $u_k \in H$ is a sequence such that $\rho_k = \|u_k\|_H \rightarrow 0$, $\tilde{u}_k = u_k/\rho_k$, and $\tilde{u}_k \rightarrow \tilde{u}$ in $L^2(\Omega)$ and a.e., then*

$$\int_{\Omega} [\ln(1 + V + u_k^2) - \ln(1 + V)]dx/\|u_k\|_H^2 \rightarrow \int_{\Omega} \frac{1}{1 + V} \tilde{u}^2(x)dx. \quad (19)$$

Proof Suppose $u_k \in H$ is a sequence such that $\rho_k = \|u_k\|_H \rightarrow 0$. In particular, there is a renamed subsequence such that $u_k \rightarrow 0$ a.e. Let $\tilde{u}_k = u_k/\rho_k$. Then, $\|\tilde{u}_k\|_H = 1$. Hence, there is a renamed subsequence such that $\tilde{u}_k \rightarrow \tilde{u} \in H$, and $\tilde{u}_k \rightarrow \tilde{u}$ in $L^2(\Omega)$ and a.e. Now,

$$\begin{aligned} \frac{\ln(1 + V + u_k^2) - \ln(1 + V)}{\rho_k^2} &= \frac{\ln(1 + u_k^2/(1 + V))}{u_k^2/(1 + V)} \frac{\tilde{u}_k^2}{1 + V} \\ &\rightarrow \frac{1}{1 + V} \tilde{u}^2 \text{ a.e.,} \end{aligned}$$

and it is dominated a.e. by $\tilde{u}_k^2/(1 + V) \rightarrow \tilde{u}^2/(1 + V)$ in $L^1(\Omega)$. Thus,

$$\int_{\Omega} \frac{\ln(1 + V + u_k^2) - \ln(1 + V)}{\rho_k^2} dx \rightarrow \int_{\Omega} \frac{1}{1 + V} \tilde{u}^2 dx.$$

3 Proof of Theorem 2

Let

$$I_V(v) = \frac{1}{P} \|\nabla v\|^2 + \frac{\lambda}{P} \|v\|^2 + \int_{\Omega} \ln\{1 + v^2 + V\} dx, \quad v \in H. \quad (20)$$

Then,

$$(I'_V(v), g)/2 = \frac{1}{P} (\nabla v, \nabla g) + \frac{\lambda}{P} (v, g) + \int_{\Omega} \frac{vg}{1 + v^2 + V} dx. \quad (21)$$

If $I'_V(v) = 0$, then v satisfies

$$\Delta v = \frac{Pv}{1 + v^2 + V} + \lambda v, \quad (22)$$

which is (1). If we can find a solution $v \neq 0$ of $I'_V(v) = 0$, then we shall have a solution of (1). To find such a solution, we first note that by Lemma 7

$$PI_V(v) \rightarrow \infty \text{ as } \|v\|_H \rightarrow \infty. \tag{23}$$

Let the sequence $u_k \in H$ satisfy

$$PI_V(u_k) \searrow \alpha = \inf_H PI_V$$

(which may be $-\infty$). By (23), $\rho_k = \|u_k\|_H$ is bounded. Hence, there is a renamed subsequence such that $u_k \rightharpoonup u_0$ in H , and $u_k \rightarrow u_0$ in $L^2(\Omega)$, and a.e. Thus,

$$\begin{aligned} \|\nabla u_k\|^2 &= PI_V(u_k) - \lambda \|u_k\|^2 - P \int_{\Omega} \ln\{1 + u_k^2 + V\} dx \\ &\rightarrow \alpha - \lambda \|u_0\|^2 - P \int_{\Omega} \ln\{1 + u_0^2 + V\} dx, \end{aligned}$$

showing that α is finite. By Lemma 1, there is a sequence $\{v_k\}$ such that

$$PI_V(v_k) \rightarrow \alpha, \quad (1 + \|v_k\|_H) \|I'_V(v_k)\| \rightarrow 0. \tag{24}$$

By Lemma 4, it has a subsequence converging in H . Consequently, there is a $v \in H$ such that $PI_V(v) = \alpha$ and $I'_V(v) = 0$. We must show that $v \neq 0$. Let

$$\psi(t) = \lambda |\Omega| t + P \int_{\Omega} \ln(1 + V + t) dx.$$

Then,

$$\psi(0) = P \int_{\Omega} \ln(1 + V) dx, \quad \psi'(0) = \lambda |\Omega| + P \int_{\Omega} \frac{1}{1 + V} dx < 0.$$

Hence, there is a constant $t > 0$ such that $\psi(t) < \psi(0)$. Thus, there is a constant c such that

$$\begin{aligned} PI_V(v) &= \alpha \leq PI_V(c) = \psi(c^2) \\ &< \psi(0) = P \int_{\Omega} \ln(1 + V) dx = PI_V(0). \end{aligned}$$

This shows that $v \neq 0$. Thus, we have a solution $v \neq 0$ of (1). This completes the proof.

4 Proof of Theorem 3

First, we show that

$$\alpha = \liminf_{\|u\|_H \rightarrow 0} [PI_V(u) - PI_V(0)]/\|u\|_H^2 > 0, \quad u \in H.$$

Suppose $v_k \in H$ is a sequence such that $\rho_k = \|v_k\|_H \rightarrow 0$ and

$$PI_V(v_k) - PI_V(0)]/\|v_k\|^2 \rightarrow \alpha.$$

Let $\tilde{v}_k = v_k/\rho_k$. Then, $\|\tilde{v}_k\|_H = 1$. Hence, there is a renamed subsequence such that $\tilde{v}_k \rightharpoonup \tilde{v}$ in H , and $\tilde{v}_k \rightarrow \tilde{v}$ in $L^2(\Omega)$ and a.e., and

$$PI_V(v_k) - PI_V(0)]/\|v_k\|_H^2 \rightarrow \alpha.$$

By Lemma 8,

$$\int_{\Omega} [\ln(1 + V + v_k^2) - \ln(1 + V)]dx/\|v_k\|_H^2 \rightarrow \int_{\Omega} \frac{1}{1 + V} \tilde{v}^2(x)dx. \tag{25}$$

Hence, we have

$$\begin{aligned} PI_V(v_k) - PI_V(0)]/\|v_k\|_H^2 &= 1 + (\lambda - 1)\|\tilde{v}_k\|^2 \\ &\quad + \int_{\Omega} [\ln(1 + V + v_k^2) - \ln(1 + V)]dx/\rho_k^2 \\ &\rightarrow 1 + (\lambda - 1)\|\tilde{v}\|^2 + \int_{\Omega} \frac{1}{1 + V} \tilde{v}^2(x)dx \\ &= \alpha. \end{aligned}$$

Thus,

$$(1 - \|\tilde{v}\|_H^2) + \|\nabla \tilde{v}\|^2 + \int_{\Omega} [\lambda + P/(1 + V)]\tilde{v}^2 dx = \alpha.$$

Since each of these terms is nonnegative, we see that $\alpha \geq 0$. Moreover, the only way it can vanish is if $1 = \|\tilde{v}\|_H^2$, $\|\nabla \tilde{v}\| = 0$, and

$$\int_{\Omega} [\lambda + P/(1 + V)]\tilde{v}^2 dx = 0.$$

But this would mean that $\tilde{v} = c$, $c^2|\Omega| = 1$, and

$$\lambda|\Omega| + P \int_{\Omega} \frac{1}{1+V} dx = 0,$$

contrary to hypothesis. Since $\alpha > 0$, there are positive numbers ρ, ε such that

$$PI_V(v) - PI_V(0) > \varepsilon, \quad \|v\|_H = \rho.$$

Moreover, we have

$$PI_V(c) = c^2[\lambda|\Omega| + \int_{\Omega} \frac{\ln(1+V+c^2)}{c^2} dx] \rightarrow -\infty, \quad c \rightarrow \infty.$$

Thus, we have functions $v_1, v_2 \in H$ such that $\|v_1\|_H < \rho$, $\|v_2\|_H > \rho$, and

$$PI_V(v_i) < \inf_{\|u\|_H=\rho} PI_V(u).$$

We can now apply Lemmas 2, 3, and 4 to reach the desired conclusion.

References

1. G. Bartal, O. Manela, O. Cohen, J.W. Fleischer, M. Segev, Observation of second-band vortex solitons in 2D photonic lattices. *Phys. Rev. Lett.* **95**, 053904 (2005)
2. S. Chen, Y. Lei, Existence of steady-state solutions in a nonlinear photonic lattice model. *J. Math. Phys.* **52**(6), 063508 (2011)
3. W. Chen, D.L. Mills, Gap solitons and the nonlinear optical response of superlattices. *Phys. Rev. Lett.* **62**, 1746–1749 (1989)
4. N.K. Efremidis, S. Sears, D.N. Christodoulides, Discrete solitons in photorefractive optically-induced photonic lattices. *Phys. Rev. Lett.* **85**, 1863–1866 (2000)
5. J.W. Fleischer, M. Segev, N.K. Efremidis, D.N. Christodoulides, Observation of two-dimensional discrete solitons in optically induced nonlinear photonic lattices. *Nature* **422**, 147–149 (2003)
6. J.W. Fleischer, G. Bartal, O. Cohen, O. Manela, M. Segev, J. Hudock, D.N. Christodoulides, Observation of vortex-ring discrete solitons in 2D photonic lattices. *Phys. Rev. Lett.* **92**, 123904 (2004)
7. H. Martin, E.D. Eugenieva, Z. Chen, Discrete solitons and soliton-induced dislocations in partially coherent photonic lattices. *Martin et al. Phys. Rev. Lett.* **92**, 123902 (2004)
8. D.N. Neshev, T.J. Alexander, E.A. Ostrovskaya, Y.S. Kivshar, H. Martin, I. Makasyuk, Z. Chen, Observation of discrete vortex solitons in optically induced photonic lattices. *Phys. Rev. Lett.* **92**, 123903 (2004)
9. M. Schechter, *Linking Methods in Critical Point Theory* (Birkhauser, Boston, 1999)
10. M. Schechter, in *An Introduction to Nonlinear Analysis*. Cambridge Studies in Advanced Mathematics, vol. 95 (Cambridge University Press, Cambridge, 2004)
11. M. Schechter, The use of Cerami sequences in critical point theory. *Abstr. Appl. Anal.* **2007**, 58948 (2007)
12. M. Schechter, *Minimax Systems and Critical Point Theory* (Birkhauser, Boston, 2009)
13. M. Schechter, Steady state solutions for Schrödinger equations governing nonlinear optics. *J. Math. Phys.* **53**, 043504 (2012)

14. M. Schechter, *Critical Point Theory, Sandwich and Linking Systems* (Birkhauser, Boston, 2020)
15. Y. Yang, *Soliton in Field Theory and Nonlinear Analysis* (Springer, New York, 2001)
16. J. Yang, A. Bezryadina, Z. Chen, I. Makasyuk. Observation of two-dimensional lattice vector solitons. *Opt. Lett.* **29**, 1656 (2004)
17. J. Yang, I. Makasyuk, A. Bezryadina, Z. Chen, in *Dipole and Quadrupole Solitons in Optically Induced Two-Dimensional Photonic Lattices: Theory and Experiment*. *Studies in Applied Mathematics*, vol. 113 (2004) pp. 389–412
18. Y. Yang, R. Zhang. Steady state solutions for nonlinear Schrödinger equation arising in optics. *J. Math. Phys.* **50**, 053501-9 (2009)
19. Y. Yang, R. Zhang. Erratum. Steady state solutions for nonlinear Schrödinger equation arising in optics. *J. Math. Phys.* **50**, 053501-9 (2009). *J. Math. Phys.* **51**, 049902 (2010)

Ekeland Variational Principles in 2-Local Branciari Metric Spaces



Mihai Turinici

Abstract An Ekeland Variational Principle is stated over a class of local and 2-local Branciari metric spaces, and its relationships with the Dependent Choice Principle are discussed. Applications to Caristi–Kirk fixed point theorems over such a setting are also being considered.

AMS Subject Classification 49J53 (Primary), 54H25 (Secondary)

1 Introduction

Let X be a nonempty set, and $d : X \times X \rightarrow R_+ := [0, \infty[$ be a *metric* over it; then (X, d) is called a *metric space*. Further, let $\varphi : X \rightarrow R \cup \{\infty\}$ be *regular*:

- (r-1) φ is inf-proper ($\text{Dom}(\varphi) \neq \emptyset$ and $\inf[\varphi(X)] > -\infty$),
- (r-2) φ is d -lsc ($\liminf_n \varphi(x_n) \geq \varphi(x)$, whenever $x_n \xrightarrow{d} x$).

The following 1974 statement in Ekeland [16] (referred to as Ekeland’s variational principle; in short, EVP) is our starting point.

Theorem 1 *Let the precise conditions hold; and X be d -complete. Then, for each $u \in \text{Dom}(\varphi)$, there exists $v = v(u) \in \text{Dom}(\varphi)$, with*

- (11-a) $d(u, v) \leq \varphi(u) - \varphi(v)$ (hence $\varphi(u) \geq \varphi(v)$),
- (11-b) $d(v, x) > \varphi(v) - \varphi(x)$, for all $x \in X \setminus \{v\}$.

M. Turinici (✉)

A Myller Mathematical Seminar, A I Cuza University, Iași, Romania
e-mail: mturi@uaic.ro

© Springer Nature Switzerland AG 2021

T. M. Rassias, P. M. Pardalos (eds.), *Nonlinear Analysis and Global Optimization*,
Springer Optimization and Its Applications 167,
https://doi.org/10.1007/978-3-030-61732-5_23

461

Note that with respect to the Brøndsted (quasi-) order [8]

(Br-ord) $(x, y \in X) x \leq y$ iff $d(x, y) + \varphi(y) \leq \varphi(x)$,

the point $v \in X$ appearing in (11-b) is *maximal*; so that (EVP) is nothing but a variant of the Zorn–Bourbaki maximal statement [5, 45] in the way proposed by Brezis–Browder ordering principle [7] (in short, BB); hence, (EVP) is deductible from (BB). Concerning the reverse inclusion, note that (BB) is obtainable from the Dependent Choice Principle (in short, DC) due to Bernays [4] and Tarski [35] on the one hand, and (EVP) implies (DC), on the other hand; hence, summing up (cf. Brunner [9] and Turinici [42])

(DC) \implies (BB) \implies (EVP) \implies (DC)

(wherefrom, all these principles are mutually equivalent).

As a consequence of such practical and theoretical conclusions, (EVP) found some basic applications to control and optimization, generalized differential calculus, critical point theory, and global analysis; we refer to the 1979 paper by Ekeland [17] for a survey of these. So, it cannot be surprising that, soon after its formulation, many extensions of (EVP) were proposed. For example, the (*pseudo-*) *metrical* one consists of the conditions imposed upon the ambient metric d over X being relaxed. A basic result in this direction has been stated in Tataru [36], via Ekeland-type techniques; subsequent extensions of it were obtained by Altman [2], Turinici [37], Kang and Park [23], and Kada et al. [22], among many others. Since all these are obtainable from (DC), it follows from the above that a deduction of them from (EVP) is possible; see Turinici [42] for details. Further, a *functional* extension of (EVP) was obtained by Zhong [44] and refined in Bao and Khanh [3]; however, as precise by Turinici [40], it is nothing but a variant of (EVP). Finally, the *dimensional* way of extension refers to the ambient space (R) of $\varphi(X)$ being substituted by a (topological or not) vector space; an account of the results in this area is to be found in the 1986 paper by Nemeth [30], and the 2003 monograph by Goepfert et al. [18, Ch 3]; see also Chen et al. [11, 12]. Note that the scalarization-type method used there allows us to reducing most “sequential” statements in the area to (BB) [hence, ultimately, to (EVP)]. However, this device cannot cover the 1989 variational principle in Khanh [25]; but, for “higher order” versions of (DC) taken as in Wolk [43], this works.

In the following, a variant of Ekeland Variational Principle is formulated over a class of local Branciari metric spaces introduced as in Turinici [41]. The outlines of this method were implicitly discussed (under the Caristi–Kirk setting [10]) in a recent paper by Alamri et al. [1], at the level of Branciari metric spaces [6]; but, the proposed setting seems to be new. As we will see, the obtained variational principle is nothing but a logical equivalent of the standard (EVP); however, it may be useful in practice. Further aspects of these facts will be discussed elsewhere.

2 Dependent Choice Principles

Throughout this exposition, the axiomatic system in use is Zermelo–Fraenkel’s (abbreviated: ZF); cf. Cohen [13, Ch 2]. The notations and basic facts to be considered are standard. Some important ones are described below.

(A) Let X be a nonempty set. By a *relation* over X , we mean any (nonempty) part $\mathcal{R} \subseteq X \times X$; then, (X, \mathcal{R}) will be referred to as a *relational structure*. For simplicity, we sometimes write $(x, y) \in \mathcal{R}$ as $x\mathcal{R}y$. Note that \mathcal{R} may be regarded as a mapping between X and $\exp[X]$ (=the class of all subsets in X). In fact, denote

$$X(x, \mathcal{R}) = \{y \in X; x\mathcal{R}y\} \text{ (the section of } \mathcal{R} \text{ through } x), x \in X:$$

then, the desired mapping representation is $(\mathcal{R}(x) = X(x, \mathcal{R}); x \in X)$. A basic example of such object is

$$\mathcal{I} = \{(x, x); x \in X\} \text{ [the identical relation over } X \text{].}$$

Given the relations \mathcal{R}, \mathcal{I} over X , define their *product* $\mathcal{R} \circ \mathcal{I}$ as

$$(x, z) \in \mathcal{R} \circ \mathcal{I}, \text{ if there exists } y \in X \text{ with } (x, y) \in \mathcal{R}, (y, z) \in \mathcal{I}.$$

Also, for each relation \mathcal{R} on X , denote

$$\mathcal{R}^{-1} = \{(x, y) \in X \times X; (y, x) \in \mathcal{R}\} \text{ (the inverse of } \mathcal{R}\text{).}$$

Finally, given the relations \mathcal{R} and \mathcal{I} over X , let us say that \mathcal{R} is *coarser* than \mathcal{I} (or, equivalently: \mathcal{I} is *finer* than \mathcal{R}), provided

$$\mathcal{R} \subseteq \mathcal{I}, \text{ that is, } x\mathcal{R}y \text{ implies } x\mathcal{I}y.$$

Given a relation \mathcal{R} on X , the following properties are to be discussed here:

- (P1) \mathcal{R} is *reflexive*: $\mathcal{I} \subseteq \mathcal{R}$,
- (P2) \mathcal{R} is *irreflexive*: $\mathcal{R} \cap \mathcal{I} = \emptyset$,
- (P3) \mathcal{R} is *transitive*: $\mathcal{R} \circ \mathcal{R} \subseteq \mathcal{R}$,
- (P4) \mathcal{R} is *symmetric*: $\mathcal{R}^{-1} = \mathcal{R}$,
- (P5) \mathcal{R} is *antisymmetric*: $\mathcal{R}^{-1} \cap \mathcal{R} \subseteq \mathcal{I}$.

This yields the classes of relations to be used; the following ones are important for our developments:

- (C0) \mathcal{R} is *amorphous* (i.e., it has no specific properties),
- (C1) \mathcal{R} is a *quasi-order* (reflexive and transitive),
- (C2) \mathcal{R} is a *strict order* (irreflexive and transitive),
- (C3) \mathcal{R} is an *equivalence* (reflexive, transitive, symmetric),
- (C4) \mathcal{R} is a (*partial*) *order* (reflexive, transitive, antisymmetric),
- (C5) \mathcal{R} is the *trivial relation* (i.e., $\mathcal{R} = X \times X$).

(B) A basic example of relational structure is to be constructed as below. Let

$$N := \{0, 1, \dots\}, \text{ where } (0 = \emptyset, 1 = \{0\}, 2 = \{0, 1\}, \dots),$$

denote the set of *natural* numbers. Technically speaking, the basic (algebraic and order) structures over N may be obtained by means of the (*immediate*) *successor* function $\text{suc} : N \rightarrow N$, and the following Peano properties (deductible in our axiomatic system (ZF)):

(pea-1) $(0 \in N \text{ and } 0 \notin \text{suc}(N))$,

(pea-2) $\text{suc}(\cdot)$ is injective ($\text{suc}(n) = \text{suc}(m)$ implies $n = m$),

(pea-3) if $M \subseteq N$ is such that $[0 \in M]$ and $[\text{suc}(M) \subseteq M]$, then $M = N$.

[Note that, in the absence of our axiomatic setting, these properties become the well-known Peano axioms, as described in Halmos [19, Ch 12]; we do not give details.] In fact, starting from these properties, one may construct, in a recurrent way, an *addition* $(a, b) \mapsto a + b$ over N , according to

$$(\forall m \in N): m + 0 = m; m + \text{suc}(n) = \text{suc}(m + n).$$

This, in turn, makes possible the introduction of a (partial) order relation (\leq) over N , as

$$(m, n \in N): m \leq n \text{ iff } m + p = n, \text{ for some } p \in N.$$

Concerning the properties of this structure, the most important one writes

(N, \leq) is well ordered:

any (nonempty) subset of N has a first element.

hence, in particular, (N, \leq) is (partially) ordered. Denote, for simplicity,

$$N(r, \leq) = \{n \in N; r \leq n\} = \{r, r + 1, \dots\}, r \in N,$$

$$N(r, >) = \{n \in N; r > n\} = \{0, \dots, r - 1\}, r \in N(1, \leq);$$

the latter one is referred to as the *initial interval* (in N) induced by r . Any set P with $P \sim N$ (in the sense: there exists a bijection from P to N) will be referred to as *effectively denumerable*. In addition, given some natural number $n \geq 1$, any set Q with $Q \sim N(n, >)$ will be said to be *n-finite*; when n is generic here, we say that Q is *finite*. Finally, the (nonempty) set Y is called (at most) *denumerable* iff it is either effectively denumerable or finite.

Let X be a nonempty set. By a *sequence* in X , we mean any mapping $x : N \rightarrow X$; where, as already precise, $N := \{0, 1, \dots\}$ is the set of *natural* numbers. For simplicity reasons, it will be useful to denote it as $(x(n); n \geq 0)$, or $(x_n; n \geq 0)$; moreover, when no confusion can arise, we further simplify this notation as $(x(n))$ or (x_n) , respectively. Given such an object, $(x_n; n \geq 0)$, any sequence $(y_n := x_{i(n)}; n \geq 0)$ with $(i(n)) = \text{divergent}$ ($i(n) \rightarrow \infty$ as $n \rightarrow \infty$) will be referred to as a *subsequence* of $(x_n; n \geq 0)$. Note that the relation “subsequence of” is transitive:

$(z_n) = \text{subsequence of } (y_n)$ and $(y_n) = \text{subsequence of } (x_n)$

imply $(z_n) = \text{subsequence of } (x_n)$.

(C) 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 $F : J \rightarrow \exp(X)$, there exists a (selective) function $f : J \rightarrow X$, with $f(v) \in F(v)$, for each $v \in J$.

(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 all choice reasonings; and the 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 \mathcal{R} over X , *proper* when

$$(X(x, \mathcal{R}) =) \mathcal{R}(x) \text{ is nonempty, for each } x \in X.$$

Then, \mathcal{R} is to be viewed as a mapping between X and $\exp(X)$, and the couple (X, \mathcal{R}) will be referred to as a *proper relational structure*. Further, given $a \in X$, let us say that the sequence $(x_n; n \geq 0)$ in X is $(a; \mathcal{R})$ -*iterative*, provided

$$x_0 = a \text{ and } x_n \mathcal{R} x_{n+1} \text{ (i.e., } x_{n+1} \in \mathcal{R}(x_n)), \text{ for all } n.$$

Proposition 1 *Let the relational structure (X, \mathcal{R}) be proper. Then, for each $a \in X$, there is at least one (a, \mathcal{R}) -iterative sequence in X .*

This principle—proposed, independently, by Bernays [4] and Tarski [35]—is deductible from (AC), but not conversely; cf. Wolk [43]. Moreover, by the developments in Moskhovakis [29, Ch 8] and Schechter [33, Ch 6], the reduced system (ZF-AC+DC) is comprehensive enough so as to cover the “usual” mathematics; see also Moore [28, Appendix 2].

Let $(\mathcal{R}_n; n \geq 0)$ 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 $(a; (\mathcal{R}_n; n \geq 0))$ -*iterative*, provided

$$x_0 = a \text{ and } x_n \mathcal{R}_n x_{n+1} \text{ (i.e., } x_{n+1} \in \mathcal{R}_n(x_n)), \text{ for all } n.$$

The following *Diagonal Dependent Choice* principle (in short, DDC) is available.

Proposition 2 *Let $(\mathcal{R}_n; n \geq 0)$ be a sequence of proper relations on X . Then, for each $a \in X$, there exists at least one $(a; (\mathcal{R}_n; n \geq 0))$ -iterative sequence in X .*

Clearly, (DDC) includes (DC), to which it reduces when $(\mathcal{R}_n; n \geq 0)$ is constant. The reciprocal of this is also true. In fact, letting the premises of (DDC) hold, put $P = N \times X$, and let \mathcal{S} be the relation over P introduced as

$$\mathcal{S}(i, x) = \{i + 1\} \times \mathcal{R}_i(x), \text{ (} i, x) \in P.$$

It will suffice applying (DC) to (P, \mathcal{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 $F : N \rightarrow \exp(X)$ and the relation \mathcal{R} over X fulfill*

$$(\forall n \in N): \mathcal{R}(x) \cap F(n + 1) \neq \emptyset, \text{ for all } x \in F(n).$$

Then, for each $a \in F(0)$, there exists a sequence $(x(n); n \geq 0)$ in X , with

$$x(0) = a, x(n) \in F(n), x(n + 1) \in \mathcal{R}(x(n)), \forall n.$$

As before, $(\text{SDC}) \implies (\text{DC}) (\iff (\text{DDC}))$; just take $(F(n) = X, n \in N)$. But, the reciprocal is also true, in the sense: $(\text{DDC}) \implies (\text{SDC})$. This follows from

Proof of Proposition 3 Let the premises of (SDC) be true. Define a sequence of relations $(\mathcal{R}_n; n \geq 0)$ over X as: for each $n \geq 0$,

$$\begin{aligned} \mathcal{R}_n(x) &= \mathcal{R}(x) \cap F(n + 1), \text{ if } x \in F(n), \\ \mathcal{R}_n(x) &= \{x\}, \text{ otherwise } (x \in X \setminus F(n)). \end{aligned}$$

Clearly, \mathcal{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 $(a, (\mathcal{R}_n; n \geq 0))$ -iterative sequence $(x(n); n \geq 0)$ in X . This, along with the very definition above, gives all desired conclusions.

In particular, when $\mathcal{R} = X \times X$, the regularity condition imposed in (SDC) holds. The corresponding variant of the underlying statement is just $(\text{AC}(N))$ (=the *Denumerable Axiom of Choice*). Precisely, we have

Proposition 4 Let $F : N \rightarrow \exp(X)$ be a function. Then, for each $a \in F(0)$, there exists a function $f : N \rightarrow X$ with $f(0) = a$ and $f(n) \in F(n), \forall n \in N$.

As a consequence of the above facts, $(\text{DC}) \implies (\text{AC}(N))$ in (ZF-AC) . A direct verification of this is obtainable by taking $Q = N \times X$ and introducing the relation \mathcal{R} over it, according to:

$$\mathcal{R}(n, x) = \{n + 1\} \times F(n + 1), n \in N, x \in X;$$

we do not give details. The reciprocal of the written inclusion is not true; see Moskhovakis [29, Ch 8, Sect 8.25] for details.

3 Conv-Cauchy Structures

Let X be a nonempty set. 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 $Y = \{y\}$, for some $y \in X$.

Let $\mathcal{S}(X)$ stand for the class of all sequences (x_n) in X . By a (sequential) *convergence structure* on X , we mean any part \mathcal{C} of $\mathcal{S}(X) \times X$, with the properties (cf. Kasahara [24]):

- (conv-1) \mathcal{C} is *hereditary*:
 $((x_n); x) \in \mathcal{C} \implies ((y_n); x) \in \mathcal{C}$, for each subsequence (y_n) of (x_n) ,
- (conv-2) \mathcal{C} is *reflexive*: for each $u \in X$,
the constant sequence $(x_n = u; n \geq 0)$ fulfills $((x_n); u) \in \mathcal{C}$.

For (x_n) in $\mathcal{S}(X)$ and $x \in X$, we write $((x_n); x) \in \mathcal{C}$ as $x_n \xrightarrow{\mathcal{C}} x$; this reads

(x_n) , \mathcal{C} -converges to x (also referred to as: x is the \mathcal{C} -limit of (x_n)).

The set of all such x is denoted $\mathcal{C} - \lim_n(x_n)$; when it is nonempty, we say that (x_n) is \mathcal{C} -convergent. The following condition is to be optionally considered here:

(conv-3) \mathcal{C} is separated:

$\mathcal{C} - \lim_n(x_n)$ is an asingleton, for each sequence (x_n) ;

when it holds, $x_n \xrightarrow{\mathcal{C}} z$ will be also written as $\mathcal{C} - \lim_n(x_n) = z$.

Further, by a (sequential) *Cauchy structure* on X , we shall mean any part \mathcal{H} of $\mathcal{S}(X)$ with (cf. Turinici [38]):

(Cauchy-1) \mathcal{H} is hereditary:

$(x_n) \in \mathcal{H} \implies (y_n) \in \mathcal{H}$, for each subsequence (y_n) of (x_n)

(Cauchy-2) \mathcal{H} is reflexive: for each $u \in X$,

the constant sequence $(x_n = u; n \geq 0)$ fulfills $(x_n) \in \mathcal{H}$.

Each element of \mathcal{H} will be referred to as a \mathcal{H} -Cauchy sequence in X .

Finally, given the couple $(\mathcal{C}, \mathcal{H})$ as before, we shall say that it is a *conv-Cauchy structure* on X . The natural conditions about the conv-Cauchy structure $(\mathcal{C}, \mathcal{H})$ to be considered here are

(CC-1) $(\mathcal{C}, \mathcal{H})$ is regular: each \mathcal{C} -convergent sequence is \mathcal{H} -Cauchy,

(CC-2) $(\mathcal{C}, \mathcal{H})$ is complete: each \mathcal{H} -Cauchy sequence is \mathcal{C} -convergent.

A standard way of introducing such structures is the (*pseudo*) *metrical* one. Let $d : X \times X \rightarrow R_+$ be a (r -*s*)-symmetric over X ; i.e.,

(rss-1) d is symmetric: $d(x, y) = d(y, x), \forall x, y \in X$,

(rss-2) d is reflexive sufficient: $x = y \iff d(x, y) = 0$;

in this case, (X, d) is called a (r -*s*)-symmetric space. 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 $d(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$; i.e.,

$\forall \varepsilon > 0, \exists i = i(\varepsilon) : i \leq n \implies d(x_n, x) < \varepsilon$.

By this very definition, we have the hereditary and reflexive properties:

(d-conv-1) (\xrightarrow{d}) is hereditary)

$x_n \xrightarrow{d} x$ implies $y_n \xrightarrow{d} x$, for each subsequence (y_n) of (x_n) ,

(d-conv-2) (\xrightarrow{d}) is reflexive) for each $u \in X$,

the constant sequence $(x_n = u; n \geq 0)$ fulfills $x_n \xrightarrow{d} u$.

As a consequence, (\xrightarrow{d}) is a sequential convergence on X . The set of all such limit points of (x_n) will be denoted $\lim_n(x_n)$; if it is nonempty, then (x_n) is called d -convergent. The following condition about this structure is to be considered:

(d-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, by the conditions imposed upon d , this is not in general true. However, under the extra property

(tri) d is triangular: $d(x, y) \leq d(x, z) + d(z, y), \forall x, y, z \in X$;

(when d becomes a *metric* on X), the separated property holds.

Further, call the sequence (x_n) , d -Cauchy when

$$\lim_n \sup\{d(x_n, x_{n+m}); m \geq 1\} = 0; \text{ that is,}$$

$$\forall \varepsilon > 0, \exists j = j(\varepsilon): j \leq n < p \implies d(x_n, x_p) < \varepsilon;$$

the class of all these will be denoted as $Cauchy(d)$. As before, we have the hereditary and reflexive properties:

- (d-Cauchy-1) ($Cauchy(d)$ is hereditary)
 (x_n) is d -Cauchy implies (y_n) is d -Cauchy,
 for each subsequence (y_n) of (x_n) ,
- (d-Cauchy-2) ($Cauchy(d)$ is reflexive) for each $u \in X$,
 the constant sequence $(x_n = u; n \geq 0)$ is d -Cauchy;

hence, $Cauchy(d)$ is a Cauchy structure on X .

Now—according to the general setting—call the couple $((\xrightarrow{d}), Cauchy(d))$, a *conv-Cauchy structure* induced by d . 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.

Generally, none of these is holding under our framework; however, the former one is retainable if (in addition) d is triangular (see above).

Let again (X, d) be a (r-s)-symmetric space. In the following, some classes of sequences (related to the d -Cauchy ones) are introduced:

(I) Let K be some nonempty subset of $N(1, \leq)$. Given the sequence (x_n) , call it (d, K) -asymptotic, provided

$$\lim_n d(x_n, x_{n+i}) = 0, \text{ for each } i \in K.$$

In particular, the $(d, \{1\})$ -asymptotic property will be called d -asymptotic; and the $(d, N(1, \leq))$ -asymptotic one is referred to as d -strongly-asymptotic. Clearly,

$$\text{(for each sequence } (x_n) \text{ in } X):$$

$$d\text{-Cauchy} \implies d\text{-strongly-asymptotic} \implies d\text{-asymptotic};$$

but, none of the converse properties is available. Concerning this aspect, a basic situation to be discussed is the metrical one.

Proposition 5 *Supposed that, in addition, d is triangular (hence, a metric) on X . Then, for each sequence (x_n) in X ,*

$$(x_n) \text{ is } d\text{-asymptotic iff } (x_n) \text{ is } d\text{-strongly-asymptotic.}$$

Proof Let $i \in N(1, \leq)$ be arbitrary fixed. By the triangular inequality,

$$d(x_n, x_{n+i}) \leq \rho_n + \dots + \rho_{n+i-1}, \forall n, \text{ where } (\rho_n = d(x_n, x_{n+1}); n \geq 0).$$

By the imposed hypothesis, the right member of this relation tends to zero as $n \rightarrow \infty$; wherefrom, all is clear.

(II) Given $\nu \in N(1, \leq)$, let us say that (x_n) is (d, ν) -Cauchy, provided

for each $\varepsilon > 0$, there exists $h = h(\varepsilon) \geq 0$, such that $n \geq h$ and $j \geq 0$ imply $d(x_n, x_{n+1+j\nu}) < \varepsilon$.

By this definition, we have

(for each sequence (x_n) in X):
 d -Cauchy implies (d, ν) -Cauchy, for each $\nu \in N(1, \leq)$;

but, the converse is not in general true. Concerning this aspect, a natural question is to establish conditions under which the reciprocal inclusion holds. The simplest one is again d -triangular; as results from

Proposition 6 *Supposed that, in addition, d is triangular (hence, a metric) on X . Then, for each sequence (x_n) in X , and each $\nu \geq 1$,*

(x_n) is d -Cauchy iff (x_n) is (d, ν) -Cauchy.

Proof The case $\nu = 1$ is clear; so, without loss, one may assume that $\nu \geq 2$. Suppose that (x_n) is (d, ν) -Cauchy; hence, in particular, d -asymptotic. Given $\varepsilon > 0$, there must be some $h = h(\varepsilon) \geq 0$, such that

$d(x_n, x_{n+1+j\nu}) < \varepsilon/3\nu$, for all $n \geq h, j \geq 0$;
 hence, in particular, $\rho_n := d(x_n, x_{n+1}) < \varepsilon/3\nu$, for all $n \geq h$.

Let $m > n$ be arbitrary fixed. The case of $m \in \{n + 1 + j\nu; j \geq 0\}$ yields

$$d(x_n, x_m) = d(x_n, x_{n+1+j\nu}) < \varepsilon/3\nu < \varepsilon.$$

It remains the case of $m \notin \{n + 1 + j\nu; j \geq 0\}$; when, one has the representation

$$m = n + 1 + j\nu + k, \text{ for some } j \geq 0, k \in \{1, \dots, \nu - 1\}.$$

In this case, the triangular inequality gives (under the notation $q = n + 1 + j\nu$)

$$d(x_n, x_m) \leq d(x_n, x_q) + \rho_q + \dots + \rho_{q+k-1} < (k + 1)\varepsilon/3\nu \leq \varepsilon/3 < \varepsilon,$$

and the conclusion follows.

(III) Let us say that the sequence (x_n) is d -telescopic Cauchy, when

$$\sum_n d(x_n, x_{n+1}) (= d(x_0, x_1) + d(x_1, x_2) + \dots) < \infty.$$

The relationship with the d -Cauchy property is to be clarified in a triangular setting.

Proposition 7 *Supposed that, in addition, d is triangular (hence, a metric) on X . Then, for each sequence (x_n) in X ,*

(x_n) is d -telescopic Cauchy implies (x_n) is d -Cauchy.

The reciprocal is not in general true.

Proof

(i) By hypothesis, we have under the notation $(\rho_n := d(x_n, x_{n+1}); n \geq 0)$

$$\sigma_n := \sum_{k \geq n} \rho_k \rightarrow 0 \text{ as } n \rightarrow \infty.$$

In this case, for each $n, m \geq 0$ with $n < m$,

$$d(x_n, x_m) \leq \rho_n + \rho_{n+1} + \dots + \rho_{m-1} \leq \sigma_n,$$

and this, along with the property of (σ_n) , gives the needed conclusion.

(ii) Let (X, d) be the real axis with its standard metric. The (real) sequence $(x_n = (-1)^n/(n + 1); n \geq 0)$ fulfills

$$\lim_n(x_n) = 0; \text{ hence, } (x_n) \text{ is } d\text{-Cauchy.}$$

On the other hand, by the properties of harmonic series,

$$\sum_n d(x_n, x_{n+1}) = \sum_n [1/(n + 1) + 1/(n + 2)] \geq \sum_n 1/(n + 1) = \infty;$$

hence, (x_n) is not d -telescopic Cauchy. The proof is thereby complete.

4 Local and 2-Local Branciari Metric Spaces

In the following, some technical facts about local and 2-local Branciari metric spaces are being discussed. Their exposition will necessitate some conventions and auxiliary facts.

Let X be a nonempty set, and $d : X \times X \rightarrow R_+$ be a mapping with

(symm) d is symmetric [$d(x, y) = d(y, x), \forall x, y \in X$];

(r-s) d is reflexive sufficient [$x = y \iff d(x, y) = 0$];

we then say that d is a reflexive sufficient symmetric (in short, $(r-s)$ -symmetric), and (X, d) is a $(r-s)$ -symmetric space.

Given $k \geq 1$, any ordered system $C = (x_1, \dots, x_k)$ in X^k will be called a k -chain of X ; the class of all these is denoted as $\text{chain}(X; k)$. Given such an object, put $[C] = \{x_1, \dots, x_k\}$ (the set of all points belonging to this k -chain); clearly, the alternative $\text{card}([C]) < k$ (when $k > 1$) cannot be avoided. If $\text{card}([C]) = k$, then C will be referred to as a full k -chain (in X); denote the class of all these as $\text{fchain}(X; k)$. In particular, any point $a \in X$ may be identified with a full 1-chain of X . For any $C \in \text{chain}(X; k)$, where $k \geq 2$, denote

$$\Lambda(C) = d(x_1, x_2) + \dots + d(x_{k-1}, x_k), \text{ if } C = (x_1, \dots, x_k)$$

(the “length” of C). Given $h \geq 1$ and the h -chain $D = (y_1, \dots, y_h)$ in X , let $(C; D)$ stand for the $(k + h)$ -chain $E = (z_1, \dots, z_{k+h})$ in X introduced as

$$z_i = x_i, 1 \leq i \leq k; z_{k+j} = y_j, 1 \leq j \leq h;$$

it will be referred to as the “product” between C and D . This operation may be extended to a finite family of such objects.

Let (X, d) be a $(r-s)$ -symmetric space. For an efficient handling of convergence and Cauchy properties, a sort of triangular condition upon d must be added. The simplest one writes

(1-tri) d is 1-triangular [$d(x, y) \leq d(x, z) + d(z, y), \forall x, y, z \in X$];

in this case, (X, d) will be called a 1-triangular metric space. Clearly, 1-triangular $(r-s)$ -symmetric is nothing else than *triangular* $(r-s)$ -symmetric; and 1-triangular metric space is identical with *(standard) metric space*. A natural extension of 1-triangular condition is the one introduced as in Branciari [6]

(nu-tri) $d(., .)$ is ν -triangular (where $\nu \geq 1$):

$$d(x, y) \leq \Lambda(x; C; y), \text{ for all } x, y \in X,$$

and all $C \in \text{fchain}(X; \nu)$, with $(x; C; y) \in \text{fchain}(X; \nu + 2)$;

then, the couple (X, d) is referred to as a ν -triangular metric space. When $\nu \geq 1$ is generic here, we say that $d(., .)$ is a *Branciari metric* on X ; and (X, d) will be called a *Branciari metric space*.

A local version of these conventions is as follows. Given $M \in \text{exp}(X)$, let us say that $h \in N(1, \leq)$ is a *Branciari constant* for it, provided

$$d(x, y) \leq \Lambda(x; C; y), \text{ for all } x, y \in M, x \neq y,$$

and all $C \in \text{fchain}(M; h)$, with $(x; C; y) \in \text{fchain}(M; h + 2)$.

Denote, for simplicity,

(B-M) $\mathcal{B}(M)$ = the class of all Branciari constants for M ,

(B-M-min) $\nu_{\mathcal{B}}(M) = \min \mathcal{B}(M)$ (where $\min(\emptyset) = -\infty$);

the latter of these will be referred to as the *minimal Branciari constant* for M . When $\mathcal{B}(M) \neq \emptyset$, we say that $M \in \text{exp}(X)$ is *Branciari compatible*; then, clearly, $\nu_{\mathcal{B}}(M)$ exists as an element of $N(1, \leq)$.

Having these precise, let us say that the $(r-s)$ -symmetric $d(., .)$ is a *local Branciari metric* when it has the “local” triangular property:

each effectively denumerable part $M \in \text{exp}(X)$ is Branciari compatible:

there exists $h \geq 1$ such that $d(x, y) \leq \Lambda(x; C; y)$, for all $x, y \in M, x \neq y$, and all $C \in \text{fchain}(M; h)$, with $(x; C; y) \in \text{fchain}(M; h + 2)$;

in this case, (X, d) will be referred to as a *local Branciari metric space*. It follows by this very definition that (for any $(r-s)$ -symmetric d)

(d is Branciari metric) implies (d is local Branciari metric),

and a corresponding relationship is to be retained between their associated spaces. The reciprocal is not in general true; because, the index appearing in the definition of local Branciari metric $d(., .)$ depends on the (effectively denumerable) subset M of X involved there.

A 2-local version of this concept is the following. Let M be some nonempty part of X . Given $x, y \in M, x \neq y$, let us say that $k \in N(1, \leq)$ is a *Branciari constant* for (x, y) over M , provided

$$d(x, y) \leq \Lambda(x; C; y),$$

for all $C \in \text{fchain}(M; k)$ with $(x; C; y) \in \text{fchain}(M; k + 2)$.

Denote, for $M \in \text{exp}(X)$, and $(x, y \in M, x \neq y)$,

(B-M-2) $\mathcal{L}_M(x, y)$ = the class of all Branciari constants for (x, y) over M ,

(B-M-2-min) $\nu_M(x, y) = \min \mathcal{L}_M(x, y)$ (where $\min(\emptyset) = -\infty$);

the later of these will be referred to as the *minimal Branciari constant* for (x, y) on M . When $\mathcal{L}_M(x, y) \neq \emptyset$, we say that the couple (x, y) is *Branciari compatible* over M ; then, clearly, $\nu_M(x, y)$ exists as an element of $N(1, \leq)$.

Having these precise, let us say that the (r-s)-symmetric $d(., .)$ is a *2-local Branciari metric*, when it has the “2-local” triangular property:

for each effectively denumerable part $M \in \text{exp}(X)$ and each $x, y \in M$ with $x \neq y$, we have that (x, y) is Branciari compatible over M : there exists $k \geq 1$ such that $d(x, y) \leq \Lambda(x; C; y)$, for all $C \in \text{chain}(M; k)$, with $(x; C; y) \in \text{fchain}(M; k + 2)$;

in this case, (X, d) will be referred to as a *2-local Branciari metric space*. It follows by this very definition that (for any (r-s)-symmetric d)

$(d$ is local Branciari metric) implies $(d$ is 2-local Branciari metric);

and a corresponding relationship is to be retained between their associated spaces. The reciprocal inclusion is not in general true; because, the index appearing in the definition of 2-local Branciari metric $d(., .)$ depends on the couples (x, y) taken from the (effectively denumerable) subset M of X involved there.

Now, according to Suzuki [34], a topological study of Branciari metric spaces is not ultimately possible; and the conclusion remains valid for local and 2-local Branciari metric spaces as well. As a consequence, the natural way of handling these objects is the conv-Cauchy one; its basic lines are to be sketched as follows. Let (X, d) be a 2-local Branciari metric space. Define a d -convergence structure (\xrightarrow{d}) and d -Cauchy structure *Cauchy*(d) over X in the above discussed way. Namely, given the sequence (x_n) in X and the point $x \in X$, we say that (x_n) , d -converges to x (written as $x_n \xrightarrow{d} x$) if $d(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$; i.e.,

$$\forall \varepsilon > 0, \exists i = i(\varepsilon): i \leq n \implies d(x_n, x) < \varepsilon.$$

The set of all such points x will be denoted $\lim_n(x_n)$; when it is nonempty, (x_n) is called *d-convergent*. Note that (\xrightarrow{d}) is not (in general) separated even if the 2-triangular inequality holds; precisely—cf. Samet [32]—there exist sequences (x_n) in X with $\lim_n(x_n)$ having at least two (distinct) points. Further, call the sequence (x_n) , *d-Cauchy* when $d(x_m, x_n) \rightarrow 0$ as $m, n \rightarrow \infty, m < n$; i.e.,

$$\forall \varepsilon > 0, \exists j = j(\varepsilon): j \leq m < n \implies d(x_m, x_n) < \varepsilon.$$

Finally, let us say that the couple $((\xrightarrow{d}), \text{Cauchy}(d))$ is the *conv-Cauchy structure* generated by d . Note that (in general) this conv-Cauchy structure is not regular,

even if d is 2-triangular; precisely (cf. the quoted paper), there exist d -convergent sequences that are not d -Cauchy. Finally, call $(x_n; n \geq 0)$

- (asy) d -asymptotic, when $\lim_n d(x_n, x_{n+1}) = 0$,
- (s-asy) d -strongly-asymptotic, if $\lim_n d(x_n, x_{n+i}) = 0, \forall i \geq 1$.

Clearly, the generic relationships are valid

$$(\forall \text{ sequence}): d\text{-Cauchy} \implies d\text{-strongly-asymptotic} \implies d\text{-asymptotic},$$

but the converse relations are not in general true.

As already precise, the (nonempty) set of limit points for a convergent sequence is not in general a singleton. So, we may ask of which supplementary conditions upon this sequence are needed so as to retain such a property. The following answer to this is available. (See also Jleli and Samet [21] or Kirk and Shahzad [26, 27], for a number of related aspects.)

Theorem 2 *Let (X, d) be a 2-local Branciari metric space. Then, for each sequence (x_n) in X ,*

- (41-a) $\lim_n(x_n)$ is an asingleton, if (x_n) is d -asymptotic and full,
- (41-b) $\lim_n(x_n)$ is an asingleton, whenever (x_n) is d -Cauchy.

Proof There are two parts to be discussed.

Part 1. Let (x_n) be a d -asymptotic full sequence in X . Further, let u, v be two points in $\lim_n(x_n)$ with $u \neq v$. By the full property of (x_n) ,

$$(\exists h \geq 0): \{u, v\} \cap \{x_n; n \geq h\} = \emptyset.$$

Denote for simplicity $M = \{u, v\} \cup \{x_n; n \geq h\}$. From the 2-local Branciari property of $d(., .)$, there exists the minimal Branciari constant $\alpha := \nu_M(u, v) \geq 1$ of (u, v) on M . But then, combining with the α -triangular inequality for (u, v) , one gets

$$d(u, v) \leq d(u, x_{n+1}) + \dots + d(x_{n+\alpha}, v), \text{ for all } n \geq h.$$

Passing to limit as $n \rightarrow \infty$ yields $d(u, v) = 0$ (by the choice of u, v and the d -asymptotic property of (x_n)). But then, $u = v$; an impossible situation.

Part 2. Let (x_n) be a d -Cauchy sequence in X . Assume by contradiction that $\lim_n(x_n)$ has at least two distinct points:

$$\exists u, v \in X \text{ with } u \neq v, \text{ such that } x_n \xrightarrow{d} u, x_n \xrightarrow{d} v.$$

Step 2-1. Denote $A = \{n \in N; x_n = u\}, B = \{n \in N; x_n = v\}$. We claim that both A and B are finite. In fact, if A is effectively denumerable, then $A = \{i(n); n \geq 0\}$, where $(i(n); n \geq 0)$ is strictly ascending and $(x_{i(n)} = u, \forall n \geq 0)$. Since, on the other hand, $x_{i(n)} \rightarrow v$ as $n \rightarrow \infty$, we must have $d(u, v) = 0$, so that $u = v$, contradiction. An identical reasoning is applicable when B is effectively denumerable, hence the claim. As a consequence, there exists some index $p \in N$, such that

$$\{x_n; n \geq p\} \cap \{u, v\} = \emptyset \quad (x_n \neq u \text{ and } x_n \neq v, \text{ for all } n \geq p).$$

Step 2-2. Put $h(0) = p$, and

$$S[0] = \{n \geq h(0); x_n = x_{h(0)}\}; \text{ hence, } h(0) \in S[0].$$

We claim that the set $S[0]$ is finite. For, otherwise, it has the representation $S[0] = \{j(n); n \geq 0\}$, where $(j(n); n \geq 0)$ is strictly ascending and $(x_{j(n)} = x_{h(0)}, \forall n)$. Combining with the convergence hypothesis, one derives

$$x_{h(0)} = u, x_{h(0)} = v; \text{ wherefrom, } u = v, \text{ contradiction.}$$

Hence, $S[0]$ is indeed finite; wherefrom, $H(0) = \max S[0]$ exists; moreover, the rank $h(1) := H(0) + 1$ fulfills

$$h(0) < h(1), \{x_n; n \geq h(1)\} \cap \{x_{h(0)}\} = \emptyset, \text{ so that } x_{h(1)} \notin \{x_{h(0)}\}.$$

Further, denote

$$S[1] = \{n \geq h(1); x_n = x_{h(1)}\}; \text{ hence, } h(1) \in S[1].$$

By a very similar argument, $S[1]$ is finite too; wherefrom, $H(1) = \max S[1]$ exists; moreover, the rank $h(2) := H(1) + 1$ fulfills

$$h(0) < h(1) < h(2), \{x_n; n \geq h(2)\} \cap \{x_{h(0)}, x_{h(1)}\} = \emptyset; \\ \text{hence, in particular, } x_{h(2)} \notin \{x_{h(0)}, x_{h(1)}\}.$$

The procedure may continue indefinitely; it yields—inductively—a strictly ascending rank sequence $(h(n); n \geq 0)$, such that the subsequence $(y_n := x_{h(n)}; n \geq 0)$ of $(x_n; n \geq 0)$ fulfills

- (p-1) (y_n) is full, $u, v \in \lim_n (y_n)$, and $\{u, v\} \cap \{y_n; n \geq 0\} = \emptyset$,
- (p-2) (y_n) is d -Cauchy; hence, d -asymptotic $(\lim_n d(y_n, y_{n+1}) = 0)$.

But then, from the preceding stage, we get a contradiction, hence the conclusion.

Remember that if $e : X \times X \rightarrow R_+$ is a metric on X , the mapping $(x, y) \mapsto e(x, y)$ is continuous, in the sense:

$$x_n \xrightarrow{e} x, y_n \xrightarrow{e} y \text{ imply } e(x_n, y_n) \rightarrow e(x, y).$$

Unfortunately, for a local and/or 2-local Branciari metric $d(., .)$, this property is no longer valid. However, two partial versions of it are available.

The first of these refers to 2-local Branciari metric spaces.

Proposition 8 *Let (X, d) be a 2- local Branciari metric space. If the full d -telescopic d -Cauchy sequence $(x_n; n \geq 0)$ in X and the points $u, v \in X$ are such that*

$$\{x_n; n \geq 0\} \cap \{u, v\} = \emptyset \text{ (i.e., } x_n \neq u \text{ and } x_n \neq v, \forall n),$$

then, necessarily,

$$x_n \xrightarrow{d} u \text{ implies } d(x_n, v) \rightarrow d(u, v).$$

Proof The case $u = v$ is clear; so, without any loss, one may assume that $u \neq v$. By the d -telescopic Cauchy property of (x_n) ,

$$\sum_{n \geq 0} \rho_n < \infty, \text{ where } (\rho_n := d(x_n, x_{n+1}); n \geq 0);$$

$$\text{hence, } \lim_n \sigma_n = 0, \text{ where } (\sigma_n := \sum_{k \geq n} \rho_k; n \geq 0).$$

Denote $M = \{u, v\} \cup \{x_n; n \geq 0\}$; this is an effectively denumerable part of X . Then, let $\alpha := v_M(u, v) \geq 1$ stand for the minimal Branciari constant of (u, v) over M ; note that this constant is independent of the α -chains in $M_0 = \{x_n; n \geq 0\}$. Finally, put for simplicity

$$(\gamma_n := d(x_n, u); n \geq 0) \text{ (hence, } \lim_n \gamma_n = 0),$$

$$(\Gamma_n := \sup\{\gamma_k; k \geq n\}; n \geq 0) \text{ (hence, } (\Gamma_n) \text{ is descending and } \lim_n \Gamma_n = 0).$$

For the arbitrary fixed $n \geq 0$, we have (combining with the 2-local triangular property relative to d)

$$d(u, v) \leq \Lambda(u; x_{n+1}, \dots, x_{n+\alpha}, v) =$$

$$\gamma_{n+1} + \Lambda(x_{n+1}, \dots, x_{n+\alpha}) + d(x_{n+\alpha}, v) \leq \Gamma_n + \sigma_n + d(x_{n+\alpha}, v).$$

Further, let $\beta := v_M(x_{n+\alpha}, v) \geq 1$ stand for the minimal Branciari constant of $(x_{n+\alpha}, v)$ over M ; clearly, it depends on the (starting) index n . By the 2-local triangular property once again,

$$d(x_{n+\alpha}, v) \leq \Lambda(x_{n+\alpha}, \dots, x_{n+\alpha+\beta-1}, u, v) =$$

$$\Lambda(x_{n+\alpha}, \dots, x_{n+\alpha+\beta-1}) + \gamma_{n+\alpha+\beta-1} + d(u, v) \leq$$

$$\sigma_{n+\alpha} + \gamma_{n+\alpha+\beta-1} + d(u, v) \leq \sigma_n + \Gamma_n + d(u, v).$$

Combining these inequalities yields

$$d(u, v) - \sigma_n - \Gamma_n \leq d(x_{n+\alpha}, v) \leq d(u, v) + \sigma_n + \Gamma_n, \forall n,$$

and this finally yields

$$\lim_n d(x_{n+\alpha}, v) = d(u, v) \text{ (as } \lim_n \sigma_n = \lim_n \Gamma_n = 0),$$

which is nothing else than the desired conclusion.

Further, the second of these answers refers to local Branciari metric spaces. As it would be expected, the working hypothesis about our sequence can be weakened.

Proposition 9 *Let (X, d) be a local Branciari metric space. If the full d -asymptotic sequence $(x_n; n \geq 0)$ in X and the points $u, v \in X$ are such that*

$$\{x_n; n \geq 0\} \cap \{u, v\} = \emptyset \text{ (i.e., } x_n \neq u \text{ and } x_n \neq v, \forall n),$$

then, necessarily,

$$x_n \xrightarrow{d} u \text{ implies } d(x_n, v) \rightarrow d(u, v).$$

Proof As before, the case $u = v$ is clear; so, without any loss, one may assume that $u \neq v$. Denote $M = \{u, v\} \cup \{x_n; n \geq 0\}$, and let $\alpha := v_{\mathcal{B}}(M) \geq 1$ be the minimal Branciari constant for M (existing by the local Branciari property of d). Denote

$$A_n := (x_{n+1}, \dots, x_{n+\alpha}), B_n = (x_{n+\alpha}, \dots, x_{n+1}), n \geq 0.$$

For each $n \geq 0$, A_n and B_n are full α -chains in X . By the α -triangular inequality applied to the full $(\alpha + 2)$ -chains $(u; A_n; v)$ and $(B_n; u; v)$, respectively, we have

$$d(u, v) \leq \Lambda(u; A_n; v) = \Lambda(u; A_n) + d(x_{n+\alpha}, v),$$

$$d(x_{n+\alpha}, v) \leq \Lambda(B_n; u; v) = \Lambda(B_n; u) + d(u, v);$$

or, equivalently,

$$d(u, v) - \Lambda(u; A_n) \leq d(x_{n+\alpha}, v) \leq d(u, v) + \Lambda(B_n; u).$$

Moreover, from the d -asymptotic and d -convergence (toward u) hypotheses,

$$\Lambda(u; A_n) \rightarrow 0 \text{ and } \Lambda(B_n; u) \rightarrow 0, \text{ as } n \rightarrow \infty;$$

so, by simply combining these,

$$d(x_{n+\alpha}, v) \rightarrow d(u, v), \text{ as } n \rightarrow \infty;$$

wherefrom, the desired conclusion holds. The proof is complete.

We close this section with an auxiliary fact to be used later.

Proposition 10 *Let (X, d) be a 2-local Branciari metric space. Further, let the full sequence (x_n) in X and the points $z, w \in X$ be such that*

$$z \neq w \text{ and } \{z, w\} \cap \{x_n; n \geq 0\} = \emptyset.$$

Then, for each index $m \in \mathbb{N}$, there exists a strictly ascending rank sequence $(r(n); n \geq 0)$ in $\mathbb{N}(m, <)$, with

(rela) $d(x_m, w) \leq \Lambda(x_m, \dots, x_{r(0)}, \dots, x_{r(n)}, z, w)$, for each $n \geq 0$.

Proof Denote $M = \{z, w\} \cup \{x_n; n \geq 0\}$. There are two steps to be passed:

(I) First, let us show that (rela-0) holds; that is,

$$\text{(rela-0) } \forall m, \exists r(0) > m, \text{ with } d(x_m, w) \leq \Lambda(x_m, \dots, x_{r(0)}, z, w).$$

To do this, fix some $m \geq 0$, and let $\gamma := v_M(x_m, w) \geq 1$ stand for the minimal Branciari constant of (x_m, w) over M . Two alternatives occur:

(I-1) Suppose that $\gamma > 1$; hence, $\gamma \geq 2$. Then,

$$d(x_m, w) \leq \Lambda(x_m, \dots, x_{m+\gamma-1}, z, w).$$

Putting $r(0) = m + \gamma - 1$, we derive (rela-0), because

$$r(0) \geq m + 1 > m, \text{ hence the claim.}$$

(I-2) Suppose that $\gamma = 1$. For the moment,

$$d(x_m, w) \leq \Lambda(x_m, z, w) = d(x_m, z) + d(z, w).$$

Let $\delta := v_M(x_m, z) \geq 1$ stand for the minimal Branciari constant of (x_m, z) over M . We thus have

$$d(x_m, z) \leq \Lambda(x_m, \dots, x_{m+\delta}, z);$$

so, replacing in the preceding relation,

$$d(x_m, w) \leq \Lambda(x_m, \dots, x_{m+\delta}, z, w).$$

Putting $r(0) = m + \delta$, we get (rela-0), because

$$r(0) > m \text{ (in view of } \delta \geq 1 > 0), \text{ hence the claim.}$$

(II) Starting from (rela-0), let $\alpha(0) := \nu_M(x_{r(0)}, z) \geq 1$ stand for the minimal Branciari constant of $(x_{r(0)}, z)$ over M . We have (by definition)

$$d(x_{r(0)}, z) \leq \Lambda(x_{r(0)}, \dots, x_{r(0)+\alpha(0)}, z);$$

so, combining with (rela-0), one derives

$$d(x_m, w) \leq \Lambda(x_m, \dots, x_{r(0)}, \dots, x_{r(1)}, z, w),$$

where $r(1) = r(0) + \alpha(0) > r(0)$;

so, (rela-1) holds. Further, starting from (rela-1), let $\alpha(1) = \nu_M(x_{r(1)}, z) \geq 1$ stand for the minimal Branciari constant of $(x_{r(1)}, z)$ over M . We have (by definition)

$$d(x_{r(1)}, z) \leq \Lambda(x_{r(1)}, \dots, x_{r(1)+\alpha(1)}, z);$$

so, combining with (rela-1), one derives

$$d(x_m, w) \leq \Lambda(x_m, \dots, x_{r(0)}, \dots, x_{r(1)}, \dots, x_{r(2)}, z, w),$$

where $r(2) = r(1) + \alpha(1) > r(1)$,

which tells us that (rela-2) holds. This procedure may continue indefinitely and yields—by a finite induction technique—the desired conclusion.

In particular, the 2-local Branciari metric property for (X, d) is fulfilled under a local Branciari metric property of the same. We thus have

Proposition 11 *Let (X, d) be a local Branciari metric space. Further, let the full sequence (x_n) in X and the points $z, w \in X$ be such that*

$$z \neq w \text{ and } \{z, w\} \cap \{x_n; n \geq 0\} = \emptyset.$$

Then, for each index $m \in N$, there exists a strictly ascending rank sequence $(r(n); n \geq 0)$ in $N(m, <)$, with

$$(rela) \ d(x_m, w) \leq \Lambda(x_m, \dots, x_{r(0)}, \dots, x_{r(n)}, z, w), \text{ for each } n \geq 0.$$

In particular, the local Branciari metric property for (X, d) is fulfilled under a Branciari metric property of the same. This yields a corresponding version of statement above over such structures; we do not give details.

5 Main Result

Let X be a nonempty set, and $d : X \times X \rightarrow R_+$ be a mapping with

(symm) d is symmetric [$d(x, y) = d(y, x), \forall x, y \in X$],

(r-s) d is reflexive sufficient [$x = y \iff d(x, y) = 0$];

we then say that d is a *reflexive sufficient symmetric* (in short, *(r-s)-symmetric*), and (X, d) is a *(r-s)-symmetric space*.

Suppose that we fixed such a structure, and Further, let $\varphi : X \rightarrow R$ be a function. Define a relation ∇ over X as

$$(x, y \in X): x \nabla y \text{ iff } d(x, y) \leq \varphi(x) - \varphi(y).$$

Clearly, the following properties hold (by the choice of d):

- (na-r) ∇ is reflexive: $x \nabla x, \forall x \in X,$
- (na-as) ∇ is antisymmetric: $x \nabla y$ and $y \nabla x$ imply $x = y.$

In other words: ∇ has all properties of a (partial) order, excepting transitivity. And, if d is triangular (hence, a metric), ∇ is transitive as well; hence, a (partial) order. Let also $\tilde{\nabla}$ stand for the *irreflexive* part of ∇ , namely,

$$x \tilde{\nabla} y \text{ iff } x \nabla y \text{ and } x \neq y.$$

This convention comes from $\tilde{\nabla}$ being irreflexive. Moreover, if d is triangular (hence, a metric), $\tilde{\nabla}$ is transitive too; hence, a strict order on X .

Having this precise, define a maximality property over X under the standard way: call the point $z \in X, \nabla$ -maximal if

$$X(z, \nabla) = \{z\}; \text{ or, equivalently: } X(z, \tilde{\nabla}) = \emptyset;$$

the class of all such elements will be denoted as $\max(X, \nabla)$. For practical reasons, it would be useful to determine conditions under which

(a-Zorn) ∇ is an almost Zorn relation: $\max(X, \nabla)$ is nonempty.

For the metrical framework, the conditions in question involve completeness of our ambient space (X, d) and lower semicontinuity of the objective function φ . It is natural then to ask whether these are in effect over the (local and) 2-local Branciari setting too. As we will see, a positive answer to this is essentially available. Further aspects occasioned by our developments are also discussed.

Let (X, d) be a 2-local Branciari metric space. Remember that the sequence (x_n) is *d-telescopic Cauchy*, when

$$\sum_n d(x_n, x_{n+1})(= d(x_0, x_1) + d(x_1, x_2) + \dots) < \infty.$$

Then, let us say that X is *full d-telescopic complete*, when

(f-tele-com) each full *d-telescopic* Cauchy sequence is *d-convergent*.

Note that, by a previous result, the set $\lim_n(x_n)$ for any such sequence (x_n) is an asingleton. Further, let us say that the function $\varphi : X \rightarrow R$ is *full d-lsc*, provided

$$\lim \inf_n \varphi(x_n) \geq \varphi(z), \text{ for each full sequence } (x_n) \text{ in } X \text{ with } x_n \xrightarrow{d} z.$$

We may now state the main result of this exposition (called: Ekeland Variational Principle for 2-local Branciari metric spaces; in short, (EVP-2-loc-Bms)).

Theorem 3 *Let the 2-local Branciari metric space (X, d) and the (objective) function $\varphi : X \rightarrow R$ be such that*

- (51-i) X is full d -telescopic complete,
- (51-ii) φ is bounded from below and full d -lsc.

Then, in the reduced axiomatic system (ZF-AC+DC),

the associated relation ∇ is an almost Zorn one: $\max(X, \nabla) \neq \emptyset$.

Proof Suppose by contradiction that this is not true:

- (51-iii) $X(x, \tilde{\nabla})$ is nonempty, for each $x \in X$.

Define a sequence of relations $(\mathcal{R}_n; n \geq 0)$ over X according to

$$(\forall n): x\mathcal{R}_n y \text{ iff } x\tilde{\nabla} y \text{ and } \varphi(y) < \inf \varphi(X(x, \tilde{\nabla})) + 2^{-n}.$$

We claim that

$$(\forall n): \mathcal{R}_n(x) \neq \emptyset, \forall x \in X \text{ (i.e., } \mathcal{R}_n \text{ is proper).}$$

In fact, let $x \in X$ be arbitrary fixed, and put $M = X(x, \tilde{\nabla})$. By the infimum definition, we have

$$\forall \varepsilon > 0, \exists z \in M \text{ (hence, } x\tilde{\nabla} z), \text{ such that } \varphi(z) < \inf \varphi(M) + \varepsilon;$$

and this, applied to the sequence $(\varepsilon_n := 2^{-n}; n \geq 0)$, proves our claim. As a consequence, the Diagonal Dependent Choice principle (DDC) is applicable here; so, given the starting point $u_0 \in X$, there exists a sequence $(u_n; n \geq 0)$ in X with

$$(\forall n): u_n\mathcal{R}_n u_{n+1}; \text{ that is, } u_n\tilde{\nabla} u_{n+1} \text{ and } \varphi(u_{n+1}) < \inf \varphi(X(u_n, \tilde{\nabla})) + 2^{-n}.$$

By the former of these relations, we have

$$(\forall n): u_n \neq u_{n+1} \text{ and } d(u_n, u_{n+1}) \leq \varphi(u_n) - \varphi(u_{n+1}).$$

Note that, as a first consequence of this,

$$\varphi(u_n) > \varphi(u_{n+1}), \forall n; \text{ whence, } (\varphi(x_n); n \geq 0) \text{ is strictly descending, so that } (u_n; n \geq 0) \text{ is full.}$$

On the other hand, as a second consequence of this,

$$(\varphi(x_n); n \geq 0) \text{ is (strictly descending) bounded, so that } (u_n) \text{ is } d\text{-telescopic Cauchy in } X.$$

Combining with the full d -telescopic completeness property of X , we have

$$u_n \xrightarrow{d} z \text{ as } n \rightarrow \infty, \text{ for some } z \in X;$$

in addition, since (u_n) is d -asymptotic, the point $z \in X$ is uniquely determined. On the other hand, as φ is full d -lsc, we get (via $(\varphi(u_n)) =$ strictly descending)

$$\varphi(z) \leq \liminf_n \varphi(u_n) = \lim_n \varphi(u_n); \text{ whence, } \varphi(z) < \varphi(u_n), \forall n.$$

Note that, until now, the 2-local Branciari metric space condition was not used.

Finally, take some $w \in X(z, \tilde{\nabla})$ (nonempty, by hypothesis); we have

$$0 < d(z, w) \leq \varphi(z) - \varphi(w); \text{ wherefrom (by the above)}$$

$$(\forall n): (\varphi(u_n) >) \varphi(z) > \varphi(w), \text{ so that } u_n \neq z \text{ and } u_n \neq w.$$

Let $m \geq 0$ be arbitrary fixed. By the 2-local Branciari metric condition upon d (and an auxiliary fact), there exists a strictly ascending rank sequence $(r(n))$ in $N(m, <)$, such that

$$(\text{rela}) d(u_m, w) \leq \Lambda(u_m, \dots, u_{r(n)}, z, w) =$$

$$\Lambda(u_m \dots, u_{r(n)}) + d(u_{r(n)}, z) + d(z, w), \text{ for each } n \geq 0.$$

But then, by the very definition of our relation (∇) ,

$$d(u_m, w) \leq \varphi(u_m) - \varphi(u_{r(n)}) + d(u_{r(n)}, z) + \varphi(z) - \varphi(w)$$

$$\leq \varphi(u_m) - \varphi(z) + d(u_{r(n)}, z) + \varphi(z) - \varphi(w) =$$

$$\varphi(u_m) - \varphi(w) + d(u_{r(n)}, z), \text{ for each } n \geq 0.$$

Passing to limit as $n \rightarrow \infty$, one derives

$$(\forall m): d(u_m, w) \leq \varphi(u_m) - \varphi(w), \text{ so that } w \in X(u_m, \tilde{\nabla}) \text{ (via } u_m \neq w).$$

But then, according to the definition of (u_n) , one gets

$$\varphi(z) < \varphi(u_{m+1}) < \varphi(w) + 2^{-m}, \text{ for all } m; \text{ whence}$$

$$\varphi(z) \leq \lim_m \varphi(u_m) \leq \varphi(w) \text{ (passing to limit as } m \rightarrow \infty).$$

This, however, is impossible, in view of $w \in X(z, \tilde{\nabla}) \implies \varphi(w) < \varphi(z)$. Hence, our working hypothesis (51-iii) cannot be accepted, and conclusion follows.

As a direct consequence of this, the following fixed point statement (referred to as Caristi–Kirk fixed point theorem on 2-local Branciari metric spaces; in short, (CK-2-loc-Bms)) is available.

Theorem 4 *Let the 2-local Branciari metric space (X, d) , the function $\varphi : X \rightarrow R$, and the self-map $T : X \rightarrow X$ be such that*

- (52-i) X is full d -telescopic complete,
- (52-ii) φ is bounded from below and full d -lsc,
- (52-iii) T is φ -progressive: $d(x, Tx) \leq \varphi(x) - \varphi(Tx), \forall x \in X$.

Then, in the reduced axiomatic system (ZF-AC+DC),

- (52-a) *the associated relation ∇ is an almost Zorn one: $\max(X, \nabla) \neq \emptyset$,*
- (52-b) *we necessarily have $\max(X, \nabla) \subseteq \text{Fix}(T)$; whence, T has at least one fixed point in X .*

Proof Let $z \in \max(X, \nabla)$ be arbitrary fixed. As $z \nabla Tz$, we must have $z = Tz$, and the proof is complete.

In particular, the 2-local Branciari metric space property is deductible from the local Branciari metric one. Combining with Ekeland Variational Principle for 2-local Branciari metric spaces (EVP-2-loc-Bms), one gets the following practical

statement (called: Ekeland Variational Principle for local Branciari metric spaces; in short, (EVP-loc-Bms)).

Theorem 5 *Let the local Branciari metric space (X, d) and the function $\varphi : X \rightarrow R$ be such that*

- (53-i) *X is full d -telescopic complete,*
- (53-ii) *φ is bounded from below and full d -lsc.*

Then, in the reduced axiomatic system (ZF-AC+DC),

the associated relation ∇ is an almost Zorn one: $\max(X, \nabla) \neq \emptyset$.

At the same time, combining with Caristi–Kirk fixed point theorem on 2-local Branciari metric spaces (CK-2-loc-Bms), one gets the practical statement below (referred to as Caristi–Kirk fixed point theorem on local Branciari metric spaces; in short, (CK-loc-Bms)).

Theorem 6 *Let the local Branciari metric space (X, d) , the function $\varphi : X \rightarrow R$, and the self-map $T : X \rightarrow X$ be such that*

- (54-i) *X is full d -telescopic complete,*
- (54-ii) *φ is bounded from below and full d -lsc,*
- (54-iii) *T is φ -progressive: $d(x, Tx) \leq \varphi(x) - \varphi(Tx), \forall x \in X$.*

Then, in the reduced axiomatic system (ZF-AC+DC),

- (54-a) *the associated relation ∇ is an almost Zorn one: $\max(X, \nabla) \neq \emptyset$,*
- (54-b) *we necessarily have $\max(X, \nabla) \subseteq \text{Fix}(T)$; whence, T has at least one fixed point in X .*

In particular, the local Branciari metric property for d is fulfilled when d is a Branciari metric. The corresponding version of Caristi–Kirk fixed point theorem on local Branciari metric spaces (CK-loc-Bms) is just the statement in Alamri et al. [1], proved via similar methods.

Finally, note that various circumstances under which the full d -telescopic completeness condition is to be fulfilled were indicated in the quoted paper. On the other hand, we stress that functional versions of these results in the way described by Park and Bae [31] or Turinici [39] are not yet possible in this setting. Further aspects will be delineated elsewhere.

6 Equivalence Statements

Let (X, d) be a (standard) metric space, and $\varphi : X \rightarrow R$ be a function. The relation (\leq) over X introduced as

$$x \leq y \text{ iff } d(x, y) \leq \varphi(x) - \varphi(y)$$

is reflexive, transitive, and antisymmetric; hence, a (partial) order on X . Let also $(<)$ stand for the *irreflexive* part of (\leq) ; namely,

$$x < y \text{ iff } x \nabla y \text{ and } x \neq y.$$

Clearly, $(<)$ is (irreflexive and) transitive; hence, a strict order on X .

Having this precise, define a maximality property over X as: call the point $z \in X$, (\leq) -*maximal*, if

$$X(z, \leq) = \{z\}; \text{ or, equivalently: } X(z, <) = \emptyset;$$

the class of all such elements will be denoted as $\max(X, \leq)$. As before, it would be useful to determine conditions under which

(a-Zorn) (\leq) is an almost Zorn (partial) order: $\max(X, \leq)$ is nonempty.

In this direction, as a direct consequence of our previous developments, the following statement (referred to as *generic Ekeland Variational Principle*; in short, (EVP-gen)) is available.

Theorem 7 *Suppose that (in the precise context)*

- (61-i) X is d -complete: each d -Cauchy sequence is d -convergent,
- (61-ii) φ is bounded from below and d -lsc.

Then, necessarily, (\leq) is an almost Zorn relation, in the reduced axiomatic system (ZF-AC+DC).

Proof Clearly, (X, d) is a local Branciari metric space. In addition, by the posed hypothesis, one trivially has

$$X \text{ is full } d\text{-telescopic complete and } \varphi \text{ is bounded below, full } d\text{-lsc.}$$

Summing up, Ekeland Variational Principle for local Branciari metric spaces (EVP-loc-Bms) is applicable here, and, from this, we are done.

It remains now to establish the relationships between this generic principle and (EVP). An appropriate answer to this is contained in

Proposition 12 *Under these conventions, we have*

$$(EVP\text{-gen}) \iff (EVP), \text{ in (ZF-AC).}$$

Proof The right to left inclusion is clear; so, it remains to establish the left to right inclusion. Let the metric space (X, d) be such that X is d -complete, and let the (extended valued) function $\varphi : X \rightarrow R \cup \{\infty\}$ be *regular*; i.e.,

- (r-1) φ is inf-proper ($\text{Dom}(\varphi) \neq \emptyset$ and $\inf[\varphi(M)] > -\infty$),
- (r-2) φ is d -lsc ($\liminf_n \varphi(x_n) \geq \varphi(x)$, whenever $x_n \xrightarrow{d} x$).

Denote by (\leq) the Brøndsted quasi-order

$$(Br\text{-ord}) \quad x \leq y \text{ iff } d(x, y) + \varphi(y) \leq \varphi(x);$$

clearly, (\leq) is a (partial) order on $\text{Dom}(\varphi)$. Let $u \in \text{Dom}(\varphi)$ be fixed in the sequel, and put $X_u = X(u, \leq)$; as $X_u \subseteq \text{Dom}(\varphi)$, one has that (\leq) is a (partial) order on X_u . Let again φ stand for the restriction of φ over X_u . We claim that (EVP-gen) is applicable over X_u and (the restriction of) φ .

- (I) By the imposed conditions, X_u is a closed part of X . Let (y_n) be a d -Cauchy sequence in X_u ; hence, in particular,

$$d(u, y_n) \leq \varphi(u) - \varphi(y_n), \text{ for all } n.$$

By the completeness hypothesis,

$$y_n \xrightarrow{d} z \text{ as } n \rightarrow \infty, \text{ for some } z \in X.$$

Passing to limit in the equivalent relation

$$\varphi(y_n) \leq \varphi(u) - d(u, y_n), \text{ for all } n,$$

gives (by the metrical properties of d)

$$\begin{aligned} \varphi(z) &\leq \liminf_n \varphi(y_n) \leq \varphi(u) - \lim_n d(u, y_n) = \varphi(u) - d(u, z); \\ &\text{that is, } z \in X_u. \end{aligned}$$

This, by the arbitrariness of our sequence, proves that X_u is d -complete.

- (II) Clearly, the restriction $\varphi : X_u \rightarrow R$ is a bounded from below function endowed with the d -lsc property on X_u .

Putting these together, it follows that (EVP-gen) is indeed applicable over X_u and (the restriction of) φ . This tells us that there exists some $v \in X_u$ with

$$d(v, w) > \varphi(v) - \varphi(w), \text{ for all } w \in X_u, w \neq v.$$

Now, the very relation $v \in X_u$ yields the first conclusion in (EVP). For the second one, let $x \in X \setminus \{v\}$ be arbitrary fixed. If, by absurd,

$$d(v, x) \leq \varphi(v) - \varphi(x) \text{ (i.e., } v \leq x),$$

it results (via $u \leq v$) that $u \leq x$; or, equivalently, $x \in X_u$. This, however, contradicts the previous relation involving v , and, then, we are done.

By the developments above, we have [in the strongly reduced Zermelo–Fraenkel system (ZF-AC)] the inclusions

- (i-1) (DC) \implies (EVP-2-loc-Bms) \implies (EVP-loc-Bms),
- (i-2) (EVP-loc-Bms) \implies (EVP-gen) \implies (EVP).

So, it is natural asking whether these inclusion chains may be reversed. At a first glance, a negative answer is expectable, because (DC) is “too general” with respect to (EVP). However, the situation is exactly opposite; i.e., (EVP) includes (DC); and then we closed the circle between all such principles. An early result of this type was provided in 1987 by Brunner [9]; for a different answer to the same, we refer to the 1999 paper by Dodu and Morillon [15]. It is our aim in the following to show that a further extension of this last result is possible, in the sense: (DC) is deductible

from a certain Lipschitz-bounded countable version of (EVP). Some preliminaries are needed.

Let (X, \leq) be a partially ordered structure. Remember that $z \in X$ is (\leq) -maximal, if $z \leq w \in X$ implies $z = w$; the class of all these will be denoted as $\max(X, \leq)$. In this case, we say that (\leq) is a *Zorn order* when

$\max(X, \leq)$ is (nonempty and) cofinal in X
 (for each $u \in X$, there exists $v \in \max(X, \leq)$ with $u \leq v$).

In particular, when $d(., .)$ is a (standard) metric on X and $\varphi : X \rightarrow R_+$ is some function, a good example of partial order on X is that introduced by the convention

$$x \leq_{(d,\varphi)} y \text{ iff } d(x, y) \leq \varphi(x) - \varphi(y);$$

referred to as the *Brøndsted order* [8] attached to the couple (d, φ) . Further, let us say that φ is *d-Lipschitz*, provided

$$|\varphi(x) - \varphi(y)| \leq Ld(x, y), \forall x, y \in X, \text{ for some } L > 0;$$

note that any such function is uniformly continuous on X .

The following stronger variant of (EVP) enters into this discussion.

Theorem 8 *Let the metric space (X, d) and the function $\varphi : X \rightarrow R_+$ satisfy*

- (62-i) X is d -bounded and d -complete,
- (62-ii) φ is d -Lipschitz (hence, bounded),
- (62-iii) $\varphi(X)$ is (at most) countable.

Then, $(\leq_{(d,\varphi)})$ is a Zorn order.

We call this, the *Lipschitz-bounded countable* version of (EVP) (in short, (EVP-Lbc)). By the above developments, we thus have

$$(DC) \implies (EVP) \implies (EVP-Lbc).$$

The remarkable fact to be added is that this last principle yields (DC); so—as precise—it completes the circle between all these.

Proposition 13 *We have, in the strongly reduced system (ZF-AC),*

$$(EVP-Lbc) \implies (DC); \text{ or: } (DC) \text{ is deductible in } (ZF-AC)+(EVP-Lbc).$$

As a consequence of this,

- (62-1) *the variational principles (EVP-2-loc-Bms), (EVP-loc-Bms), and (EVP-gen) are equivalent to both (DC) and (EVP); hence, necessarily, equivalent to each other;*
- (62-2) *any maximal/variational principle (VP) with $(DC) \implies (VP) \implies (EVP)$ is equivalent to both (DC) and (EVP).*

In particular, when the boundedness and Lipschitz properties are ignored, this result is just the one in Dodu and Morillon [15]. Further aspects may be found in the paper by Turinici [42].

Summing up, all variational principles in this exposition (derived from (DC))—as well as many other ones, described in Hyers et al. [20, Ch 5]—are nothing but logical equivalents of (EVP). So, it is natural to ask whether the remaining (sequential) ones—including the Smooth Variational Principle in Deville and Ghoussoub [14]—are endowed as well with such a property. The answer to this is affirmative; further aspects will be delineated elsewhere.

References

1. B. Alamri, T. Suzuki, L.A. Khan, Caristi's fixed point theorem and Subrahmanyam's fixed point theorem in ν -generalized metric spaces. *J. Funct. Spaces* 2015, 709391 (2015)
2. M. Altman, A generalization of the Brezis-Browder principle on ordered sets. *Nonlinear Anal.* **6**, 157–165 (1982)
3. T.Q. Bao, P.Q. Khanh, Are several recent generalizations of Ekeland's variational principle more general than the original principle? *Acta Math. Vietnamica* **28**, 345–350 (2003)
4. P. Bernays, A system of axiomatic set theory: Part III. Infinity and enumerability analysis. *J. Symb. Log.* **7**, 65–89 (1942)
5. N. Bourbaki, Sur le théorème de Zorn. *Archiv Math.* **2**, 434–437 (1949/1950)
6. A. Branciari, A fixed point theorem of Banach-Caccioppoli type on a class of generalized metric spaces. *Publ. Math. Debrecen* **57**, 31–37 (2000)
7. H. Brezis, F.E. Browder, A general principle on ordered sets in nonlinear functional analysis. *Adv. Math.* **21**, 355–364 (1976)
8. A. Brøndsted, Fixed points and partial orders. *Proc. Am. Math. Soc.* **60**, 365–366 (1976)
9. N. Brunner, Topologische Maximalprinzipien. *Zeitschr. Math. Logik Grundl. Math.* **33**, 135–139 (1987)
10. J. Caristi, W.A. Kirk, Geometric fixed point theory and inwardness conditions, in *The Geometry of Metric and Linear Spaces*. Lecture Notes Mathematics, vol. 490 (Springer, Berlin, 1975), pp. 74–83
11. G.Y. Chen, X.X. Huang, S.H. Hou, General Ekeland's variational principle for set-valued mappings. *J. Optim. Theory Appl.* **106**, 151–164 (2000)
12. G.Y. Chen, X.X. Huang, S.H. Hou, General Ekeland's variational principle for set-valued mappings [Errata Corrige]. *J. Optim. Theory Appl.* **117**, 217–218 (2003)
13. P.J. Cohen, *Set Theory and the Continuum Hypothesis* (Benjamin, New York, 1966)
14. R. Deville, N. Ghoussoub, Perturbed minimization principles and applications, in *Handbook of the Geometry of Banach Spaces*, ed. by W.B. Johnson, J. Lindenstrauss, vol. I (Elsevier Science B.V., Amsterdam, 2001), pp. 399–435
15. J. Dodu, M. Morillon, The Hahn-Banach property and the axiom of choice. *Math. Logic Quart.* **45**, 299–314 (1999)
16. I. Ekeland, On the variational principle. *J. Math. Anal. Appl.* **47**, 324–353 (1974)
17. I. Ekeland, Nonconvex minimization problems. *Bull. Am. Math. Soc.* **1**, 443–474 (1979)
18. A. Goepfert, H. Riahi, C. Tammer, C. Zălinescu, *Variational Methods in Partially Ordered Spaces*. Canadian Mathematical Society Books Mathematics, vol. 17 (Springer, New York, 2003)
19. P.R. Halmos, *Naive Set Theory* (Van Nostrand Reinhold, New York, 1960)
20. D.H. Hyers, G. Isac, T.M. Rassias, *Topics in Nonlinear Analysis and Applications* (World Scientific, Singapore, 1997)
21. M. Jleli, B. Samet, The Kannan's fixed point theorem in a cone rectangular metric space. *J. Nonlinear Sci. Appl.* **2**, 161–167 (2009)
22. O. Kada, T. Suzuki, W. Takahashi, Nonconvex minimization theorems and fixed point theorems in complete metric spaces. *Math. Japon.* **44**, 381–391 (1996)

23. B.G. Kang, S. Park, On generalized ordering principles in nonlinear analysis. *Nonlinear Anal.* **14**, 159–165 (1990)
24. S. Kasahara, On some generalizations of the Banach contraction theorem. *Publ. Res. Inst. Math. Sci. Kyoto Univ.* **12**, 427–437 (1976)
25. P.Q. Khanh, On Caristi-Kirk's theorem and Ekeland's variational principle for Pareto extrema. *Bull. Polish Acad. Sci.* **37**, 33–39 (1989)
26. W.A. Kirk, N. Shahzad, Generalized metrics and Caristi's theorem. *Fixed Point Theory Appl.* **2013**, 129 (2013)
27. W.A. Kirk, N. Shahzad, Correction: generalized metrics and Caristi's theorem. *Fixed Point Theory Appl.* **2014**, 177 (2014)
28. G.H. Moore, *Zermelo's Axiom of Choice: Its Origin, Development and Influence* (Springer, New York, 1982)
29. Y. Moskhovakis, *Notes on Set Theory* (Springer, New York, 2006)
30. A.B. Nemeth, A nonconvex vector minimization problem. *Nonlinear Anal.* **10**, 669–678 (1986)
31. S. Park, J.S. Bae, On the Ray-Walker extension of the Caristi-Kirk fixed point theorem. *Nonlinear Anal.* **9**, 1135–1136 (1985)
32. B. Samet, Discussion on "A fixed point theorem of Banach-Caccioppoli type on a class of generalized metric spaces" by A. Branciari. *Publ. Math. Debrecen* **76**, 493–494 (2010)
33. E. Schechter, *Handbook of Analysis and its Foundation* (Academic Press, New York, 1997)
34. T. Suzuki, Generalized metric spaces do not have the compatible topology. *Abstr. Appl. Anal.* **2014**, 458098 (2014)
35. A. Tarski, Axiomatic and algebraic aspects of two theorems on sums of cardinals. *Fund. Math.* **35**, 79–104 (1948)
36. D. Tataru, Viscosity solutions of Hamilton-Jacobi equations with unbounded nonlinear terms. *J. Math. Anal. Appl.* **163**, 345–392 (1992)
37. M. Turinici, A generalization of Altman's ordering principle. *Proc. Am. Math. Soc.* **90**, 128–132 (1984)
38. M. Turinici, Function pseudometric VP and applications. *Bul. Inst. Polit. Iași* **53**(57), 393–411 (2007)
39. M. Turinici, Variational statements on KST-metric structures. *Ann. Şt. Univ. Ovidius Constanța* **17**, 231–246 (2009)
40. M. Turinici, Function variational principles and coercivity over normed spaces. *Optimization* **59**, 199–222 (2010)
41. M. Turinici, Functional contractions in local Branciari metric spaces. *Romai J.* **8**(2), 189–199 (2012)
42. M. Turinici, Sequential maximality principles, in *Mathematics Without Boundaries* ed. by T. M. Rassias, P.M. Pardalos (Springer, New York, 2014), pp. 515–548
43. E.S. Wolk, On the principle of dependent choices and some forms of Zorn's lemma. *Can. Math. Bull.* **26**, 365–367 (1983)
44. C.K. Zhong, A generalization of Ekeland's variational principle and application to the study of the relation between the weak P.S. condition and coercivity. *Nonlinear Anal.* **29**, 1421–1431 (1997)
45. M. Zorn, A remark on method in transfinite algebra. *Bull. Am. Math. Soc.* **41**, 667–670 (1935)