Springer Proceedings in Mathematics & Statistics

Gerard Olivar Tost
Olga Vasilieva *Editors*

Analysis, Modelling, Optimization, and Numerical Techniques

ICAMI, San Andres Island, Colombia, November 2013

Springer

Springer Proceedings in Mathematics & Statistics

Volume 121

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

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

Gerard Olivar Tost • Olga Vasilieva
Editors

Analysis, Modelling, Optimization, and Numerical Techniques

ICAMI, San Andres Island, Colombia, November 2013

Springer

Editors
Gerard Olivar Tost
Universidad Nacional de Colombia
Department of Electric, Electronic and Computer Science
Manizales
Colombia

Olga Vasilieva
Department of Mathematics
Universidad del Valle
Cali
Colombia

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-319-12582-4 ISBN 978-3-319-12583-1 (eBook)
DOI 10.1007/978-3-319-12583-1

Library of Congress Control Number: 2015933271

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

Contributors

Carlos D. Acosta Departamento de Matemáticas y Estadística, sede Manizales, Universidad Nacional de Colombia, Manizales, Colombia

Primitivo B. Acosta-Humánez Department of Mathematics, Universidad del Atlántico and Intelectual. Co, Barranquilla, Colombia

Raha Akhavan-Tabatabaei Universidad de los Andes, Bogotá, Colombia

Pedro A. Amador Departamento de Matemáticas y Estadística, sede Manizales, Universidad Nacional de Colombia, Manizales, Colombia

Boris D. Annin Lavrentyev Institute of Hydrodynamics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk State University, Novosibirsk, Russia

Anatoly Antipin Computing Center, Russian Academy of Sciences, Moscow, Russia

Jaime Arango Departamento de Matemáticas, Universidad del Valle, Cali, Colombia

Rodrigo Assar ICBM Escuela de Medicina, Universidad de Chile, Santiago, Chile

Christian Autenried Department of Mathematics, University of Bergen, Bergen, Norway

Leonardo Castañeda Universidad Nacional de Colombia, Observatorio Astronómico Nacional, Colombia

Benito M. Chen-Charpentier University of Texas at Arlington, Arlington, TX, USA

Juan Manuel Cordovez Departamento de Ingeniería Biomédica, Universidad de los Andes, Bogotá D.C., Colombia

Richard A. De la cruz Guerrero School of Mathematics and Statistics, Universidad Pedagógica y Tecnológica de Colombia-UPTC, Tunja, Colombia

Jairo Delgado Posgrado en Ciencias Matemáticas, Universidad del Valle, Cali, Colombia

Ibrahim Diakite University of Texas at Arlington, Arlington, TX, USA

Isaac Dyner Universidad Nacional de Colombia, Bogotá, Colombia
Universidad Jorge Tadeo Lozano, Bogotá, Colombia

Joaquín Fernández Departamento de Control, FCEIA, Universidad Nacional de Rosario, Rosario, Argentina

Dominique Fortin INRIA, Domaine de Voluceau, Le Chesnay Cedex, France

Carlos Jaime Franco Department of Computer Science and Decision, School of Mines, Universidad Nacional de Colombia, Medellín, Colombia

Mauricio Godoy Molina Department of Mathematics, University of Bergen, Bergen, Norway

Departamento de Matemática y Estadística, Universidad de la Frontera, Temuco, Casilla, Chile

Cindy González Université de Valenciennes et du Hainaut-Cambrésis, Valenciennes cedex 9, France

Eduardo González-Olivares Grupo de Ecología Matemática, Pontificia Universidad Católica de Valparaiso, Valparaiso, Chile

Francisco Guerra-Vásquez Escuela de Ciencias, Universidad de las Américas, Puebla, Puebla, México

Juan C. Juajibioy Department of Mathematics, Universidad Nacional de Colombia - UN, Bogotá, Colombia

Ernesto Kofman Departamento de Control, FCEIA, Universidad Nacional de Rosario, Rosario, Argentina

Victor A. Kovtunenko Lavrentyev Institute of Hydrodynamics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

Institute for Mathematics and Scientific Computing, Karl-Franzens University of Graz, Graz, Austria

Vladik Kreinovich University of Texas at El Paso, El Paso, TX, USA

Department of Geological Sciences, University of Texas at El Paso (UTEP), El Paso, TX, USA

Alejandro Maass Departamento de Ingeniería Matemática, Universidad de Chile, Santiago, Chile

Gisela L. Mazzieri Instituto de Matemática Aplicada del Litoral, IMAL, CCT CONICET Santa Fe, Santa Fe, Argentina

Departamento de Matemática, Facultad de Bioquímica y Ciencias Biológicas, Universidad Nacional del Litoral, Santa Fe, Argentina

Carlos E. Mejía Escuela de Matemáticas, sede Medellín, Universidad Nacional de Colombia, Medellín, Colombia

Mónica Jhoana Mesa Mazo Universidad del Quindío, Quindío, Colombia

Daniel Molano Universidad Sergio Arboleda, Universidad Nacional de Colombia, Bogotá, Colombia

Universidad Nacional de Colombia, Universidad Sergio Arboleda, Observatorio Astronómico Nacional, Colombia

Martín A. Montecino Centro de Investigaciones Biomédicas, Universidad Andrés Bello, Santiago, Chile

Akihiro Morinaga Mechanical Engineering Department, Kyushu University, Fukuoka, Japan

Gerard Olivar Tost Department of Electric, Electronic and Computer Science, Universidad Nacional de Colombia, Manizales, Colombia

Marco Paluszny Escuela de Matemáticas, Universidad Nacional de Colombia, Medellín, Colombia

Enrique Ponce Depto. Matemática Aplicada II, Camino Descubrimientos, Sevilla, Spain

Jan-Joachim Rückmann Department of Informatics, University of Bergen, Bergen, Norway

Leonardo D. Restrepo-Alape Grupo Gedes, Universidad del Quindío, Quindío, Colombia

Javier Ros Depto. Matemática Aplicada II, Camino Descubrimientos, Sevilla, Spain

Vladimir M. Sadovskii Institute of Computational Modeling, Siberian Branch of the Russian Academy of Sciences, Krasnoyarsk, Russia

Camilo Sanabria Departamento de Matemáticas, Universidad de los Andes, Bogotá D.C., Colombia

Jorge Soliz Universidad Sergio Arboleda, Bogotá, Colombia

Ruben D. Spies Instituto de Matemática Aplicada del Litoral, IMAL, CCT CONICET Santa Fe, Santa Fe, Argentina

Departamento de Matemática, Facultad de Ingeniería Química, Universidad Nacional del Litoral, Santa Fe, Argentina

Erwin Suazo Department of Mathematical Sciences, University of Puerto Rico, Puerto Rico, USA

School of Mathematics and Statistics, Arizona State University, AZ, USA

Mikhail Svinin Mechanical Engineering Department, Kyushu University, Fukuoka, Japan

Karina G. Temperini Instituto de Matemática Aplicada del Litoral, IMAL, CCT CONICET, Santa Fe, Santa Fe, Argentina

Departamento de Matemática, Facultad de Humanidades y Ciencias, Universidad Nacional del Litoral, Santa Fe, Argentina

Lennox Thompson Department of Geological Sciences, University of Texas at El Paso (UTEP), El Paso, TX, USA

Paulo C. Tintinago-Ruíz Universidad del Quindío, Quindío, Colombia

Hernán Darío Toro-Zapata Universidad del Quindío, Quindío, Colombia

Carlos Andrés Trujillo-Salazar Universidad del Quindío, Quindío, Colombia

Ider Tseveendorj Laboratoire PRiSM, UMR, Université de Versailles, Versailles Cedex, France

Johnny Valencia Department of Computer Science and Decision, School of Mines, Universidad Nacional de Colombia, Medellín, Colombia

Sergio A. Vargas Universidad de los Andes, Bogotá, Colombia

Olga Vasilieva Department of Mathematics, Universidad del Valle, Cali, Colombia

Elísabet Vela Depto. Matemática Aplicada II, Camino Descubrimientos, Sevilla, Spain

Aaron A. Velasco Department of Geological Sciences, University of Texas at El Paso (UTEP), El Paso, TX, USA

Motoji Yamamoto Mechanical Engineering Department, Kyushu University, Fukuoka, Japan

Shengfan Zhang University of Arkansas, Fayetteville, AR, USA

Augmented Lagrangian Method for Optimal Control Problems

Anatoly Antipin and Olga Vasilieva

Abstract This chapter describes a method for solving optimal control problems with boundary conditions at the right-hand end point of system trajectories. Such boundary conditions are expressed by means of finite-dimensional problem of convex programming. The trajectories are generated by a system of linear differential equations with control, and the latter is treated as an ordinary linear constraint. The proposed approach is based on solving the dual maximization problem engendered by an augmented Lagrangian of convex programming problem formulated in the infinite-dimensional functional space. There is also an analog in finite-dimensional convex programming, known as "augmented Lagrangian method." The convergence of the proposed method is proved in the infinite-dimensional functional space. This convergence has the additional property of monotonicity of the norm with respect to controls, trajectories, and adjoint functions.

Keywords Optimal control, Augmented Lagrangian method, Strong and weak convergence

1 Introduction

In the optimal control theory, there are two general trends for the development of solution methods without a priori discretization. First group of methods includes various techniques directly based on the Pontryagin maximum principle where one seeks to obtain an approximation of the control function that maximizes the Hamiltonian at (almost) each instant along the time interval. This group can be figuratively referred to as Hamiltonian-type methods.

A. Antipin (✉)
Computing Center, Russian Academy of Sciences, 40 Vavilov Str., 119991 Moscow, Russia
e-mail: antipin@ccas.ru

O. Vasilieva
Department of Mathematics, Universidad del Valle, Calle 13 No. 100-00, Cali, Colombia
e-mail: olga.vasilieva@correounivalle.edu.co

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_1

Another group of methods exploits the notion of Lagrangian (or augmented Lagrangian) and treats the dynamic system as intertemporal constraint. Therefore, this group can be regarded as Lagrangian-type methods. Under linear dynamics and with convex objective, Lagrangians are susceptible to duality and engender a minimization problem for primal variables and a dual problem of maximization. In other words, they always possess saddlepoints whose primal components provide the optimal solution to underlying problem of optimal control. Therefore, in order to solve an original optimal control problem, one may eventually seek for a saddlepoint of the associated Lagrangian.

Saddlepoint methods were originally developed in order to deal with convex programming problems in finite-dimensional spaces [11–13, 22–25], and were also called "methods of multipliers" or "methods of modified/augmented Lagrangian." In essence, they are variants of (sub)gradient ascent applied to a dual function engendered by the Lagrangian.

Subsequently, extra-gradient methods [1, 19] and extra-proximal methods [2, 5] were designed. Later, these methods have been adapted to the problems of equilibrium programming, dealing with computation of fixed points of extremal mappings [4, 6].

Extremal mappings help to describe many real-life situations of decision making involving an external human factor. The preference of decision making is modeled by an objective function that defines the best option from a variety of available alternatives.

Other situations can be described by mathematical models which characterize different kinds of balances or equilibria (biological, economical, ecological, energetic, financial, etc.). These models usually involve variational inequalities and equations.

These two classes of mathematical models have common features; however, they may differ due to some specificities or particularities of the objective function that is used in the preference system of decision making. Therefore, solution methods applicable to these models may possess dissimilar structures and may have different convergence properties.

Generally speaking, the efficiency of solution techniques is characterized by the solution nature that is expressed by the phase portrait in the vicinity of underlying solution. There are two main types of possible phase portraits. First type can be referred to as *(sub)gradient* (or *potential*) since it is usually associated with potential gradient vector field around minimum points of convex functions. Another type can be regarded as *saddle* (or *non-potential*) since it is usually associated with nonpotential vector field around saddlepoints of convex–concave functions.

In both cases, there exist solution methods generating explicit trajectories (continuous or iterative), which converge monotonically in the norm to the problem solution [7, 9].

In this chapter, an optimal control problem (formulated in Sect. **??**) is treated as a *saddle model*. This problem has linear dynamics and boundary conditions at the right-hand end points given in the form of finite-dimensional problem of convex programming. Section 3 provides some background results from finite-dimensional convex programming together with general formulation of proximal-type methods

aimed at finding a saddlepoint of an augmented Lagrangian. In Sect. 3, the optimal control problem is transformed into a saddle model. Section 4 describes the augmented Lagrangian approach in infinite-dimensional functional spaces and Sect. 5 presents a saddlepoint method of proximal type. The properties of weak convergence in functional spaces are briefly revised in Sect. 6 and the convergence of the proposed method is formally proved in Sect. 8.

2 Problem Formulation and Preliminaries

We consider an optimal control problem with linear dynamics on a fixed time interval $[t_0, t_1]$ and with conditions at the end points of the interval, which are implicitly specified as solutions to optimization problems. If the controls $u(t) \in U$ run through the entire set of controls, the left- and right-hand end points of the corresponding trajectories $x(t_0) = x_0$ and $x(t_1) = x_1$ describe the terminal (or boundary) sets $X_0 \subseteq \mathbb{R}^n$ and $X_1 \subseteq \mathbb{R}^n$, respectively. Both such sets X_0 and X_1 can be, in particular, convex and closed. The "left" set X_0 is called the *set of initial conditions*, while the "right" one X_1 is referred to as the *set of attainability*. We consider the direct (or Cartesian) product of these sets $\{x_0, x_1 \mid x_0, x_1 \in X(t_0) \times X(t_1)\}$, on which a convex terminal function $\varphi(x_0, x_1)$ is defined. Two components of the optimum of this function $x_0^*, x_1^* \in X(t_0) \times X(t_1)$ engender, respectively, the initial and terminal conditions for a controlled dynamic system.

In this situation, the optimal control problem with boundary conditions can be formulated as follows: Find an optimal control $u^*(t)$ whose underlying trajectory $x^*(t)$ engenders the left- and right-hand end points x_0^*, x_1^* representing together the coordinates of the point that minimizes the terminal function $\varphi(x_0, x_1)$ over the set $X(t_0) \times X(t_1)$.

Now we provide a formal statement of the problem. Since the main focus of the work will be placed on developing methods of solving the problem, let us formulate it in a Hilbert space:

$$x_0^*, x_1^* \in \operatorname{Argmin} \left\{ \varphi(x_0, x_1) \;\middle|\; A_0 x_0 + A_1 x_1 = a,\ (x_0, x_1) \in X_0 \times X_1 \right\}. \tag{1}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} x(t) = D(t)x(t) + B(t)u(t), \quad x^*(t_0) = x_0^* \in X_0,\ x^*(t_1) = x_1^* \in X_1, \tag{2}$$

$$U = \left\{ u(t) \in L_2^r[t_0, t_1] \;\middle|\; \int_{t_0}^{t_1} u^2(t)\mathrm{d}t \le \rho^2 \right\}, \quad t_0 \le t \le t_1. \tag{3}$$

Here $D(t)$, $B(t)$ are matrix functions (continuously dependent on time) of dimensions $n \times n$ and $n \times r$, respectively, and A_0, A_1 are fixed numerical matrices of dimension $n \times n$. The control functions belong to a convex set U which is closed and bounded in the norm of Hilbert space $L_2^r[t_0, t_1]$ and $\rho > 0$ is a given constant. It should

be noted that some functions of such a set can be unbounded. Therefore, any pair $(x(\cdot), u(\cdot)) \in L_2^n[t_0, t_1] \times U$ that satisfies identically the following condition

$$x(t) = x(t_0) + \int_{t_0}^{t} [D(\tau)x(\tau) + B(\tau)u(\tau)]\,\mathrm{d}\tau, \quad t_0 \le t \le t_1. \tag{4}$$

should be understood as a solution to the differentional equation (3), (4).

It is shown in ([27], vol. 1, p. 443) that for any control $u(\cdot) \in U \subseteq L_2^r[t_0, t_1]$ there is a unique underlying trajectory $x(\cdot)$ of the linear differential system, such that a resulting pair satisfies the identity (4). In many applications, the control $u(t)$ is often defined as a *piecewise continuous* function. The presence of finite discontinuities in control functions does not affect the continuity of underlying state trajectory $x(t)$. Strictly speaking, all state trajectories are *absolutely continuous functions*[1]. A set of absolutely continuous functions denoted as $AC^n[t_0, t_1]$ is actually a dense subset in the functional Hilbert space $L_2^n[t_0, t_1]$, that is $AC^n[t_0, t_1] \subset L_2^n[t_0, t_1]$. Moreover, the trajectory $x(t)$ will remain absolutely continuous even if we change the values of $u(t)$ on a set of measure zero. The latter implies that Newton–Leibniz formulas are fulfilled over a linear manifold of pairs of functions $(x(\cdot), u(\cdot))$ and an integration by parts can be performed.

Actually, problem (1)–(3) can be treated as a convex programming problem formulated in functional spaces, more precisely on the Cartesian product of subsets $X(t_0) \times X(t_1) \times AC^n[t_0, t_1] \times U$ of the corresponding spaces $\mathbb{R}^n \times \mathbb{R}^n \times L_2^n[t_0, t_1] \times L_2^r[t_0, t_1]$.

It is worthwhile to recall that the scalar products and norms in these spaces are defined as:

$$\langle x(\cdot), y(\cdot)\rangle = \int_{t_0}^{t_1} \langle x(t), y(t)\rangle \mathrm{d}t, \quad |x(\cdot)|^2 = \int_{t_0}^{t_1} |x(t)|^2 \mathrm{d}t,$$

where

$$\langle x(t), y(t)\rangle = \sqrt{\sum_{i=1}^{n} x_i(t) y_i(t)}, \quad |x(t)|^2 = \sum_{i=1}^{n} x_i^2(t)$$

and

$$x(\cdot) = (x_1(t), ..., x_n(t))^{\mathrm{T}}, \; y(\cdot) = (y_1(t), ..., y_n(t))^{\mathrm{T}}.$$

Since this chapter is focused on the convergence of solution methods, it is assumed that solutions $\left(x_0^*, x_1^*\right) \in X_1 \times X_2$, $x^*(\cdot) \in AC^n[t_0, t_1] \subseteq L_2^n[t_0, t_1], u^*(\cdot) \in U \subseteq L_2^r[t_0, t_1]$ always exist [10].

The system (1)–(3) works as follows. The linear controllable system (2) and (3) is treated as a linear constraint, which selects a linear manifold of functions (processes)

[1] For formal definitions of absolutely continuous functions and their properties, refer to [31, p. 270], [18, p. 361] or other similar text books.

$(x(\cdot), u(\cdot))$, defined on $[t_0, t_1]$. As already mentioned, the right- and left-hand end points of the system trajectories engender the direct product $X(t_0) \times X(t_1)$. On this set the function $\varphi(x_0, x_1)$ allocates (or "highlights") a single point of minimum or a closed convex set of minimum points. Each point has two coordinates x_0 and x_1. They will be treated as initial and terminal conditions for dynamic system (2), respectively. Now the optimal control problem can be formulated in the following way: find a control $u^*(\cdot) \in U$ under which both endpoints of the underlying trajectory $x^*(\cdot)$ coincide with the coordinates of the minimum point of the function $\varphi(x_0, x_1)$ over the set $X_0 \times X_1$, where X_0 is a set of initial conditions, X_1 is a set of attainability of the dynamical system (2).

This formulation includes a variety of particular options. For example, if the terminal function and terminal constraints are split with respect to its variables (that is, $\varphi(x_0, x_1) = \varphi_0(x_0) + \varphi_1(x_1)$ and $A_0 x_0 = a_0,\ A_0 x_1 = a_1,\ a_0 + a_1 = a$), then the problem (1)–(3) takes a form where the initial condition at the left-hand end points is defined as a solution to the problem of convex programming:

$$x_0^* \in \operatorname{Argmin}\{\varphi_0(x_0) \mid A_0 x_0 = a_0,\ x_0 \in X_0\}, \tag{5}$$

while at the right-hand end points, a problem of optimal control is posed:

$$x_1^* \in \operatorname{Argmin}\left\{\varphi_1(x_1) \mid A_1 x_1 = a_1,\ x_1 \in X_1\right\}. \tag{6}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = D(t)x(t) + B(t)u(t), \quad x(t_0) = x_0^* \in X_0,\ x(t_1) = x_1^* \in X_1, \tag{7}$$

$$U = \left\{u(t) \in L_2^r[t_0, t_1] \mid \int_{t_0}^{t_1} |u(t)|^2 \mathrm{d}t \le \rho^2\right\}, \quad t_0 \le t \le t_1. \tag{8}$$

Here, problem (5) provides a solution that will effectively serve as initial condition for the dynamics (7). Then we choose a control $u^*(\cdot) \in U \subseteq L_2^r[t_0, t_1]$, with underlying trajectory $x^*(\cdot) \in AC^n[t_0, t_1] \subset L_2^n[t_0, t_1]$ that transforms the initial state x_0^* into the final state x_1^*, i.e., connects by a "curve" the solutions of problems (5) and (6). Thus, it turns out that the finite-dimensional problem is transferred by the system dynamics from one state (initial state) to another (implicitly defined terminal state). It should be noted that $x^*(t)$, as a solution to (7), is also an *absolutely continuous* function (see formal definition and properties in ([31], p. 270) or [18], p. 361)).

In the simplest case, that is, when the boundary condition at the left-hand end points of the time interval is specified as a fixed initial condition, we have:

$$x_1^* \in \operatorname{Argmin}\left\{\varphi_1(x_1) \mid A_1 x_1 = a_1,\ x_1 \in X_1\right\}, \tag{9}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = D(t)x(t) + B(t)u(t), \quad x(t_0) = x_0 \in X_0,\ x^*(t_1) = x_1^* \in X_1, \tag{10}$$

$$U = \left\{u(t) \in L_2^r[t_0, t_1] \mid \int_{t_0}^{t_1} |u(t)|^2 \mathrm{d}t \le \rho^2\right\}, \quad t_0 \le t \le t_1. \tag{11}$$

A meaningful picture of (9)–(11) is simplified even more, and can be interpreted in the following way. It is required to choose a control $u^*(\cdot) \in U \subseteq L_2^r[t_0, t_1]$ with underlying trajectory $x^*(\cdot) \in AC^n[t_0, t_1] \subset L_2^n[t_0, t_1]$, generated by the dynamics (10) with initial condition $x(t_0) = x_0$ whose right-hand end point will eventually hit the terminal point indicated by the problem (9).

In the terminal problems (1), (5), (6), (9), the equality-type constraints can be also replaced by inequality-type constraints.

3 Finite-Dimensional Convex Programming

Our approach for solving the problems formulated in Sect. 2 is based on the analysis and reasoning, primarily developed for the finite convex programming problems (see, e.g., [11, 22]). Recall that the concept of convexity and Lagrange function are fundamental features of this theory. It should be also noted that optimal control problems with convex terminal functions and linear controllable ordinary differential equation (ODE) systems are, in effect, convex programming problems formulated in infinite-dimensional function spaces.

These problems are fundamental in control theory. To solve such a problem implies to find a synthesis of control function in the form of feedback. In this chapter, we propose another approach to solving this type of problems.

At first, we recall the main concepts related to finite-dimensional convex programming. For simplicity sake, we consider a problem with equality constraints [27]:

$$x^* \in \operatorname{Argmin}\{\varphi(x) \mid Ax + a = 0,\ x \geq 0\} \tag{12}$$

where $\varphi(x)$ is a convex function, $x \in \mathbb{R}^n_+$ and A is a numerical matrix of dimension $m \times n$. The minimum point $x^* \geq 0$ of the above problem is characterized by the following feature. There exists a set of multipliers $p^* \in \mathbb{R}^m$ (not simultaneously zeros) which are coordinates of the gradient of the objective function evaluated at x^* corresponding to the basis formed by the gradients of constraints, namely, $\nabla\varphi(x^*) = A^{\mathrm{T}} p^*$. Such vector $p^* \in \mathbb{R}^m$ is referred to as *dual solution* of the convex programming problem (12).

Therefore, together with problem (12), it is worthwhile to consider its scalarization (or linear convolution), known as function of Lagrange (or Lagrangian):

$$L(p, x) = \varphi(x) + \langle p, Ax + a\rangle \tag{13}$$

defined for all $x \in \mathbb{R}^n_+,\ p \in \mathbb{R}^m$. As a rule, this function has a saddlepoint $x^*, p^* \in \mathbb{R}^n_+ \times \mathbb{R}^m$ that satisfies the inequalities:

$$\varphi(x^*) + \langle p, Ax^* + a\rangle \leq \varphi(x^*) + \langle p^*, Ax^* + a\rangle \leq \varphi(x) + \langle p^*, Ax + a\rangle \tag{14}$$

for all $(x, p) \in \mathbb{R}^n_+ \times \mathbb{R}^m$. As stated by the theory of convex programming, in all regular cases (when Slater's condition holds), direct and dual solutions to a convex

programming problem effectively determine a saddlepoint of the Lagrange function, and vice versa.

Moreover, the saddle system can be written in other equivalent forms. For example, as:

$$\varphi(x^*) + \langle p^*, Ax^* + a\rangle \le \varphi(x) + \langle p^*, Ax + a\rangle, \tag{15}$$

$$Ax^* + a = 0 \tag{16}$$

for all $x \in \mathbb{R}^n_+$ or, alternatively, as:

$$p^* = p^* + k(Ax^* + a) \tag{17}$$

where $k > 0$.

It is known that the first component of the saddlepoint $p^* \in \mathbb{R}^m$ of (14) is a solution to maximization problem of the function of minima (i.e., dual function [27])

$$\max\vartheta_1(p) = \max\left\{\min\{L(p,x) \mid x \in \mathbb{R}^n_+\}\right\}, \tag{18}$$

and the second component $x^* \in \mathbb{R}^n_+$ (solution of 12) represents a solution to minimization problem of the function of maxima:

$$\min\vartheta_2(x) = \min\left\{\max\{L(p,x) \mid p \in \mathbb{R}^m\}\right\}. \tag{19}$$

Since the function $L(p,x)$ is convex–concave with respect to its variables then, according to [27], it is fulfilled that

$$\max\min L(p,x) = \min\max L(p,x). \tag{20}$$

It should be noted that $\vartheta_1(p)$ is concave while $\vartheta_2(x)$ is convex. The latter implies that instead of finding a saddlepoint of the function $L(p,x)$, one may also try to solve the problem (18), i.e., to maximize the concave dual function using, for example, the gradient method. However, in this case, other difficulties may arise. Namely, the dual function $\vartheta_1(p)$ is not differentiable, it is only *subdifferentiable*; therefore, solution methods based on the property of subdifferentiability are rather ineffective since they do not converge monotonically (in the space norm) to the optimum.

This obstacle can be overcome by using the idea of regularization of the Lagrangian with respect to its direct variables:

$$M(p,x) = \varphi(x) + \frac{1}{2k}\,|p + k(Ax + a)|^2 - \frac{1}{2k}|p|^2,\ x \ge 0,\ p \in \mathbb{R}^m \tag{21}$$

for all $(x, p) \in \mathbb{R}^n_+ \times \mathbb{R}^m$. The above function is easily obtained by adding a squared residual of functional constraints to the Lagrangian (13), and then turning the underlying expression into a perfect square. The result of such synthesis, which is usually called *augmented Lagrangian*, represents a "quadratic convolution" of the objective function and underlying constraints.

Augmented Lagrangian has a number of advantages. The most important of them consists in the fact that both traditional and augmented Lagrangians have the same set of saddlepoints. Another important property is that the dual function:

$$\vartheta_1(p) = \min\left\{M(p,x) \;\middle|\; x \in \mathbb{R}^n_+\right\}, \tag{22}$$

generated by the augmented Lagrangian is *differentiable* and its gradient satisfies Lipschitz condition with constant $k > 0$ (see more details in [11]).

Actually, for augmented Lagrangian the saddle system of inequalities similar to (14) can be written as:

$$\begin{aligned}
&\varphi(x^*) + \frac{1}{2k}|p + k(Ax^* + a)|^2 - \frac{1}{2k}|p|^2 \\
&\leq \varphi(x^*) + \frac{1}{2k}|p^* + k(Ax^* + a)|^2 - \frac{1}{2k}|p^*|^2 \\
&\leq \varphi(x) + \frac{1}{2k}|p^* + k(Ax + a)|^2 - \frac{1}{2k}|p^*|^2,
\end{aligned} \tag{23}$$

for all $(x, p) \in \mathbb{R}^n_+ \times \mathbb{R}^m$. From the left-hand inequality of this system, we have

$$Ax^* + a = 0, \tag{24}$$

while the right-hand inequality yields:

$$\varphi(x^*) \leq \varphi(x), \tag{25}$$

provided that the variable x satisfies scalar constraint:

$$\frac{1}{2k}|p^* + k(Ax + a)|^2 \leq \frac{1}{2k}|p^*|^2.$$

Two last formulas lead to the verification of (12), meaning that a pair (p^*, x^*) is also a saddlepoint of the traditional Lagrangian (14).

It is known [11] that in order to compute the gradient of the function of minima generated by augmented Lagrangian at the point $p^{'}$, one should find an optimum $x^{'} \in \mathbb{R}^n_+$ of $M(p,x)$ with respect to the variable $x \in \mathbb{R}^n_+$ for fixed $p^{'} \in \mathbb{R}^m$, and then calculate the value of the partial derivative $\dfrac{\partial M(p^{'}, x^{'})}{\partial p} = \nabla_p M\left(p^{'}, x^{'}\right)$ at the point $\left(p^{'}, x^{'}\right)$. In this case, we obtain the gradient of the dual function:

$$\nabla_{p'} M\left(p^{'}, x^{'}\right) = \frac{1}{k}\left(p^{'} + k\left(Ax^{'} + a\right)\right) - \frac{1}{k}p^{'} = Ax^{'} + a,$$

which also satisfies the Lipschitz condition. The proof of this fact plainly follows from (23).

Using the resulting gradient, we can formulate the so-called *method of augmented Lagrangian*, also referred to as "multipliers method" or, more precisely, gradient

method for maximization of the dual function engendered by augmented Lagrangian. This method has the form:

$$x^{n+1} \in \operatorname{Argmin}\left\{M\left(x, p^n\right) \mid x \in \mathbb{R}^n_+\right\},$$
$$p^{n+1} = p^n + k\left(Ax^{n+1} + a\right), \tag{26}$$

where $k > 0$.

The convergence of this method had been proved by several authors (see, e.g., [11, 13, 22–25]). Their assertions on convergence were confined to the fact that, under an additional condition of boundedness of the sequence $x^n \in \mathbb{R}^n_+$, the iterative process converges monotonically with respect to dual variable $p \in \mathbb{R}^m$ in (an adequate) norm. To ensure convergence with respect to direct variable $x \in \mathbb{R}^n_+$, one should introduce a mechanism of approximation to an optimum with respect to this variable. For example,

$$x^{n+1} \in \operatorname{Argmin}\left\{\frac{1}{2}\left|x - x^n\right|^2 + kM\left(x, p^n\right) \mid x \in \mathbb{R}^n_+\right\},$$
$$p^{n+1} = p^n + k\left(Ax^{n+1} + a\right). \tag{27}$$

To obtain some estimates which will be needed for formal proofs, it is useful to write the process (27) in the form of variational inequalities, namely:

$$\langle x^{n+1} - x^n + k\left(\nabla\varphi(x^{n+1}) + A^{\mathrm{T}}\left(p^n + k\left(Ax^{n+1} + a\right)\right)\right), x - x^{n+1}\rangle \geq 0,$$
$$p^{n+1} = p^n + k\left(Ax^{n+1} + a\right) \tag{28}$$

for all $x \in \mathbb{R}^n_+$, $p \in \mathbb{R}^m$. In view of Eq. (26), the inequality (28) can be written as:

$$\langle x^{n+1} - x^n + k\left(\nabla\varphi(x^{n+1}) + A^{\mathrm{T}}p^{n+1}\right), x - x^{n+1}\rangle \geq 0,$$
$$p^{n+1} = p^n + k\left(Ax^{n+1} + a\right) \tag{29}$$

Using the saddle systems (14), (23), we can write the variational inequalities similar to (29) for the saddlepoint $p^*, x^* \in \mathbb{R}^m \times \mathbb{R}^n_+$:

$$\langle \nabla\varphi(x^*) + kA^{\mathrm{T}}p^*, x - x^*\rangle \geq 0,$$
$$Ax^* + a = 0, \tag{30}$$

for all $p \in \mathbb{R}^m, x \geq 0$.

Now, we can prove a theorem on convergence of the gradient-type method (27).

Theorem 1 *If the set of solutions of (12) is nonempty, then iterative process (27) converges monotonically in the space norm to some direct and dual solutions of this problem, i.e., to a saddlepoint of Lagrangian x^*, p^*. In other words, $|x^n - x^*|^2 + |p^n - p^*|^2 \to 0$ when $n \to \infty$ and for all $x^0, p^0 \in \mathbb{R}^n_+ \times \mathbb{R}^m$.*

Proof Let $x = x^*$ in (29) and set $x = x^{n+1}$ in (30), then

$$\left\langle x^{n+1} - x^n + k\left(\nabla\varphi(x^{n+1}) + A^{\mathrm{T}} p^{n+1}\right), x^* - x^{n+1}\right\rangle \geq 0,$$
$$\left\langle \nabla\varphi(x^*) + kA^{\mathrm{T}} p^*, x^{n+1} - x^*\right\rangle \geq 0. \tag{31}$$

Summing up the above inequalities we obtain:

$$\left\langle x^{n+1} - x^n, x^* - x^{n+1}\right\rangle + \langle k\nabla\varphi(x^{n+1}) - \nabla\varphi(x^*), x^* - x^{n+1}\rangle$$
$$+ \left\langle p^{n+1} - p^*, (Ax^* + a) - (Ax^{n+1} + a)\right\rangle \geq 0. \tag{32}$$

From (30), for $p = p^{n+1}$, we have

$$\langle Ax^* + a, p^* - p^{n+1}\rangle \geq 0, \tag{33}$$

Using this estimate, from (32) we obtain:

$$\left\langle x^{n+1} - x^n, x^* - x^{n+1}\right\rangle + \langle k\nabla\varphi(x^{n+1}) - \nabla\varphi(x^*), x^* - x^{n+1}\rangle$$
$$- \left\langle p^{n+1} - p^*, Ax^{n+1} + a\right\rangle \geq 0. \tag{34}$$

Setting $p = p^*$ in (29), yields:

$$\left\langle p^{n+1} - p^n - k(Ax^{n+1} + a), p^* - p^{n+1}\right\rangle \geq 0$$

and hence

$$\left\langle p^{n+1} - p^n, p^* - p^{n+1}\right\rangle - k\left\langle Ax^{n+1} + a, p^* - p^{n+1}\right\rangle \geq 0. \tag{35}$$

By summing up (34) and (35) and taking into account the monotonicity of the gradient, that is, $\langle\nabla\varphi(x^{n+1}) - \nabla\varphi(x^*), x^{n+1} - x^*\rangle \geq 0$, we get:

$$\left\langle x^{n+1} - x^n, x^* - x^{n+1}\right\rangle + \left\langle p^{n+1} - p^n, p^* - p^{n+1}\right\rangle \geq 0. \tag{36}$$

Using the identity:

$$|p_1 - p_2|^2 = |p_1 - p_3|^2 + 2\langle p_1 - p_3, p_3 - p_2\rangle + |p_3 - p_2|^2,$$

the scalar product can be expanded into the sum of the squares:

$$|x^{n+1} - x^*|^2 + |p^{n+1} - p^*|^2 + |x^{n+1} - x^n|^2 + |p^{n+1}$$
$$- p^n|^2 \leq |x^n - x^*|^2 + |p^n - p^*|^2.$$

The latter indicates that by increasing n the quantity $|x^n - x^*|^2 + |p^n - p^*|^2$ will decrease monotonically. By summing up the above inequality from $n = 0$ to $n = N$, it is obtained that

$$|x^{N+1} - x^*|^2 + |p^{N+1} - p^*|^2 + \sum_{k=0}^{N}(|x^{k+1}$$
$$- x^k|^2 + |p^{k+1} - p^k|^2) \leq |x^0 - x^*|^2 + |p^0 - p^*|^2.$$

This last inequality implies the boundedness of the trajectory $|x^{N+1}-x^*|^2+|p^{N+1}-p^*|^2 \le |x^0-x^*|^2+|p^0-p^*|^2$, the convergence of the series $\sum_{k=0}^{\infty}|x^{k+1}-x^k|^2 < \infty$, $\sum_{k=0}^{\infty}|p^{k+1}-p^k|^2 < \infty$ and, therefore, the tendency to zero of the quantities $|x^{n+1}-x^n|^2 \to 0$ and $|p^{n+1}-p^n|^2 \to 0$ when $n \to \infty$.

Since the sequence $\{x^n, p^n\}$ is bounded, then there exists an element $\left(x^{'}, p^{'}\right)$, such that $x^{n_i} \to x^{'}, p^{n_i} \to p^{'}$ when $n_i \to \infty$, while $|x^{n_i+1}-x^{n_i}|^2 \to 0$ and $|p^{n_i+1}-p^{n_i}|^2 \to 0$.

Finally, by passing to limit in the inequality (29) for all $n_i \to \infty$, we arrive to

$$\left\langle \nabla\varphi(x^{'})+A^{\mathrm{T}}p^{'}, x-x^{'}\right\rangle \ge 0,\ x \ge 0,$$
$$Ax^{'}+a=0.$$

These inequalities coincide with (30); therefore, $x^{'}=x^*, p^{'}=p^* \in \mathbb{R}^n_+ \times \in \mathbb{R}^m$, that is, any limit point of the sequence $\{x^n, p^n\}$ is a solution to our problem. Additionally, monotone decreasing of the quantity $|x^n-x^*|^2+|p^n-p^*|^2$ ensures uniqueness of the limit point, and also convergence in the sense $x^n \to x^*, p^n \to p^*$ for $n \to \infty$.

Thus, we have shown that, in regular case, the problem of a finite-dimensional convex programming is equivalent to the calculation of the saddlepoint of the Lagrange function. We have also established that by using the so-called augmented (or regularized) Lagrangian, one can construct the gradient-type methods which converge monotonically (in the space norm) to the saddlepoint of Lagrange function. It is worth noting that first component of the saddlepoint provides solution of the direct (primary) problem, while its second component engenders solution of the dual one. Other methods of similar type can be found in [5].

Problem (5)–(8) and its particular cases are classified as problems of convex programming formulated in functional spaces; therefore, this approach can be applied to optimal control problems with boundary conditions at the end points of the time interval.

4 Reduction of the Optimal Control Problem to Computation of Lagrangian's Saddlepoint

Problems (6)–(8) can be formally viewed as convex programming problems formulated in the Hilbert space $\mathbb{R}^n \times L_2^n[t_0,t_1] \times L_2^r[t_0,t_1]$ of infinite dimension.

Let us introduce the Lagrange function:

$$L\left(p_1,x_1,\psi(\cdot),x(\cdot),u(\cdot)\right)=\varphi_1(x_1)+\langle p_1, A_1x_1+a_1\rangle + \int_{t_0}^{t_1}\left\langle \psi(t), D(t)x(t)+B(t)u(t)-\frac{\mathrm{d}}{\mathrm{d}t}x(t)\right\rangle \mathrm{d}t, \quad (37)$$

for all $p_1 \in \mathbb{R}^m, x_1 \in \mathbb{R}^n, (x(\cdot),u(\cdot)) \in AC[t_0,t_1] \times U \subseteq L_2^n[t_0,t_1] \times L_2^r[t_0,t_1]$, and $\psi(\cdot) \in \Psi_2^n[t_0,t_1] \subseteq L_2^n[t_0,t_1]^\star$. Here, $\Psi_2^n[t_0,t_1]$ is a linear manifold of *absolutely*

continuous functions which is dense in $L_2^n[t_0, t_1]$. In other words, the closure of this manifold $\Psi_2^n[t_0, t_1]$ in the norm of $L_2^n[t_0, t_1]$ coincides with $L_2^n[t_0, t_1]$. Each element $\psi(\cdot) \in \Psi_2^n[t_0, t_1]$ is treated as a normal of a linear functional:

$$\int_{t_0}^{t_1} \langle \psi(t), x(t)) \rangle \, \mathrm{d}t \tag{38}$$

of the *dual* space $L_2^n[t_0, t_1]^\star$ of linear functionals (38) defined on $AC^n[t_0, t_1] \subset L_2^n[t_0, t_1]$ for all admissible $u(\cdot) \in U \subset L_2^r[t_0, t_1]$. Alternatively, $\Psi_2^n[t_0, t_1]$ can be viewed as a "dual image" of all possible trajectories $x(\cdot) \in AC^n[t_0, t_1]$ for underlying $u \in U$ of the primal differential equation (10) in the *dual* space $L_2^n[t_0, t_1]^\star$ of linear functionals (38). Therefore, $\Psi_2^n[t_0, t_1]$ describes a linear subspace of solutions to homogeneous differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\psi(t) + D^{\mathrm{T}}(t)\psi(t) = 0, \tag{39}$$

also known as *adjoint equation*. Additionally, the kernel of primal Eq. (10) is orthogonal to the image of adjoint equation (39) and the kernel of (39) is orthogonal to the image of (10). The latter is attributed to the fact that Hilbert space $L_2^n[t_0, t_1]$ is self-adjoint.

According to [17], in the regular case (i.e., under Slater's condition), problem (6)–(8) always has a solution reducible to computation of a saddlepoint of the Lagrangian (37).

Naturally, a saddlepoint $\big((p_1^*, \psi^*(\cdot)), (x_1^*, x^*(\cdot), u^*(\cdot))\big)$ of Lagrange function is formed by *dual* $\big(p_1^*, \psi^*(\cdot)\big)$ and *direct* $\big(x_1^*, x^*(\cdot), u^*(\cdot)\big)$ components and must satisfy by definition the system of inequalities:

$$\begin{aligned}
&\varphi_1(x_1^*) + \langle p_1, A_1 x_1^* + a_1 \rangle + \int_{t_0}^{t_1} \left\langle \psi(t), D(t)x^*(t) + B(t)u^*(t) - \frac{\mathrm{d}}{\mathrm{d}t}x^*(t) \right\rangle \mathrm{d}t \\
&\leq \varphi_1(x_1^*) + \langle p_1^*, A_1 x_1^* + a_1 \rangle + \int_{t_0}^{t_1} \left\langle \psi^*(t), D(t)x^*(t) + B(t)u^*(t) - \frac{\mathrm{d}}{\mathrm{d}t}x^*(t) \right\rangle \mathrm{d}t \\
&\leq \varphi_1(x_1) + \langle p_1^*, A_1 x_1 + a_1 \rangle + \int_{t_0}^{t_1} \left\langle \psi^*(t), D(t)x(t) + B(t)u(t) - \frac{\mathrm{d}}{\mathrm{d}t}x(t) \right\rangle \mathrm{d}t,
\end{aligned} \tag{40}$$

for all $p_1 \in \mathbb{R}^m, \psi(\cdot) \in \Psi_2^n[t_0, t_1] \subset L_2^n[t_0, t_1]^\star, x_1 \in \mathbb{R}^n, (x(\cdot), u(\cdot)) \in AC^n[t_0, t_1] \times U \subset L_2^n[t_0, t_1] \times L_2^n[t_0, t_1]^\star, x(t_0) = x_0$.

The left-hand inequality of this system represents a maximization problem of a linear function depending on the variables $(p_1, \psi(\cdot))$ over the domain $\mathbb{R}^m \times \Psi_2^n[t_0, t_1]$. This inequality yields:

$$\begin{aligned}
&\langle p_1 - p_1^*, A_1 x_1^* + a_1 \rangle \\
&\quad + \int_{t_0}^{t_1} \left\langle \psi(t) - \psi^*(t), D(t)x^*(t) + B(t)u^*(t) - \frac{\mathrm{d}}{\mathrm{d}t}x^*(t) \right\rangle \mathrm{d}t \leq 0,
\end{aligned} \tag{41}$$

where $p_1 \in \mathbb{R}^m, \psi(\cdot) \in \Psi_2^n[t_0, t_1]$. Due to linearity with respect to $p_1, \psi(\cdot)$, the last inequality holds, if and only if,

$$A_1 x_1^* + a_1 = 0, \quad D(t)x^*(t) + B(t)u^*(t) - \frac{\mathrm{d}}{\mathrm{d}t}x^*(t) = 0, \quad x(t_0) = x_0. \tag{42}$$

The right-hand inequality of the system (40) represents a minimization problem of Lagrange function with respect to the variables $x_1, x(\cdot), u(\cdot)$ for fixed values $p_1 = p_1^*, \psi(\cdot) = \psi^*(t)$. Let us show that the system of vectors $(p_1^*, x_1^*, \psi^*(\cdot), x^*(\cdot), u^*(\cdot))$ is a solution of the problem (6)–(8). In view of (42), it follows from the right-hand inequality of (40) that

$$\varphi_1(x_1^*) \le \varphi_1(x_1) + \langle p_1^*, A_1 x_1 + a_1 \rangle + \int_{t_0}^{t_1} \left\langle \psi^*(t), D(t)x(t) + B(t)u(t) - \frac{\mathrm{d}}{\mathrm{d}t}x(t) \right\rangle \mathrm{d}t, \tag{43}$$

for all $x_1 \in \mathbb{R}^n, (x(\cdot), u(\cdot)) \in AC^n[t_0, t_1], \times U$. Under the conditions

$$\langle p_1^*, A_1 x_1 + a_1 \rangle = 0, \quad \int_{t_0}^{t_1} \left\langle \psi^*(t), D(t)x(t) + B(t)u(t) - \frac{\mathrm{d}}{\mathrm{d}t}x(t) \right\rangle \mathrm{d}t = 0,$$

the inequality (43) results in an optimization problem:

$$\varphi_1(x_1^*) \le \varphi_1(x_1), \tag{44}$$

subject to equality-type constraints

$$\langle p^*, A_1 x_1 + a_1 \rangle = 0, \quad \int_{t_0}^{t_1} \left\langle \psi^*(t), D(t)x(t) + B(t)u(t) - \frac{\mathrm{d}}{\mathrm{d}t}x(t) \right\rangle \mathrm{d}t = 0 \tag{45}$$

for all $x_1 \in \mathbb{R}^n, (x(\cdot), u(\cdot)) \in AC^n[t_0, t_1], \times U$.

From (42), it follows that $x_1^*, x^*(t), u^*(t)$ are solutions to these vector equations. On the other hand, there exists $u^*(t)$ that engenders $x^*(t)$ and x_1^* which is a minimizer of (44) under scalar constraints (45). Solution sets of (44) and (45) are "wider" than those engendered by vector constraints (42) in the sense that all solutions of (42) are contained in the solution sets of (44) and (45).

Therefore, $x_1^*, x^*(t), u^*(t)$ (as elements of "wider" solution sets of (44) and (45) under scalar constraints) will also belong to a "narrower" *subset* of solutions corresponding to vector constraints (43), that is,

$$\varphi_1(x_1^*) \le \varphi_1(x_1) \tag{46}$$

subject to

$$A_1 x_1 + a_1 = 0, \quad x_1 \in \mathbb{R}^n,$$
$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = D(t)x(t) + B(t)u(t), \quad x(t_0) = x_0, \quad u(\cdot) \in U, \tag{47}$$

for all $x_1 \in \mathbb{R}^n$, $(x(\cdot), u(\cdot)) \in AC^n[t_0, t_1], \times U$. In other words, the saddlepoint of the Lagrange function (37) is a solution to the original problem (6)–(8).

The converse statement (including infinite-dimensional cases) is known as *Karush–Kuhn–Tucker theorem* or *KKT theorem* (see, e.g., [17, 20]). In consequence, if the Lagrangian (37) has a saddlepoint, then its components provide both primal and dual solutions of the original convex programming problem in infinite-dimensional space.

5 Augmented Lagrangian Approach

Lagrange function (37), as well as any function of two variables, always engenders a function of minima (and a function of maxima):

$$\Theta(p_1, \psi(t)) = \min\left\{ L(p_1, x_1, \psi(\cdot), x(\cdot), u(\cdot)) \middle| (x_1, x(\cdot), u(\cdot)) \in \mathbb{R}^n \times AC^n[t_0, t_1] \times U \right\}.$$

In regular finite-dimensional case, the function of minima is concave; moreover, its maximum coincides with the saddle value of the Lagrangian and is attained at $p_1 = p_1^*, \psi(\cdot) = \psi^*(\cdot)$. This point is referred to as *dual solution* of the convex programming problem. Function $\Theta(p_1, \psi(\cdot))$ is also known as *dual function* (see [27]). It may seem that by employing the properties of dual function, one could have applied the gradient approach for calculation of its maximum point and thus to find a solution of the original problem. However, in general (even in finite-dimensional) case this function is *not* differentiable (only *sub*differentiable); therefore, its gradient does not satisfy the Lipschitz condition, as compelled by convergence requirements of the gradient methods. A similar approach was effectively applied to some game-theoretical problems (see [28–30]).

However, Lagrange function is tolerable to various modifications (see, e.g., [1, 11, 22]) due to its linearity in dual variable; in particular, it admits a quadratic regularization with respect to constraints. In this case, the dual function acquires good properties of smoothness, becomes differentiable and its gradient satisfies Lipschitz condition. All this avails for the formulation of gradient-type methods aimed at maximization of the dual function. Around this fact, various authors (see, e.g., [1, 11, 22] and references therein) had developed an extensive theory for calculation of saddlepoints of convex–concave functions; the latter also fits into more general framework of the theory aimed at the calculation of fixed points in the context of equilibrium programming problems [4–6].

Naturally, all stated above refers to the finite-dimensional theory. A similar approach for optimal control problems was initially proposed in [8]. Generally speaking, such an approach appears to be rather promising and evokes various generalizations.

This chapter is focused on an optimal control problem with boundary conditions at both endpoints of the time interval which are given in the form of convex programming problems. In view of the foregoing, we introduce an *augmented Lagrangian*

for optimal control problem (6)–(8) in the form:

$$M\left(p_1, x_1, \psi(\cdot), x(\cdot), u(\cdot)\right) = \varphi_1(x_1) + \frac{1}{2k}\left|p_1 + k\left(A_1 x_1 + a_1\right)\right|^2 - \frac{1}{2k}|p_1|^2$$
$$+ \frac{1}{2k}\int_{t_0}^{t_1}\left|\psi(t) + k\left(D(t)x(t) + B(t)u(t) - \frac{\mathrm{d}}{\mathrm{d}t}x(t)\right)\right|^2 \mathrm{d}t - \frac{1}{2k}\int_{t_0}^{t_1}|\psi(t)|^2 \mathrm{d}t \tag{48}$$

which is defined for all $p_1 \in \mathbb{R}^m$, $x_1 \in \mathbb{R}^n$, $\psi(\cdot) \in \Psi_2^n[t_0, t_1]^\star$, $x(\cdot) \in AC^n[t_0, t_1]$, $u(\cdot) \in U, x(t_0) = x_0$.

Its saddle point $\left(p_1^*, x_1^*, \psi^*(\cdot), x^*(\cdot), u^*(\cdot)\right)$, formed by the primal $\left(x_1^*, x^*(\cdot), u^*(\cdot)\right)$ and dual $\left(p_1^*, \psi^*(\cdot)\right)$ solutions of the original optimal control problem (6)–(8), must satisfy (according to definition) the following system of inequalities:

$$M\left(p_1, x_1^*, \psi(\cdot), x^*(\cdot), u^*(\cdot)\right) \le M\left(p_1^*, x_1^*, \psi^*(\cdot), x^*(\cdot), u^*(\cdot)\right)$$
$$\le M\left(p_1^*, x_1, \psi^*(\cdot), x(\cdot), u(\cdot)\right),$$

for all $p_1 \in \mathbb{R}^m, x_1 \in \mathbb{R}^n, \psi(\cdot) \in \Psi_2^n[t_0, t_1]^\star, x(\cdot) \in AC^n[t_0, t_1], u(\cdot) \in U$.

In finite-dimensional case, the dual function (or function of minima), engendered by an augmented Lagrangian is concave, differentiable, and its gradient satisfies the Lipschitz condition [11]. It could be rather interesting to analyze the behavior and properties of the function:

$$\Theta_M(p_1, \psi(\cdot)) = \min\left\{M\left(p_1, x_1, \psi(\cdot), x(\cdot), u(\cdot)\right) \;\middle|\; \left(x_1, x(\cdot), u(\cdot)\right) \in \mathbb{R}^n \times AC^n[t_0, t_1] \times U\right\} \tag{49}$$

under (regular) infinite-dimensional settings. Effectively, there are two possible approaches for further development of solution methods for optimal control problems, namely:

1. To seek for a saddlepoint $\left((p_1^*, \psi^*(\cdot)), (x_1^*, x^*(\cdot), u^*(\cdot))\right)$ of the augmented Lagrangian (48).
2. To seek for a maximum point $(p_1^*, \psi^*(\cdot))$, of the dual function (49). Then, using $(p_1^*, \psi^*(\cdot))$, it will be easy to find $(x_1^*, x^*(\cdot), u^*(\cdot))$.

In the second item, we have a maximization problem of dual function that plays a role of Lyapunov function, and thus ensures stability of the maximum (unlike the first item, where no classical Lyapunov function can appear in principle, except for some symmetry cases).

Formally, this maximization problem with respect to dual variables can be articulated as a solution of the following system: find $p_1 = p_1^*, \psi(\cdot) = \psi^*(\cdot)$ such that

$$x_1^*, x^*(\cdot), u^*(\cdot) \in \operatorname{Argmin}\left\{\varphi_1(x_1) + \frac{1}{2k}\left|p_1^* + k\left(A_1 x_1 + a_1\right)\right|^2 - \frac{1}{2k}\left|p_1^*\right|^2\right.$$
$$\left. + \frac{1}{2k}\int_{t_0}^{t_1}\left|\psi^*(t) + k\left(D(t)x(t) + B(t)u(t) - \frac{\mathrm{d}}{\mathrm{d}t}x(t)\right)\right|^2 \mathrm{d}t\right.$$

$$-\frac{1}{2k}\int_{t_0}^{t_1}\left|\psi^*(t)\right|^2\mathrm{d}t \;\Big|\; x_1\in X_1,(x(\cdot),u(\cdot))\in AC^n[t_0,t_1]\times U\Big\}, \tag{50}$$

and

$$p_1^*=p_1^*+k\left(A_1x_1^*+a_1\right), \tag{51}$$

$$\psi^*(t)=\psi^*(t)+k\left(D(t)x^*(t)+B(t)u^*(t)-\frac{\mathrm{d}}{\mathrm{d}t}x^*(t)\right). \tag{52}$$

In other words, one should choose $p_1=p_1^*$ and $\psi(\cdot)=\psi^*(\cdot)$ such that

$$A_1x_1^*+a_1=0,\quad D(t)x^*(t)+B(t)u^*(t)=\frac{\mathrm{d}}{\mathrm{d}t}x^*(t). \tag{53}$$

Let us formulate a method of simple iteration, which in our case is, in effect, a gradient method aimed at maximization of the dual function $\Theta(p,\psi(\cdot))$.

Let an approximation $\left(p_1^n,\psi^n(t)\right)\in\mathbb{R}^m\times\Psi_2^n[t_0,t_1]^\star$ be given; then by solving a problem of quadratic optimization:

$$x_1^{n+1},x^{n+1}(\cdot),u^{n+1}(\cdot)\in\operatorname{Argmin}\Big\{M(p_1^n,x_1,\psi^n(\cdot),x(\cdot),u(\cdot))\;\Big|\;(x_1,x(\cdot),u(\cdot))\in\mathbb{R}^n\times AC^n[t_0,t_1]\times U\Big\}, \tag{54}$$

we can find its solution $x_1^{n+1},x^{n+1}(\cdot),u^{n+1}(\cdot)$. Then we should calculate the gradient values of the dual function, that is, $A_1x_1^{n+1}+a_1$, $D(t)x^{n+1}(t)+B(t)u^{n+1}(t)-\frac{\mathrm{d}}{\mathrm{d}t}x^{n+1}(t)$ and finally perform a gradient step according to the formulas:

$$p_1^{n+1}=p_1^n+k\left(A_1x_1^{n+1}+a_1\right), \tag{55}$$

$$\psi^{n+1}(t)=\psi^n(t)+k\left(D(t)x^{n+1}(t)+B(t)u^{n+1}(t)-\frac{\mathrm{d}}{\mathrm{d}t}x^{n+1}(t)\right). \tag{56}$$

Strictly speaking, this routine is known in finite-dimensional spaces as an *augmented Lagrangian method*. It transforms an optimal control problem into customary optimization problem. However, the convergence properties of this method are not good enough for the invention of efficient numerical routines. This method has a property of monotone decreasing (in the space norm) only with respect to dual variables; moreover, it imposes some boundedness restrictions on underlying sequences of primal variables.

6 Saddlepoint Method of Augmented Lagrangian

In this section, we consider a variant of the process (54)–(56) under regularization with respect to primal variables at each step of the process (see also [4]), namely:

$$x_1^{n+1},u^{n+1}(\cdot),x^{n+1}(\cdot)\in\operatorname{argmin}\Big\{|x_1-x_1^n|^2+\int_{t_0}^{t_1}|x(t)-x^n(t)|^2\mathrm{d}t \tag{57}$$

$$+\int_{t_0}^{t_1}|u(t)-u^n(t)|^2\mathrm{d}t+k\,M(p_1^n,x_1,\psi^n(\cdot),x(\cdot),u(\cdot)\ \Big|$$

$$x_1\in\mathbb{R}^n,(x(\cdot),u(\cdot))\in AC^n[t_0,t_1]\times U\Big\},$$

$$p_1^{n+1}=p_1^n+k(A_1x_1^{n+1}+a_1),\tag{58}$$

$$\psi^{n+1}(t)=\psi^n(t)+k\left(D(t)x^{n+1}(t)+B(t)u^{n+1}(t)-\frac{\mathrm{d}}{\mathrm{d}t}x^{n+1}(t)\right),\tag{59}$$

where $x(t_0)=x_0$. Naturally, when the first three quadratic terms of the objective function (57) are missing, we arrive to the problem (54)–(56).

At each step of this process, we solve a problem of quadratic optimization, whose *unique* minimizer is then used in order to recalculate the following approximation with respect to dual variables $(p^{n+1},\psi^{n+1}(\cdot))$. To perform these operations, we need some variational inequalities characterizing the optimal solutions. These inequalities involve the calculation of gradients of the augmented Lagrangian.

Let us consider in more details the differentiation of a quadratic function. In fact, this procedure is well-known and involves transition to the conjugate linear operators. Despite of peculiarities of linear differential operators and bulkiness of resulting formulas, we have to perform the differentiation of a quadratic function. Using the increment of quadratic function in the form:

$$\frac{1}{2}|A\,(x+\triangle x)+a|^2-\frac{1}{2}|Ax+a|^2=\langle A^{\mathrm{T}}(Ax+a),\triangle x\rangle+\frac{1}{2}|A\triangle x|^2,$$

let us write down a similar increment of augmented Lagrangian with respect to primal variables and for fixed values of $p,\psi(t)$:

$$\begin{aligned}
&M(p_1,x_1+\triangle x_1,\psi(\cdot),x(\cdot)+\triangle x(\cdot),u(\cdot)+\triangle u(\cdot))-M(p_1,x_1,\psi(\cdot),x(\cdot),u(\cdot))\\
&=\varphi_1(x_1+\triangle x_1)-\varphi_1(x_1)+\langle p_1+k\,(A_1x_1+a_1),A_1\triangle x_1\rangle+\frac{k}{2}|A_1\triangle x_1|^2\\
&+\int_{t_0}^{t_1}\left\langle\psi(t)+k\left(D(t)x(t)+B(t)u(t)-\frac{\mathrm{d}}{\mathrm{d}t}x(t)\right),D(t)\triangle x(t)+B(t)\triangle u(t)-\frac{\mathrm{d}}{\mathrm{d}t}\triangle x(t)\right\rangle\mathrm{d}t\\
&+\frac{k}{2}\int_{t_0}^{t_1}\left|D(t)\triangle x(t)+B(t)\triangle u(t)-\frac{\mathrm{d}}{\mathrm{d}t}\triangle x(t)\right|^2\mathrm{d}t.
\end{aligned}$$

Hence, using the transition formulas to conjugate linear operators:

$$\langle\psi,Dx\rangle=\langle D^{\mathrm{T}}\psi,x\rangle,\ \langle\psi,Bu\rangle=\langle B^{\mathrm{T}}\psi,u\rangle\tag{60}$$

together with integration by parts on the interval $[t_0,t_1]$,

$$\langle\psi(t_1),x(t_1)\rangle-\langle\psi(t_0),x(t_0)\rangle=\int_{t_0}^{t_1}\left\langle\frac{\mathrm{d}}{\mathrm{d}t}\psi(t),x(t)\right\rangle\mathrm{d}t+\int_{t_0}^{t_1}\left\langle\psi(t),\frac{\mathrm{d}}{\mathrm{d}t}x(t)\right\rangle\mathrm{d}t.\tag{61}$$

we can write the linear part of the increment in the following form:

$$\begin{aligned}
&\triangle M(\triangle x(\cdot), \triangle u(\cdot)) = \langle \nabla \varphi_1(x_1), \triangle x_1 \rangle + \langle A_1^{\mathrm{T}}(p_1 + k\,(A_1 x_1 + a_1)\,), \triangle x_1 \rangle \qquad (62)\\
&+ \int_{t_0}^{t_1} \left\langle D^{\mathrm{T}}(t) \left(\psi(t) + k \left[D(t)x(t) + B(t)u(t) - \frac{\mathrm{d}}{\mathrm{d}t} x(t) \right] \right), \triangle x(t) \right\rangle \mathrm{d}t\\
&+ \int_{t_0}^{t_1} \left\langle B^{\mathrm{T}}(t) \left(\psi(t) + k \left[D(t)x(t) + B(t)u(t) - \frac{\mathrm{d}}{\mathrm{d}t} x(t) \right] \right), \triangle u(t) \right\rangle \mathrm{d}t\\
&+ \int_{t_0}^{t_1} \left\langle \frac{\mathrm{d}}{\mathrm{d}t} \left(\psi(t) + k \left[D(t)x(t) + B(t)u(t) - \frac{\mathrm{d}}{\mathrm{d}t} x(t) \right] \right), \triangle x(t) \right\rangle \mathrm{d}t\\
&- \left(\left\langle \psi(t_1) + k \left[D(t)x(t_1) + B(t)u(t_1) - \frac{\mathrm{d}x}{\mathrm{d}t}(t_1) \right], \triangle x_1 \right\rangle \right.\\
&\quad \left. - \left\langle \psi(t_0) + k \left[D(t)x(t_0) + B(t)u(t_0) - \frac{\mathrm{d}x}{\mathrm{d}t}(t_0) \right], \triangle x_0 \right\rangle \right).
\end{aligned}$$

This linear part of the increment is a *differential*, i.e., a tangent plane to the augmented Lagrangian at any point $(x_1, x(\cdot), u(\cdot))$ under fixed values of dual variables $(p_1, \psi(\cdot))$.

Using this differential, we can write now a variational inequality to be satisfied by the solution of the problem (57) by way of the necessary and sufficient condition for a minimum:

$$\begin{aligned}
&\left\langle x_1^{n+1} - x_1^n, x_1 - x_1^{n+1} \right\rangle + \int_{t_0}^{t_1} \left\langle x^{n+1}(\cdot) - x^n(\cdot), x(\cdot) - x^{n+1}(\cdot) \right\rangle \mathrm{d}t\\
&+ k \left\langle \nabla \varphi_1(x_1^{n+1}), x_1 - x_1^{n+1} \right\rangle + k \left\langle A_1^{\mathrm{T}} \left(p_1^n + k \left(A_1 x_1^{n+1} + a_1 \right) \right), x_1 - x_1^{n+1} \right\rangle\\
&+ k \int_{t_0}^{t_1} \left\langle D^{\mathrm{T}}(t) \left(\psi^n(t) + k \left[D(t)x^{n+1}(t) + B(t)u^{n+1}(t) - \frac{\mathrm{d}}{\mathrm{d}t} x^{n+1}(t) \right] \right), x(t) - x^{n+1}(t) \right\rangle \mathrm{d}t\\
&+ k \int_{t_0}^{t_1} \left\langle \frac{\mathrm{d}}{\mathrm{d}t} \left(\psi^n(t) + k \left[D(t)x^{n+1}(t) + B(t)u^{n+1}(t) - \frac{\mathrm{d}}{\mathrm{d}t} x^{n+1}(t) \right] \right), x(t) - x^{n+1}(t) \right\rangle \mathrm{d}t\\
&- k \left(\left\langle \psi^n(t_1) + k \left[D(t)x^{n+1}(t_1) + B(t)u^{n+1}(t_1) - \frac{\mathrm{d}x^{n+1}}{\mathrm{d}t}(t_1) \right], x(t_1) - x^{n+1}(t_1) \right\rangle \right.\\
&\left. - \left\langle \psi^n(t_0) + k \left[D(t)x^{n+1}(t_0) + B(t)u^{n+1}(t_0) - \frac{\mathrm{d}x^{n+1}}{\mathrm{d}t}(t_0) \right], x(t_0) - x^{n+1}(t_0) \right\rangle \right)\\
&+ k \int_{t_0}^{t_1} \left\langle B^{\mathrm{T}}(t) \left(\psi^n(\cdot) + k \left[D(t)x^{n+1}(\cdot) + B(t)u^{n+1}(\cdot) - \frac{\mathrm{d}}{\mathrm{d}t} x^{n+1}(\cdot) \right] \right), u(\cdot) - u^{n+1}(\cdot) \right\rangle \mathrm{d}t\\
&+ \int_{t_0}^{t_1} \langle u^{n+1}(\cdot) - u^n(\cdot), u(\cdot) - u^{n+1}(\cdot) \rangle \mathrm{d}t \;\geq\; 0. \qquad (63)
\end{aligned}$$

Additionally, this differential permits to derive the necessary (and, in our case, also sufficient) condition for the minimum of augmented Lagrangian at the point $p_1 =$

$p_1^*, \psi(t) = \psi^*(t)$. Thus, a saddlepoint of augmented Lagrangian (and, therefore, of traditional Lagrangian) can be characterized by a variational inequality of the form:

$$\begin{aligned}
&\left\langle \nabla\varphi_1(x_1^*), x_1 - x_1^* \right\rangle + \left\langle A_1^{\mathrm{T}}(p_1^* + k\left(A_1 x_1^* + a_1\right)), x_1 - x_1^* \right\rangle \\
&+ \int_{t_0}^{t_1} \left\langle D^{\mathrm{T}}(t)\left(\psi^*(t) + k\left[D(t)x^*(t) + B(t)u^*(t) - \frac{\mathrm{d}}{\mathrm{d}t}x^*(t)\right]\right), x(t) - x^*(t)\right\rangle \mathrm{d}t \\
&+ \int_{t_0}^{t_1} \left\langle B^{\mathrm{T}}(t)\left(\psi^*(t) + k\left[D(t)x^*(t) + B(t)u^*(t) - \frac{\mathrm{d}}{\mathrm{d}t}x^*(t)\right]\right), u(t) - u^*(t)\right\rangle \mathrm{d}t \\
&+ \int_{t_0}^{t_1} \left\langle \frac{\mathrm{d}}{\mathrm{d}t}\left(\psi^*(t) + k\left[D(t)x^*(t) + B(t)u^*(t) - \frac{\mathrm{d}}{\mathrm{d}t}x^*(t)\right]\right), x(t) - x^*(t)\right\rangle \mathrm{d}t \\
&- \left\langle \psi^*(t_1) + k\left[D(t)x^*(t_1) + B(t)u^*(t_1) - \frac{\mathrm{d}x^*(t_1)}{\mathrm{d}t}\right], x_1 - x_1^* \right\rangle \\
&+ \left\langle \psi^*(t_0) + k\left[D(t)x^*(t_0) + B(t)u^*(t_0) - \frac{\mathrm{d}x(t_0)}{\mathrm{d}t}\right], x_0 - x_0^* \right\rangle \geq 0, \qquad (64)
\end{aligned}$$

for all $x_1 \in \mathbb{R}^n, (x(\cdot), u(\cdot)) \in AC^n[t_0, t_1] \times U$, where $x(t_0) = x_0, x(t_1) = x_1, \psi(t_0) = \psi_0, u(t_0) = u_0$.

Given the conditions (51), (52), in particular (52), for $t = t_0$ and $t = t_1$, we can reduce (64) to the form:

$$\begin{aligned}
&\langle \nabla\varphi_1(x_1^*), x_1 - x_1^* \rangle + \langle A_1^{\mathrm{T}} p_1^*, x_1 - x_1^* \rangle \\
&\quad + \int_{t_0}^{t_1} \left\langle D^{\mathrm{T}}(t)\psi^*(t) + \frac{\mathrm{d}}{\mathrm{d}t}\psi^*(t), x(t) - x^*(t)\right\rangle \mathrm{d}t \\
&\quad + \int_{t_0}^{t_1} \langle B^{\mathrm{T}}(t)\psi^*(t), u(t) - u^*(t)\rangle \mathrm{d}t - \langle \psi^*(t_1), x_1 - x_1^*\rangle + \langle \psi^*(t_0), x_0 - x_0^*\rangle \geq 0, \qquad (65)
\end{aligned}$$

for all $x_1 \in \mathbb{R}^n, (x(\cdot), u(\cdot)) \in AC^n[t_0, t_1] \times U$. The last term in (65) vanishes since $x(t_0) = x_0 = x_0^*$.

Hence,

$$\begin{aligned}
&\langle \nabla\varphi_1(x_1^*) + A_1^{\mathrm{T}} p_1^* - \psi^*(t_1), x_1 - x_1^* \rangle \\
&\quad + \int_{t_0}^{t_1} \left\langle D^{\mathrm{T}}(t)\psi^*(t) + \frac{\mathrm{d}}{\mathrm{d}t}\psi^*(t), x(t) - x^*(t)\right\rangle \mathrm{d}t \\
&\quad + \int_{t_0}^{t_1} \langle B^{\mathrm{T}}(t)\psi^*(t), u(t) - u^*(t)\rangle \mathrm{d}t \geq 0.
\end{aligned}$$

The resulting variational inequality can be viewed as a minimization problem of a linear function on the Cartesian product of $\mathbb{R}^n \times AC^n[t_0, t_1] \times U$ with respect to the

variables $x_1, x(t), u(t)$. This problem naturally splits into two independent ones, each with respect to its own variables:

$$\langle \nabla\varphi_1(x_1^*) + A_1^{\mathrm{T}} p_1^* - \psi^*(t_1), x_1 - x_1^* \rangle + \int_{t_0}^{t_1} \left\langle D^{\mathrm{T}}(t)\psi^*(t) + \frac{\mathrm{d}}{\mathrm{d}t}\psi^*(t), x(t) - x^*(t) \right\rangle \mathrm{d}t \geq 0,$$

$$\int_{t_0}^{t_1} \left\langle B^{\mathrm{T}}(t)\psi^*(t), u(t) - u^*(t) \right\rangle \mathrm{d}t \geq 0. \tag{66}$$

The upper inequality of (66) can be viewed as a minimization problem with respect to the variables $x_1, x(\cdot) \in \mathbb{R}^n \times AC^n[t_0, t_1]$; therefore, it can be rewritten as:

$$\frac{\mathrm{d}}{\mathrm{d}t}\psi^*(t) + D^{\mathrm{T}}(t)\psi^*(t) = 0, \quad \psi^*(t_1) = \nabla\varphi_1(x_1^*) + A_1^{\mathrm{T}} p_1^*. \tag{67}$$

Bringing together (51), (52), (66), and (67), we obtain:

$$\frac{\mathrm{d}}{\mathrm{d}t}x^*(t) = D(t)x^*(t) + B(t)u^*(t), \quad x(t_0) = x_0, \quad x(t_1) = x_1^*, \tag{68a}$$

$$p_1^* = p_1^* + k(A_1 x_1^* + a_1), \tag{68b}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\psi^*(t) + D^{\mathrm{T}}(t)\psi^*(t) = 0, \quad \psi^*(t_1) = \nabla\varphi_1(x_1^*) + A_1^{\mathrm{T}} p_1^*, \tag{68c}$$

$$\int_{t_0}^{t_1} \langle B^{\mathrm{T}}(t)\psi^*(t), u(t) - u^*(t) \rangle \mathrm{d}t \geq 0, \tag{68d}$$

for all $u(\cdot) \in U$. Thus, it has been shown that any solution of (50)–(52) is a solution of (60) as well. The converse assertion also holds.

System (60) was obtained parting from the augmented Lagrangian. However, the same system can be obtained by dealing with traditional Lagrangian, since both functions equally characterize the saddlepoint which is the same for both Lagrange functions. Actually, the terminal differential system (60) reflects that its solution is exactly a saddlepoint of the Lagrange function (37) or (48). Variational inequality (68d) is also known as an *integral maximum principle* (see more details in ([27], p. 671)).

If U is convex, then the *traditional maximum principle* of L.S. Pontryagin results from its integral form:

$$\frac{\mathrm{d}}{\mathrm{d}t}x^*(t) = D(t)x^*(t) + B(t)u^*(t), \tag{69a}$$

$$p_1^* = p_1^* + k(A_1 x_1^* + a_1), \tag{69b}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\psi^*(t) + D^{\mathrm{T}}(t)\psi^*(t) = 0, \quad \psi^*(t_1) = \nabla\varphi_1(x_1^*) + A_1^{\mathrm{T}} p_1^*, \tag{69c}$$

$$\left\langle B^{\mathrm{T}}(t)\psi^*(t), u - u^*(t) \right\rangle \geq 0, \tag{69d}$$

for all $u \in U$ at any instant t of the time interval $[t_0, t_1]$. Here, we can clearly see that (68d) and (69d) are different variational inequalities with respect to the variables $u(\cdot) \in U$. The first one is the problem of maximizing a linear function on the set $U \subset L_2^r[t_0, t_1]$ in functional space, while the second inequality represents a family of finite-dimensional variational inequalities depending on the parameter $t \in [t_0, t_1]$, each of which is a finite-dimensional problem of maximizing a linear function with respect to u. Undoubtedly, the system (70) is stronger than (60) in the sense that (70) can be formulated under no convexity of U. However, (60) clearly emphasizes its "saddle" nature, and permits to design solution techniques (within the frameworks of Hilbert space) possessing convergence to the solution of original problem with respect to *all* of its variables—control functions, state trajectories, adjoint functions, primal, and dual variables of the optimization problems at the end points. The authors are not apprised of any similar method based on the maximum principle.

Turning back to the boundary value problem (60), it is worth to emphasize again that even though (50)–(52) and (60) look rather different, they are equivalent indeed, and each of them can be solved by formally different methods. If these methods are properly justified, they will always yield the same result.

In this work, a numerical process of the form (57)–(59) is proposed for the solution of the system (50)–(52). Therefore, we should prove its convergence. In order to proceed, we need some essential facts from functional analysis. It seems reasonable to collect these facts in a separate section, and then get back to the proof of the convergence theorem.

7 Properties of Weak Convergence Applied to Linear Dynamics

1. Equation (68a) with initial condition, that is,

$$\frac{\mathrm{d}}{\mathrm{d}t}x(\cdot) = D(t)x(\cdot) + B(t)u(\cdot), \quad x(t_0) = x_0^*, \tag{70}$$

engenders two operators $Fu(\cdot) = x[u(\cdot)] = x(\cdot)$ and $F_1u(\cdot) = x[u(\cdot)]|_{t=t_1} = x(t_1)$, which assign to each control $u(\cdot) \in U$ an underlying trajectory $x(\cdot)$ and its right-hand end points (see ([27], vol. 2, p. 652) for more details). Similarly, equation (68c):

$$\frac{\mathrm{d}}{\mathrm{d}t}\psi(\cdot) + D^{\mathrm{T}}(\cdot)\psi(\cdot) = 0, \quad \psi(t_1) = \nabla\varphi_1(x_1) + A_1^{\mathrm{T}}p_1, \tag{71}$$

generates a linear operator $F_0\psi(t_1)$, that associates each terminal value of $\psi(t_1)$ at the right-hand end points of the interval $[t_0, t_1]$ with an underlying adjoint trajectory $\psi(\cdot) = F_0\psi(t_1)$. The operators $Fu(\cdot)$, $F_1u(\cdot)$, $F_0\psi(t_1)$ are linear and unequivocal; this is attributed to the linearity of differential equations (70) and

(71), and uniqueness of its solutions with assigned initial and terminal condition, respectively. Indeed, the linearity implies that for

$$\frac{\mathrm{d}}{\mathrm{d}t}\alpha x^{'}(\cdot) = D(t)\alpha x^{'}(\cdot) + B(t)\alpha u^{'}(\cdot), \quad x^{'}(t_0) = 0,$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\beta x^{''}(\cdot) = D(t)\beta x^{''}(\cdot) + B(t)\beta u^{''}(\cdot), \quad x^{''}(t_0) = 0,$$

we have

$$\frac{\mathrm{d}}{\mathrm{d}t}(\alpha x^{'}(\cdot) + \beta x^{''}(\cdot)) = D(t)(\alpha x^{'}(\cdot) + \beta x^{''}(\cdot)) + B(t)(\alpha x^{'}(\cdot) + \beta x^{''}(\cdot)),$$

where $\alpha x^{'}(t_0) + \beta x^{''}(t_0) = 0$. Hence, $F(\alpha u^{'}(\cdot) + \beta u^{''}(\cdot)) = \alpha F u^{'}(\cdot) + \beta F u^{''}(\cdot)$ and $u^{'}(\cdot), u^{''}(\cdot) \in U \subset PC([t_0, t_1])$; in particular, for $t = t_1$ we have $F(\alpha u^{'}(t_1) + \beta u^{''}(t_1)) = \alpha F u^{'}(t_1) + \beta F u^{''}(t_1)$. The latter implies that the operators $Fu(\cdot)$, $F_1u(\cdot)$ are linear, and their images are convex sets. The same results hold for the operator $F_0\psi(t_1)$ under similar justification.

2. The operators $Fu(\cdot)$ and $F_1u(\cdot)$ are bounded (see more details in ([27], vol. 2, p. 653)). Let us show this for the operator $F_1u(\cdot)$. Indeed, from (7), we have

$$|x[u(t)]| = \left|\int_{t_0}^{\tau} (D(\tau)x[u(\tau)] + B(\tau)u(\tau))\mathrm{d}\tau\right|$$
$$\leq D_{\max}\int_{t_0}^{\tau} \left|x[u(\tau)]\right|\mathrm{d}\tau + B_{\max}\int_{t_0}^{\tau} \left|u(\tau)\right|\mathrm{d}\tau,$$

where $D_{\max} = \|D(t)\|_{L_\infty}$, $B_{\max} = \|B(t)\|_{L_\infty}$. Hence, using the Gronwall lemma (see, e.g., ([27], vol. 2, p. 653)) together with Cauchy–Bunyakovsky–Schwarz inequality, we obtain:

$$|x[u(t)]| \leq e^{D_{\max}\cdot t_1}\cdot B_{\max}\int_{t_0}^{\tau} \left|u(\tau)\right|\mathrm{d}\tau \leq K_0\left(\int_{t_0}^{\tau} |u(\tau)|^2\mathrm{d}\tau\right)^{1/2}$$

for all $u(t) \in U$, $t \in [t_0, t_1]$. Here, $K_0 = e^{D_{\max}\cdot t_1}B_{\max}\sqrt{t_1}$. The last estimate can be written as:

$$\left|x[u(\cdot)]\right| = \left\|Fu(\cdot)\right\| \leq \|F\|\left|u(\cdot)\right|. \tag{72}$$

In particular, this estimate holds for $t = t_1$ and yields:

$$\left|x[u(t)]|_{t=t_1}\right| = \|F_1u\| \leq \|F_1\||u|, \tag{73}$$

where $\|F\| \leq K_0$, $\|F_1\| \leq K_0$. The boundedness of the set U implies the boundedness of image sets $F(U)$, $F_1(U)$, and hence the boundedness of the operators F, F_1. From the estimate (73), it follows that the set of attainability $X(t_1)$ should also be bounded. Let us recall that, for linear operators, their boundedness always implies continuity, that is, if $|u^n(\cdot) - u^{'}(\cdot)| \to 0$, then $\|Fu^n(\cdot) - Fu^{'}(\cdot)\| \to 0$

when $n \to \infty$. In this case, one speaks of *strong convergence* and *strong continuity* of $Fu(\cdot)$, i.e., convergence and continuity in the norm of Hilbert space. All said above equally applies to the operator $F_1 u(\cdot)$, where its convergence and continuity are understood in the sense of finite-dimensional Euclidean space.
We can prove the boundedness of linear operator $F_0\psi(t_1)$ in a similar way. On the other hand, its boundedness can be also proved by applying the theorems on stability of linear ODE solutions with regards to perturbations in their initial conditions.

3. The differential equation (70) can be written in terms of the operator $Fu(\cdot)$ as:

$$x(\cdot) = Fu(\cdot), \tag{74}$$

for all $u(\cdot) \in U$. Let us consider the behavior of Eq. (74) on *weakly convergent* sequences (see formal definition of weak convergence in Hilbert spaces in ([15], p. 114) or ([18], p. 208) among other books). Before doing so, it will be helpful to recall the formal definition of a weakly convergent sequence. It is said that a sequence of elements of $u^k(t) \in U$ *converges weakly* to an element $u'(t) \in U$, if the sequence of linear functionals $\int_{t_0}^{t_1} \langle u^k(t), c(\cdot)\rangle \mathrm{d}t$ converges pointwise to a linear functional $\int_{t_0}^{t_1} \langle u'(t), c(\cdot)\rangle \mathrm{d}t$ for any $c(\cdot) \in U$.
Suppose that the sequence $u^k(t) \in U \subset L_2^r[t_0, t_1]$ converges weakly to an element $u'(t) \in U$. It can be shown that the image of this sequence under the mapping F weakly converges to $Fu'(t)$. The latter is done by considering a logical chain (for details, see ([27], vol. 2, p. 651)):

$$\lim_{k\to\infty} \int_{t_0}^{t_1} \left\langle Fu^k(t), c(\cdot)\right\rangle \mathrm{d}t = \lim_{k\to\infty} \int_{t_0}^{t_1} \left\langle F^{\mathrm{T}} c(\cdot), u^k(t)\right\rangle \mathrm{d}t$$
$$= \int_{t_0}^{t_1} \left\langle F^{\mathrm{T}} c(\cdot), u'(t)\right\rangle \mathrm{d}t = \int_{t_0}^{t_1} \left\langle Fu'(t), c(\cdot)\right\rangle \mathrm{d}t.$$

Last formula implies that $Fu^k(t)$ converges weakly to $Fu'(t)$ when $k \to \infty$. In other words, $Fu(\cdot)$ is a weakly continuous mapping in Hilbert space $L_2^r[t_0, t_1]$. Its strong continuity was demonstrated above.
The class of weakly continuous functions is significantly narrower than the class of continuous functions. Actually, the first class of functions can be slightly extended to the class of *weakly lower semicontinuous functions*[2], since the latter class of functions is especially useful for optimization problems, because within it one can guarantee the existence of minima of convex optimization problems on weakly compact sets. It is worthwhile to note that a convex lower semicontinuous function on a convex set is weakly lower semicontinuous. In particular, the quadratic function $f(x) = |x|^2, x \in X$ is also weakly lower semicontinuous on the convex set [27].

[2] Formal definition can be consulted in [16, p. 209] or other similar textbooks.

4. Using the regularizing properties of quadratic functions, we can strengthen the definition of the minimum for regularized functions of the form $f(z) = \frac{1}{2}|z - x|^2 + \alpha\varphi(z), z \in X, \alpha > 0$ where $\varphi(z)$ is a convex function. Let x_α be a point of minimum of $f(z), z \in X$ for any $\alpha > 0$. It is known (see, e.g., [14, 26]) that a continuous convex function is subdifferentiable and, therefore, its subdifferential $\partial f(z)$ at the minimum point contains a positive subgradient as a necessary and sufficient condition for the minimum:

$$\langle x_\alpha - x + \alpha\nabla\varphi(x_\alpha), z - x_\alpha\rangle \geq 0, \quad x \in X. \tag{75}$$

Let us introduce an obvious identity:

$$\frac{1}{2}|z - x|^2 = \frac{1}{2}|z - x_\alpha|^2 + \langle z - x_\alpha, x_\alpha - x\rangle + \frac{1}{2}|x_\alpha - x|^2$$

together with convexity condition:

$$\varphi(z) \geq \varphi(x_\alpha) + \langle\nabla\varphi(x_\alpha), z - x_\alpha\rangle, \quad x \in X$$

and then sum up these two expressions taking into account the inequality (75) in order to arrive to

$$\frac{1}{2}|x_\alpha - x|^2 + \alpha\varphi(x_\alpha) \leq \frac{1}{2}|z - x|^2 + \alpha\varphi(z) - \frac{1}{2}|z - x_\alpha|^2. \tag{76}$$

This inequality [3, 27] enhances the usual definition of the minimum of $f(z)$ for $z \in X$.

8 Proof of Convergence

In this section, we are going to prove that numerical process (57)–(59) converges monotonically in norm to solution of the original problem (50)–(52) with respect to controls, state trajectories, adjoint functions, and terminal variables.

Theorem 2 *If the set of solutions of (60) is not empty and belongs to the space* $\mathbb{R}^m \times \mathbb{R}^n \times \Psi_2^n[t_0, t_1]^\star \times AC^n[t_0, t_1] \times U$, *and the terminal function* $\varphi_1(x_1)$ *is convex and differentiable, then the sequence* $(p_1^n, x_1^n, \psi^n(\cdot), x^n(\cdot), u^n(\cdot))$, *generated by (57)–(59) for any value of* $k > 0$ *converges weakly to the solution monotonically decreasing in the norm.*

Proof By setting $x(\cdot) = x^*(\cdot), u(\cdot) = u^*(\cdot), \psi(\cdot) = \psi^*(\cdot)$ in the variational inequality (63) we have

$$\left\langle x_1^{n+1} - x_1^n, x_1^* - x_1^{n+1}\right\rangle + \int_{t_0}^{t_1}\left\langle x^{n+1}(t) - x^n(t), x^*(t) - x^{n+1}(t)\right\rangle \mathrm{d}t$$
$$+ k\left\langle\nabla\varphi_1\left(x_1^{n+1}\right), x_1^* - x_1^{n+1}\right\rangle + k\left\langle A_1^{\mathrm{T}}\left(p_1^n + k\left(A_1 x_1^{n+1} + a_1\right)\right), x_1^* - x_1^{n+1}\right\rangle$$

$$
\begin{aligned}
&+k\int_{t_0}^{t_1}\left\langle D^{\mathrm{T}}(t)\left(\psi^n(t)+k\left(D(t)x^{n+1}(t)+B(t)u^{n+1}(t)-\frac{\mathrm{d}}{\mathrm{d}t}x^{n+1}(t)\right)\right),x^*(t)-x^{n+1}(t)\right\rangle\mathrm{d}t\\
&+k\int_{t_0}^{t_1}\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\left(\psi^n(t)+k\left(D(t)x^{n+1}(t)+B(t)u^{n+1}(t)-\frac{\mathrm{d}}{\mathrm{d}t}x^{n+1}(t)\right)\right),x^*(t)-x^{n+1}(t)\right\rangle\mathrm{d}t\\
&-k\Bigg(\left\langle\psi^n(t_1)+k\left(D(t)x^{n+1}(t_1)+B(t)u^{n+1}(t_1)-\frac{\mathrm{d}x^{n+1}}{\mathrm{d}t}(t_1)\right),x^*(t_1)-x^{n+1}(t_1)\right\rangle\\
&-\left\langle\psi^n(t_0)+k\left(D(t)x^{n+1}(t_0)+B(t)u^{n+1}(t_0)-\frac{\mathrm{d}x^{n+1}}{\mathrm{d}t}(t_0)\right),x^*(t_0)-x^{n+1}(t_0)\right\rangle\Bigg)\\
&+k\int_{t_0}^{t_1}\left\langle B^{\mathrm{T}}(t)\left(\psi^n(t)+k\left(D(t)x^{n+1}(t)+B(t)u^{n+1}(t)-\frac{\mathrm{d}}{\mathrm{d}t}x^{n+1}(t)\right)\right),u^*(t)-u^{n+1}(t)\right\rangle\mathrm{d}t\\
&+\int_{t_0}^{t_1}\langle u^{n+1}(t)-u^n(t),u^*(t)-u^{n+1}(t)\rangle\mathrm{d}t\geq 0.
\end{aligned}
\tag{77}
$$

Using the transition formulas to conjugate linear operators (60) and (61), the inequality (77) can be transformed to the form:

$$
\begin{aligned}
&\left\langle x_1^{n+1}-x_1^n,x_1^*-x_1^{n+1}\right\rangle+\int_{t_0}^{t_1}\left\langle x^{n+1}(t)-x^n(t),x^*(t)-x^{n+1}(t)\right\rangle\mathrm{d}t\\
&+k\left\langle\nabla\varphi_1(x_1^{n+1}),x_1^*-x_1^{n+1}\right\rangle+k\left\langle A_1^{\mathrm{T}}(p_1^n+k\left(A_1x_1^{n+1}+a_1\right)),x_1^*-x_1^{n+1}\right\rangle\\
&+k\int_{t_0}^{t_1}\left\langle\psi^n(t)+k\left(D(t)x^{n+1}(t)+B(t)u^{n+1}(t)-\frac{\mathrm{d}}{\mathrm{d}t}x^{n+1}(t)\right),D(t)(x^*(t)-x^{n+1}(t))\right\rangle\mathrm{d}t\\
&+k\int_{t_0}^{t_1}\left\langle\psi^n(t)+k\left(D(t)x^{n+1}(t)+B(t)u^{n+1}(t)-\frac{\mathrm{d}}{\mathrm{d}t}x^{n+1}(t)\right),\frac{\mathrm{d}}{\mathrm{d}t}(x^*(t)-x^{n+1}(t))\right\rangle\mathrm{d}t\\
&-k\Bigg(\left\langle\psi^n(t_1)+k\left(D(t)x^{n+1}(t_1)+B(t)u^{n+1}(t_1)-\frac{\mathrm{d}x^{n+1}}{\mathrm{d}t}(t_1)\right),x^*(t_1)-x^{n+1}(t_1)\right\rangle\\
&-\left\langle\psi^n(t_0)+k\left(D(t)x^{n+1}(t_0)+B(t)u^{n+1}(t_0)-\frac{\mathrm{d}x^{n+1}}{\mathrm{d}t}(t_0)\right),x^*(t_0)-x^{n+1}(t_0)\right\rangle\Bigg)\\
&+k\int_{t_0}^{t_1}\left\langle\psi^n(t)+k\left(D(t)x^{n+1}(t)+B(t)u^{n+1}(t)-\frac{\mathrm{d}}{\mathrm{d}t}x^{n+1}(t)\right),B(t)(u^*(t)-u^{n+1}(t))\right\rangle\mathrm{d}t\\
&+\int_{t_0}^{t_1}\langle u^{n+1}(t)-u^n(t),u^*(t)-u^{n+1}(t)\rangle\mathrm{d}t\geq 0.
\end{aligned}
\tag{78}
$$

Using (58) and (59), and taking into account the convexity of terminal functions, we obtain:

$$\begin{aligned}
&\langle x_1^{n+1} - x_1^n, x_1^* - x_1^{n+1}\rangle + \int_{t_0}^{t_1} \langle x^{n+1}(t) - x^n(t), x^*(t) - x^{n+1}(t)\rangle \mathrm{d}t \\
&+ k\left(\varphi_1(x_1^*) - \varphi_1(x_1^{n+1})\right) - k\left\langle A_1^{\mathrm{T}} p_1^{n+1}, x_1^* + x_1^{n+1}\right\rangle \\
&+ k\int_{t_0}^{t_1} \left\langle \psi^{n+1}(t), D(t)\big(x^*(t) - x^{n+1}(t)\big) + B(t)(u^*(t) - u^{n+1}(t)) - \frac{\mathrm{d}}{\mathrm{d}t}(x^*(t) - x^{n+1}(t))\right\rangle \mathrm{d}t \\
&+ \int_{t_0}^{t_1} \langle u^{n+1}(t) - u^n(t), u^*(t) - u^{n+1}(t)\rangle \mathrm{d}t \geq 0.
\end{aligned} \tag{79}$$

From the right-hand side of the inequality (40) for $x_1 = x_1^{n+1}$, $x(\cdot) = x^{n+1}(\cdot)$, $u(\cdot) = u^{n+1}(\cdot)$, we have

$$\begin{aligned}
&k\big(\varphi_1(x_1^{n+1}) - \varphi_1(x_1^*) + \langle p_1^*, A_1(x_1^{n+1} + x_1^*)\rangle\big) \\
&+ k\int_{t_0}^{t_1} \Big\langle \psi^*(t), D(t)(x^{n+1}(t) - x^*(t)) \\
&+ B(t)(u^{n+1}(t) - u^*(t)) - \frac{\mathrm{d}}{\mathrm{d}t}(x^{n+1}(t) - x^*(t))\Big\rangle \mathrm{d}t \geq 0.
\end{aligned} \tag{80}$$

Summing up the inequalities (79) and (80) we arrive to the following one:

$$\begin{aligned}
&\langle x_1^{n+1} - x_1^n, x_1^* - x_1^{n+1}\rangle + k\langle p_1^{n+1} - p_1^*, A_1(x_1^* + x_1^{n+1})\rangle \\
&+ \int_{t_0}^{t_1} \langle x^{n+1}(t) - x^n(t), x^*(t) - x^{n+1}(t)\rangle \mathrm{d}t + \int_{t_0}^{t_1} \langle u^{n+1}(t) - u^n(t), u^*(t) - u^{n+1}(t)\rangle \mathrm{d}t \\
&+ \int_{t_0}^{t_1} \langle \psi^{n+1}(t) - \psi^*(t), D(t)(x^*(t) - x^{n+1}(t)) + B(t)(u^*(t) - u^{n+1}(t)) \\
&- \frac{\mathrm{d}}{\mathrm{d}t}(x^*(t) - x^{n+1}(t))\rangle \mathrm{d}t \geq 0.
\end{aligned} \tag{81}$$

From (58) and (59), we have

$$\begin{aligned}
&k\langle p_1^{n+1} - p_1^n - kA_1\left(x_1^{n+1} + x_1^*\right), p_1^* - p_1^{n+1}\rangle \\
&+ k\int_{t_0}^{t_1} \Big\langle \psi^{n+1}(t) - \psi^n(t) - k\Big(D(t)(x^{n+1}(t) - x^*(t)) + B(t)(u^{n+1} - u^*(t)) \\
&- \frac{\mathrm{d}}{\mathrm{d}t}(x^{n+1}(t) - x^*(t))\Big), \psi^*(t) - \psi^{n+1}(t))\Big\rangle \mathrm{d}t \geq 0.
\end{aligned} \tag{82}$$

Now, we can sum up (81) and (82) to obtain:

$$\langle x_1^{n+1} - x_1^n, x_1^* - x_1^{n+1}\rangle + k\langle p_1^{n+1} - p_1^n, p_1^* - p_1^{n+1}\rangle$$

$$+\int_{t_0}^{t_1}\langle x^{n+1}(t)-x^n(t),x^*(t)-x^{n+1}(t)\rangle\mathrm{d}t+\int_{t_0}^{t_1}\langle u^{n+1}(t)-u^n(t),u^*(t)-u^{n+1}(t)\rangle\mathrm{d}t$$
$$+\int_{t_0}^{t_1}\langle \psi^{n+1}(t)-\psi^n(t),\psi^*(t)-\psi^{n+1}(t)\rangle\mathrm{d}t\geq 0. \tag{83}$$

Using the identity

$$|y_1-y_2|^2=|y_1-y_3|^2+2\langle y_1-y_3,y_3-y_2\rangle+|y_3-y_2|^2, \tag{84}$$

the scalar products can be expanded into the sum of the squares:

$$\begin{aligned}&|x_1^{n+1}-x_1^*|^2+|x_1^{n+1}-x_1^n|^2+|p_1^{n+1}-p_1^*|^2+|p_1^{n+1}-p_1^n|^2\\&+\int_{t_0}^{t_1}|x^{n+1}(t)-x^*(t)|^2\mathrm{d}t+\int_{t_0}^{t_1}|x^{n+1}(t)-x^n(t)|^2\mathrm{d}t\\&+\int_{t_0}^{t_1}|u^{n+1}(t)-u^*(t)|^2\mathrm{d}t+\int_{t_0}^{t_1}|u^{n+1}(t)-u^n(t)|^2\mathrm{d}t\\&+\int_{t_0}^{t_1}|\psi^{n+1}(t)-\psi^*(t)|^2\mathrm{d}t+\int_{t_0}^{t_1}|\psi^{n+1}(t)-\psi^n(t)|^2\mathrm{d}t\leq|x_1^n-x_1^*|^2+|p_1^n-p_1^*|^2\\&+\int_{t_0}^{t_1}|x^n(t)-x^*(t)|^2\mathrm{d}t+\int_{t_0}^{t_1}|u^n(t)-u^*(t)|^2\mathrm{d}t+\int_{t_0}^{t_1}|\psi^n(t)-\psi^*(t)|^2\mathrm{d}t.\end{aligned} \tag{85}$$

If the second, fourth, sixth, eighth, and tenth terms in the left-hand side of this inequality are discarded, then we obtain the property of monotonically decreasing sequence. Geometrically, this would mean that the ball of $(n+1)$st iteration is embedded in the ball nth iteration.

Summing up the inequalities of (85) with respect to n that runs from $n=0$ to $n=N$, we obtain:

$$\begin{aligned}&\left|x_1^{N+1}-x_1^*\right|^2+\sum_{n=0}^{N}\left|x_1^{n+1}-x_1^n\right|^2+\left|p_1^{N+1}-p_1^*\right|^2+\sum_{n=0}^{N}\left|p_1^{n+1}-p_1^n\right|^2\\&+\int_{t_0}^{t_1}\left|x^{N+1}(t)-x^*(t)\right|^2\mathrm{d}t+\sum_{n=0}^{N}\int_{t_0}^{t_1}\left|x^{n+1}(t)-x^n(t)\right|^2\mathrm{d}t\\&+\int_{t_0}^{t_1}\left|u^{N+1}(t)-u^*(t)\right|^2\mathrm{d}t+\sum_{n=0}^{N}\int_{t_0}^{t_1}\left|u^{n+1}(t)-u^n(t)\right|^2\mathrm{d}t\\&+\int_{t_0}^{t_1}\left|\psi^{N+1}(t)-\psi^*(t)\right|^2\mathrm{d}t+\sum_{n=0}^{N}\int_{t_0}^{t_1}\left|\psi^{n+1}(t)-\psi^n(t)\right|^2\mathrm{d}t\\&\leq\left|x_1^0-x_1^*\right|^2+\left|p_1^0-p_1^*\right|^2++\int_{t_0}^{t_1}\left|x^0(t)-x^*(t)\right|^2\mathrm{d}t\end{aligned}$$

$$+\int_{t_0}^{t_1}\left|u^0(t)-u^*(t)\right|^2\mathrm{d}t+\int_{t_0}^{t_1}\left|\psi^0(t)-\psi^*(t)\right|^2\mathrm{d}t.$$

This inequality implies the boundedness of the approximating sequence with respect to direct and dual terminal variables, state trajectories, control functions, and adjoint functions:

$$\left|x_1^{N+1}-x_1^*\right|^2+\left|p_1^{N+1}-p_1^*\right|^2+\int_{t_0}^{t_1}\left|x^{N+1}(t)-x^*(t)\right|^2\mathrm{d}t+\int_{t_0}^{t_1}\left|u^{N+1}(t)-u^*(t)\right|^2\mathrm{d}t$$
$$+\int_{t_0}^{t_1}\left|\psi^{N+1}(t)-\psi^*(t)\right|^2\mathrm{d}t\le\left|x_1^0-x_1^*\right|^2+\left|p_1^0-p_1^*\right|^2+\left|x^0(t)-x^*(t)\right|^2\mathrm{d}t$$
$$+\int_{t_0}^{t_1}\left|u^0(t)-u^*(t)\right|^2\mathrm{d}t+\int_{t_0}^{t_1}\left|\psi^0(t)-\psi^*(t)\right|^2\mathrm{d}t, \tag{86}$$

as well as convergence of the series:

$$\sum_{n=0}^{N}\left|x_1^{n+1}-x_1^n\right|^2<\infty,\quad \sum_{n=0}^{N}\left|p_1^{n+1}-p_1^n\right|^2<\infty,\quad \sum_{n=0}^{\infty}\int_{t_0}^{t_1}\left|x^{n+1}(t)-x^n(t)\right|^2\mathrm{d}t<\infty,$$
$$\sum_{n=0}^{\infty}\int_{t_0}^{t_1}\left|u^{n+1}(t)-u^n(t)\right|^2\mathrm{d}t<\infty,\quad \sum_{n=0}^{N}\int_{t_0}^{t_1}\left|\psi^{n+1}(t)-\psi^n(t)\right|^2\mathrm{d}t<\infty$$

and tendency to zero of the quantities:

$$\left|x_1^{n+1}-x_1^n\right|^2\to 0,\quad \left|P_1^{n+1}-p_1^n\right|^2\to 0,\quad \int_{t_0}^{t_1}\left|x^{n+1}(t)-x^n(t)\right|^2\mathrm{d}t\to 0,$$
$$\int_{t_0}^{t_1}\left|u^{n+1}(t)-u^n(t)\right|^2\mathrm{d}t\to 0,\quad \int_{t_0}^{t_1}\left|\psi^{n+1}(t)-\psi^n(t)\right|^2\mathrm{d}t\to 0, \tag{87}$$

when $n\to\infty$.

Since the sequence $(x_1^n,p_1^n,x^n(\cdot),u^n(\cdot),\psi^n(\cdot))$ is bounded in $\mathbb{R}^n\times\mathbb{R}^m\times AC^n[t_0,t_1]\times U\times\Psi_2^n[t_0,t_1]^\star$, it is also *weakly compact*. According to [21], the latter means that this sequence is univocally associated with a sequence of linear functionals in dual spaces:

$$\langle x_1^n,x\rangle,\quad \langle p_1^n,p\rangle,\quad \langle x^n(\cdot),x(\cdot)\rangle,\quad \langle u^n(\cdot),u(\cdot)\rangle,\quad \langle\psi^n(\cdot),\psi(\cdot)\rangle.$$

This sequence of linear functionals has a subsequence that converges pointwise (i.e., on each element of its own space), to the set of functionals:

$$l_1(x),\quad l_2(p),\quad l_3(x(\cdot)),\quad l_4(u(\cdot)),\quad l_5(\psi(\cdot)). \tag{88}$$

It is also known that these functionals are linear and bounded. Moreover, the dual spaces of Hilbert spaces are also complete Hilbert space (see more details in [21]). Therefore, all components of the family (88) are elements of the corresponding dual spaces. According to *Riesz representation theorem* (see, e.g., ([31], Theorem 18.6)

or ([18], Theorem 4.6.4)), we can conclude that *all* functionals of the family (88) have the form:

$$\langle x_1^{'}, x\rangle, \quad \langle p_1^{'}, p\rangle, \quad \langle x^{'}(\cdot), x(\cdot)\rangle, \quad \langle u^{'}(\cdot), u(\cdot)\rangle, \quad \langle \psi^{'}(\cdot), \psi(\cdot)\rangle.$$

Thus, we have shown that the sequence (57)–(59) generates a set of components $(x_1^{'}, p_1^{'}, x^{'}(\cdot), u^{'}(\cdot), \psi^{'}(\cdot))$ that represents a *weak limit* (in the sense of pointwise convergence of linear functionals) of a subsequence:

$$\left\langle x_1^{n_i}, x\right\rangle, \quad \left\langle p_1^{n_i}, p\right\rangle, \quad \left\langle x^{n_i}(\cdot), x(\cdot)\right\rangle, \quad \left\langle u^{n_i}(\cdot), u(\cdot)\right\rangle, \quad \left\langle \psi^{n_i}(\cdot), \psi(\cdot)\right\rangle.$$

It is useful to recall (see [21] for further details), that in finite-dimensional (or Euclidean) spaces, there is no difference between weak and strong convergence (with respect to the space norm).

Now we can demonstrate that the set of components $(x_1^{'}, p_1^{'}, x^{'}(\cdot), u^{'}(\cdot), \psi^{'}(\cdot))$ is a solution of the differential system (50)–(52) which is equivalent to the problems (64) and (60).

To this end, we first present the problem (57) in the form of inequality (76) that reflects more accurately the minimum properties of the regularized strongly convex objective function. Then we revise this inequality together with the system (57)–(59) on the elements of our subsequence:

$$\frac{1}{2}\left|x_1^{n+1} - x_1^n\right|^2 + \frac{1}{2}\int_{t_0}^{t_1}\left|x^{n+1}(t) - x^n(t)\right|^2 \mathrm{d}t + \frac{1}{2}\int_{t_0}^{t_1}\left|u^{n+1}(t) - u^n(t)\right|^2 \mathrm{d}t \tag{89a}$$

$$+ k\Big[\varphi_1\left(x_1^{n+1}\right) + \frac{1}{2k}\left|p_1^n + k\left(A_1 x_1^{n+1} + a_1\right)\right|^2 - \frac{1}{2k}|p_1^n|^2 - \frac{1}{2k}\int_{t_0}^{t_1}|\psi^n(t)|^2 \mathrm{d}t \tag{89b}$$

$$+ \frac{1}{2k}\int_{t_0}^{t_1}\left|\psi^n(t) + k\left(D(t)x^{n+1}(t) + B(t)u^{n+1}(t) - \frac{\mathrm{d}}{\mathrm{d}t}x^{n+1}(t)\right)\right|^2 \mathrm{d}t\Big] \tag{89c}$$

$$\leq \frac{1}{2}\left|x_1 - x_1^n\right|^2 + \frac{1}{2}\int_{t_0}^{t_1}\left|x(t) - x^n(t)\right|^2 \mathrm{d}t + \frac{1}{2}\int_{t_0}^{t_1}\left|u(t) - u^n(t)\right|^2 \mathrm{d}t \tag{89d}$$

$$+ k\Big[\varphi_1\left(x_1\right) + \frac{1}{2k}\left|p_1^n + k\left(A_1 x_1 + a_1\right)\right|^2 - \frac{1}{2k}|p_1^n|^2 - \frac{1}{2k}\int_{t_0}^{t_1}|\psi^n(t)|^2 \mathrm{d}t \tag{89e}$$

$$+ \frac{1}{2k}\int_{t_0}^{t_1}\left|\psi^n(t) + k\left(D(t)x(t) + B(t)u(t) - \frac{\mathrm{d}}{\mathrm{d}t}x(t)\right)\right|^2 \mathrm{d}t\Big] \tag{89f}$$

$$- \frac{1}{2}\left|x_1 - x_1^{n+1}\right|^2 - \frac{1}{2}\int_{t_0}^{t_1}\left|x(t) - x^{n+1}(t)\right|^2 \mathrm{d}t - \frac{1}{2}\int_{t_0}^{t_1}\left|u(t) - u^{n+1}(t)\right|^2 \mathrm{d}t, \tag{89g}$$

$$p_1^{n+1} = p_1^n + k\left(A_1 x_1^{n+1} + a_1\right), \tag{90}$$

$$\psi^{n+1}(t) = \psi^n(t) + k\left(D(t)x^{n+1}(t) + B(t)u^{n+1}(t) - \frac{\mathrm{d}}{\mathrm{d}t}x^{n+1}(t)\right), \tag{91}$$

where $x(t_0) = x_0^*$.

In Sect. 6, it was pointed out that all finite-dimensional and bounded differential operators which appear in (89), (90), and (91) are *weakly continuous* (in particular, finite-dimensional operators are also *strongly continuous*). The latter implies that if $(x_1^{n_i}, p_1^{n_i}, x^{n_i}(\cdot), u^{n_i}(\cdot), \psi^{n_i}(\cdot)) \xrightarrow{w} (x_1^{'}, p_1^{'}, x^{'}(\cdot), u^{'}(\cdot), \psi^{'}(\cdot))$ when $n_i \to \infty$, then

$$
\begin{gathered}
-\psi^{n_i+1}(t_1) + \nabla\varphi_1(x_1^{n_i+1}) + A_1^{\mathrm{T}} p_1^{n_i+1} \longrightarrow -\psi^{'}(t_1) + \nabla\varphi_1(x_1^{'}) + A_1^{\mathrm{T}} p_1^{'}, \\
\frac{\mathrm{d}}{\mathrm{d}t}\psi^{n_i+1}(t) + D^{\mathrm{T}}(t)\psi^{n_i+1}(t) \xrightarrow{w} \frac{\mathrm{d}}{\mathrm{d}t}\psi^{'}(t) + D^{\mathrm{T}}(t)\psi^{'}(t), \\
A_1 x_1^{n_i+1} + a_1 \longrightarrow A_1 x_1^{'} + a_1, \\
D(t)x^{n_i+1}(t) + B(t)u^{n_i+1}(t) - \frac{\mathrm{d}}{\mathrm{d}t}x^{n_i+1}(t) \xrightarrow{w} D(t)x^{'}(t) + B(t)u^{'}(t) - \frac{\mathrm{d}}{\mathrm{d}t}x^{'}(t).
\end{gathered}
$$

Here the symbol $\xrightarrow{w}$ denotes weak convergence.

Let us write the inequality (89) in the form:

$$
\varphi_1\left(x_1^{n+1}\right) + \frac{1}{2k^2}\left|p_1^n + k\left(A_1 x_1^{n+1} + a_1\right)\right|^2 \tag{92a}
$$

$$
+ \frac{1}{2k^2}\int_{t_0}^{t_1}\left|\psi^n(t) + k\left(D(t)x^{n+1}(t) + B(t)u^{n+1}(t) - \frac{\mathrm{d}}{\mathrm{d}t}x^{n+1}(t)\right)\right|^2 \mathrm{d}t \tag{92b}
$$

$$
\le \varphi_1(x_1) + \frac{1}{2k^2}\left|p_1^n + k\left(A_1 x_1 + a_1\right)\right|^2 + \frac{1}{k}\left(\alpha_1^n + \alpha_2^n + \alpha_3^n\right) \tag{92c}
$$

$$
+ \frac{1}{2k^2}\int_{t_0}^{t_1}\left|\psi^n(t) + k\left(D(t)x(t) + B(t)u(t) - \frac{\mathrm{d}}{\mathrm{d}t}x(t)\right)\right|^2 \mathrm{d}t \tag{92d}
$$

where

$$
\alpha_1^n = \frac{1}{2}|x_1 - x_1^n|^2 - \frac{1}{2}|x_1 - x_1^{n+1}|^2 - \frac{1}{2}|x_1^{n+1} - x_1^n|^2,
$$

$$
\alpha_2^n = \frac{1}{2}\int_{t_0}^{t_1}|x(t) - x^n(t)|^2\mathrm{d}t - \frac{1}{2}\int_{t_0}^{t_1}|x(t) - x^{n+1}(t)|^2\mathrm{d}t - \frac{1}{2}\int_{t_0}^{t_1}|x^{n+1}(t) - x^n(t)|^2\mathrm{d}t,
$$

$$
\alpha_3^n = \frac{1}{2}\int_{t_0}^{t_1}|u(t) - u^n(t)|^2\mathrm{d}t - \frac{1}{2}\int_{t_0}^{t_1}|u(t) - u^{n+1}(t)|^2\mathrm{d}t - \frac{1}{2}\int_{t_0}^{t_1}|u^{n+1}(t) - u^n(t)|^2\mathrm{d}t.
$$

In the inequality (92), when we pass to the limit for $n \to \infty$, the quantities $\alpha_i^n \to 0, i = 1, 2, 3$ in virtue of (84), (87). Therefore, for each n, the right-hand side of (92) can be treated as a constant function, and its left-hand side will be regarded as a functional, bounded from below by the designated constant function.

Then, due to continuity of φ and linear constraint with respect to x, we have for $n \to \infty$ that

$$
\varphi_1(x_1^{n+1}) + \frac{1}{2k}\left|p_1^n + k\left(A_1 x_1^{n+1} + a_1\right)\right|^2 \to \varphi_1(x_1^{'}) + \frac{1}{2k}\left|p_1^{'} + k\left(A_1 x_1^{'} + a_1\right)\right|^2.
$$

On the other hand, in virtue of *weak lower semicontinuity* (see formal definition in, e.g., ([16], p. 209)) of quadratic functions, that is,

$$\lim_{n\to\infty}\frac{1}{2k}\int_{t_0}^{t_1}\left|\psi^n(t)+k\left(D(t)x^{n+1}(t)+B(t)u^{n+1}(t)-\frac{\mathrm{d}}{\mathrm{d}t}x^{n+1}(t)\right)\right|^2\mathrm{d}t$$
$$\geq\frac{1}{2k}\int_{t_0}^{t_1}\left|\psi^{'}(t)+k\left(D(t)x^{'}(t)+B(t)u^{'}(t)-\frac{\mathrm{d}}{\mathrm{d}t}x^{'}(t)\right)\right|^2\mathrm{d}t$$

the functional in (92) at the point $\psi^{'}(\cdot),x^{'},u^{'}$ may take a value not greater than

$$\frac{1}{2k}\int_{t_0}^{t_1}\left|\psi^{'}(t)+k\left(D(t)x^{'}(t)+B(t)u^{'}(t)-\frac{\mathrm{d}}{\mathrm{d}t}x^{'}(t)\right)\right|^2\mathrm{d}t.$$

Assuming that our functional takes exactly this value, the inequality (92) for $n\to\infty$ becomes

$$\varphi_1\left(x_1^{'}\right)+\frac{1}{2k}\left|p_1^{'}+k\left(A_1x_1^{'}+a_1\right)\right|^2 \tag{93a}$$
$$+\frac{1}{2k}\int_{t_0}^{t_1}\left|\psi^{'}(t)+k\left(D(t)x^{'}(t)+B(t)u^{'}(t)-\frac{\mathrm{d}}{\mathrm{d}t}x^{'}(t)\right)\right|^2\mathrm{d}t \tag{93b}$$
$$\leq\varphi_1\left(x_1\right)+\frac{1}{2k}\left|p_1^{'}+k\left(A_1x_1+a_1\right)\right|^2 \tag{93c}$$
$$+\frac{1}{2k}\int_{t_0}^{t_1}\left|\psi^{'}(t)+k\left(D(t)x(t)+B(t)u(t)-\frac{\mathrm{d}}{\mathrm{d}t}x(t)\right)\right|^2\mathrm{d}t, \tag{93d}$$

for all $x_1,x(\cdot),u(\cdot)$. For other possible values of this functional at $\psi^{'}(\cdot),x^{'},u^{'}$, the left-hand side of (93) can only get decreased and, hence, the inequality will be strengthened and will remain valid.

We can complement (93) with the limits of Eqs. (90) and (91) when $n\to\infty$, that is,

$$A_1x_1^{'}+a_1=0,\quad\frac{\mathrm{d}}{\mathrm{d}t}x^{'}(t)=D(t)x^{'}(t)+B(t)u^{'}(t). \tag{94}$$

It is easy to see that the resulting system (93), (94) coincides with (50)–(52) which, in its turn, is equivalent to (64) and (60). Effectively, system (93), (94) can be reduced to

$$\frac{\mathrm{d}}{\mathrm{d}t}x^{'}(t)=D(t)x^{'}(t)+B(t)u^{'}(t),\quad x^{'}(t_1)=x_0^*, \tag{95a}$$
$$A_1x_1^{'}+a_1=0, \tag{95b}$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\psi^{'}(t)+D^{\mathrm{T}}(t)\psi^{'}(t)=0,\quad\psi^{'}(t_1)=\nabla\varphi_1(x_1^{'})+A_1^{\mathrm{T}}p_1^{'}, \tag{95c}$$
$$\int_{t_0}^{t_1}\langle B^{\mathrm{T}}(t)\psi^{'}(t),u(t)-u^{'}(t)\rangle\mathrm{d}t\geq0,\quad u(\cdot)\in U, \tag{95d}$$

by reiterating the argument applied between the formulas (64) and (60).

Comparing the above system with (60), one can observe that any weak limit point of the process (57)–(59) is a solution of the system (95). This essentially means that the components of $(x_1^{'}, p_1^{'}, x^{'}(\cdot), u^{'}(\cdot), \psi^{'}(\cdot))$ and the components of $(x_1^*, p_1^*, x^*(\cdot), u^*(\cdot), \psi^*(\cdot))$ describe the same point, which is a solution of (50)–(52), and thus the solution of the original problem (6)–(8).

In other words, we have shown that the process (57)–(59) generates a sequence that has limit points (in the sense of weak convergence). All these points are solutions of (50)–(52). Additionally, this process decreases monotonically in the space norm on the product $\mathbb{R}^n \times \mathbb{R}^m \times AC^n[t_0, t_1] \times U \times \Psi_2^n[t_0, t_1]^\star$ in the sense of inequality (85), that is, $(n+1)$-st iteration is embedded in a ball of nth iteration:

$$
\begin{aligned}
|x_1^{n+1} - x_1^*|^2 + |p_1^{n+1} - p_1^*|^2 + \int_{t_0}^{t_1} |x^{n+1}(\cdot) - x^*(\cdot)|^2 \mathrm{d}t \\
+ \int_{t_0}^{t_1} |u^{n+1}(\cdot) - u^*(\cdot)|^2 \mathrm{d}t + \int_{t_0}^{t_1} |\psi^{n+1}(\cdot) - \psi^*(\cdot)|^2 \mathrm{d}t \\
\leq |x_1^n - x_1^*|^2 + |p_1^n - p_1^*|^2 + \int_{t_0}^{t_1} |x^n(\cdot) - x^*(\cdot)|^2 \mathrm{d}t \\
+ \int_{t_0}^{t_1} |u^n(\cdot) - u^*(\cdot)|^2 \mathrm{d}t + \int_{t_0}^{t_1} |\psi^n(\cdot) - \psi^*(\cdot)|^2 \mathrm{d}t.
\end{aligned}
$$

Here, the component $\int_{t_0}^{t_1} |u^n(\cdot) - u^*(\cdot)|^2 \mathrm{d}t$ may not tend to zero as $n \to \infty$. However, we can affirm it regarding the the state component, that is, $\int_{t_0}^{t_1} |x^n(\cdot) - x^*(\cdot)|^2 \mathrm{d}t \to 0$. Additionally, weak convergence of $u^{n_i}(\cdot)$ to $u^{'}(\cdot)$ implies weak convergence of $x^{n_i}(\cdot)$ to $x^{'}(\cdot)$ when $n_i \to \infty$. According to the *Banach–Steinhaus theorem* (see, e.g., ([21], Chap. 4)) this sequence is bounded in norm and, being a sequence of functions defined on the interval $[t_0, t_1]$, is also uniformly bounded. Moreover, it can be shown that this sequence satisfies the Lipschitz condition:

$$
|x(t + \Delta t)[u(\cdot)] - x(t)[u(\cdot)]| \leq \mathcal{L} |\Delta t|^{\frac{1}{2}},
$$

where $\mathcal{L} = \text{const}$ is independent of $n, t, \Delta t$. A sequence with such a property is referred to as *equicontinuous* and, according to *Arzelà theorem*[3] (see, e.g., ([16], Theorem 2.12) or ([18], Theorem 2.7.4)) has a subsequence that converges pointwise to $x^{'}(\cdot)$ in the uniform norm $\max_{t \in [t_0, t_1]} |x^n(t) - x^*(t)|^2 \to 0$ (see more details in ([27], vol. 2, p. 659)). A fortiori, this sequence will be convergent in the space norm of $L_2^n[t_0, t_1]$. A similar assertion will be valid for the sequence $\max_{t \in [t_0, t_1]} |\psi^n(t) - \psi^*(t)|^2 \to 0$ when $n \to \infty$ due to the same reasoning.

Along with the main iterative process developing in the functional space, there is a subprocess that proceeds on the set of attainability. This subprocess is included in the formulas (57) and (58), and takes place in a finite-dimensional Euclidean space.

[3] also known as *Arzelà—Ascoli* theorem

Within the frameworks of general numerical scheme, this subprocess converges to a saddlepoint of the augmented Lagrangian:

$$\mathtt{m}(p_1, x_1) = \varphi_1(x_1) + \frac{1}{2k}|p_1 + k(A_1x_1 + a_1)|^2 - \frac{1}{2k}|p_1|^2, \tag{96}$$

and also of the traditional Lagrangian $\mathtt{l}(p_1, x_1) = \varphi_1(x_1) + \langle p_1, A_1x_1 + a_1\rangle$, that refers to a convex programming problem formulated on the set of attainability. The convergence of this subprocess to a saddlepoint of Lagrangian $\mathtt{l}(p_1, x_1)$ is understood in the sense of Euclidean norm, that is, $|x_1^n - x_1^*|^2 + |p_1^n - p_1^*|^2 \to 0$ when $n \to \infty$. Finally, the theorem is proved.

Thus, the global numerical process (57)–(59) takes place simultaneously in the functional and finite-dimensional spaces. The control functions, state and adjoint trajectories are moving in functional spaces, while a free right-hand end point of the state trajectory is being iteratively transformed in finite-dimensional space. The process has a weak limit point which is the solution of the original system. The primal and dual functional components of this limit point form a saddlepoint of the augmented Lagrangian (48), its primal and dual vector components produce a saddlepoint of finite-dimensional augmented Lagrangian (96). In the absence of dynamics in the original problem, the global process is reduced to an iterative process in finite-dimensional space.

In virtue of condition (20), the primal component of the global saddlepoint is also a maximizer of the dual function (49), and the method presented in this chapter can be viewed as a gradient method for the maximization of (49). It should be emphasized that this global process converges *weakly* to the solution of the original problem [4] and this convergence is of different nature for particular components of the process; namely, it is *weak* for the control function and *strong* in the sense of norm for other (functional) components of the process.

Additionally, all limit points of the process are solutions of the original system. It remains only to point out that the Theorem 2 is an accurate and natural generalization of the Theorem 1 (formulated and proved in Sect. 3 for a finite-dimensional case) to the optimal control problem in functional spaces.

The proposed method guarantees a monotone approximation to the solution in the space norm for all direct and dual variables. This regularizing property contributes to the process sustainability even without guaranteeing *strong* convergence to the solution with respect to *all* variables. This mainly refers to *weak* convergence of control variables. Namely, all controls are uniformly bounded (belong to a functional ball U), but they do admit some finite discontinuities. Each point of discontinuity can be surrounded by arbitrarily small interval inside which a discontinuous function can be approximated by a smooth curve within the limits of given accuracy. In this case, we have functions with arbitrarily large derivatives, which again lead us to *weak* convergence.

[4] This convergence is understood in the sense of subsequences.

To overcome these difficulties, one should impose additional conditions on the set of admissible controls U. For example, one may require that all $u(\cdot) \in U$ satisfy the Lipschitz condition (in integral form) and that all approximations $u^k(t)$ of the iterative process (57)–(59) remain inside U. In this case, the sequence of controls will be equicontinuous (in integral sense) and, in virtue of Riesz theorem, will possess the compactness property in the norm of $L_2^r[t_0, t_1]$. The latter will be sufficient to guarantee a monotone convergence of our method to a single limit point.

9 Conclusions

The mathematical model described in this chapter contains two well-known components: static and dynamic. The first component is related to optimization problem in finite-dimensional space. Generally speaking, static optimization models describe a situation of individual or collective decision making that does not vary in time. Among them one should recall the problems of linear, convex, and equilibrium programming, n-person games with Nash equilibrium, problems of multi-objective (or vector) optimization, problems of economic equilibrium, etc.

The second component has time-varying nature described by a linear ODE system and should be associated with dynamic optimization and optimal control. In particular, optimal control theory explores the possibility of transferring the dynamic system from one state to another under an external effect of control functions. Based on this concept, there are various methods for design of optimal control strategies that may depend on the initial or current state of the dynamic system (program and feedback controls). Though, many of them are only applicable to rather narrow classes of problems.

In this chapter, we have proposed an integrated approach for the solution of optimal control problems, whose boundary states are described by finite-dimensional models of convex optimization. In other words, our integrated model anticipates an adjustment of (finite-dimensional) decision making to possible changes over time in future decision-making situations. The latter was done by exploiting the saddlepoint conditions and methods, where the feedback to current states was described in terms of dual (finite and functional) variables.

References

1. Antipin, A.S.: A method for finding a saddle point of a modified Lagrange function. Èkonom. i Mat. Metody **13**(3), 560–565 (1977)
2. Antipin, A.S.: Extrapolation methods of computing the saddle point of a lagrange function and application to problems with block-separable structure. USSR Comput. Math. Math. Phys. **26**(1), 96 (1986)
3. Antipin, A.S.: Methods for solving systems of convex programming problems. Comput. Math. Math. Phys. **27**(3), 368–376 (1987)

4. Antipin, A.S.: Equilibrium programming: Gradient-type methods. Autom. Remote Control **58**(8), 1337–1347 (1997)
5. Antipin, A.S.: Equilibrium programming problems: Prox-regularization and prox-methods. Recent Advances in Optimization, Trier, 1996. Lecture Notes in Economics and Mathematical Systems, vol. 452, pp. 1–18. Springer, Berlin (1997)
6. Antipin, A.S.: Equilibrium programming: proximal methods. Comput. Math. Math. Phys. **37**(11), 1285–1296 (1997)
7. Antipin, A.S.: Extra-proximal methods for solving two-person nonzero-sum games. Math. Program. **120**(1, Ser. B), 147–177 (2009)
8. Antipin, A.S.: Method of modified Lagrangian for optimal control problems with free terminal end-point. Izvestiya IGU, Ser. Math. **4**(2), 27–44 (2011)
9. Antipin, A.S.: Two-person game with Nash equilibrium in optimal control problems. Optimiz. Lett. **6**, 1349–1378 (2012)
10. Galeev, E.M., Zelikin, M.I., Konyagin, S.V., Magaril-Ilyaev, G.G., Osmolovskii, N.P., Protasov, V.Yu. , Tikhomirov, V.M., Fursikov, A.V.: Optimal'noe upravlenie [Optimal control]. Moscow Center for Continuous Mathematical Education (MCCME), Moscow, Russia, 2008
11. Golshtein, E.G., Tretyakov, N.V.: Modified Lagrangians and monotone maps in optimization. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley, New York (1996). (Translated from the 1989 Russian original by Tretyakov, A Wiley-Interscience Publication)
12. Hager, W.W.. Multiplier methods for nonlinear optimal control. SIAM J. Numer. Anal. **27**(4), 1061–1080 (1990)
13. Hestenes, M.R.. Multiplier and gradient methods. J. Optim. Theory Appl. **4**, 303–320 (1969)
14. Hiriart-Urruty, J.-B., Lemaréchal, C.: Fundamentals of Convex Analysis, Grundlehren text editions. Springer, Berlin (2001)
15. Hirsch, F., Lacombe, F.H.G.. Elements of functional analysis. Graduate Texts in Mathematics Series. Springer, New York (1999)
16. Hunter, J., Nachtergaele, B.: Applied Analysis. World Scientific, Singapore (2001)
17. Ioffe, A.D., Tikhomirov, V.M.: Theory of Extremal Problems. North-Holland Publishing Company, Amsterdam (1979)
18. Kolmogorov, A.N., Fomin, S.V.: Elementy teorii funktsii i funktsionalnogo analiza [Elements of the Theory of Functions and Functional Analysis], 6th edn. Nauka, Moscow (1989)
19. Korpelevich, G.M.: An extragradient method for finding saddle points and for other problems. Èkonom. i Mat. Metody **12**(4), 747–756 (1976)
20. Luenberger, D.G.: Optimization by vector space methods. Series in Decision and Control, 2nd edn. Wiley, New York (1997)
21. Lusternik, L.A., Sobolev, V.I.: Elements of functional analysis. International Monographs on Advanced Mathematics and Physics. Hindustan Publishing Corp., Delhi (corrected edition) (1971)
22. Polyak, B.T.: Introduction to optimization. Translations series in Mathematics and Engineering. Optimization Software Inc. Publications Division, New York (1987)
23. Powell, M.J.D.: A method for nonlinear constraints in minimization problems. In: Fletcher, R. (ed.) Optimization (Sympos., Univ. Keele, Keele, 1968), pp. 283–298. Academic, London (1969)
24. Rockafellar, R.T.: The multiplier method of Hestenes and Powell applied to convex programming. J. Optimization Theory Appl. **12**, 555–562 (1973)
25. Rockafellar, R.T.: Augmented Lagrangians and applications of the proximal point algorithm in convex programming. Math. Oper. Res. **1**(2), 97–116 (1976)
26. Rockafellar, R.T., Wets, R. J.-B.: Variational analysis. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 317. Springer, Berlin (1998)
27. Vasil'ev, F.P.: Metody optimizatsii [Optimization Methods], vols. 1, 2. Moscow Center for Continuous Mathematical Education (MCCME), Moscow (2011)

28. Vasilieva, O.: The search of equilibrium strategies for controlled boundary value problem. Asian J Control **3**(1), 50–56 (2001)
29. Vasilieva, O.: Search of equilibrium controls in differential game with boundary conditions. Izvestiya IGU, Ser. Math. **1**(1):70–85 (2007)
30. Vasilieva, O., Vasil'ev, O.V.. On the search for equilibrium controls in an m-person differential game. Russ. Math. **44**(12), 7–12 (2000)
31. Yeh, J.: Real Analysis: Theory of Measure and Integration, 2nd edn. World Scientific, Singapore (2006)

Minimizing Sign Changes Rowwise: Consecutive Ones Property and Beyond

Dominique Fortin and Ider Tseveendorj

Abstract A 0–1 matrix where in each row the 1s occur consecutively is said to have the consecutive 1s property. Since this property is scarcely fulfilled in real problems and since it is non-deterministic polynomial time (NP)-hard to find the *nearest* arrangement to the property, we give a quadratic assignment formulation for optimizing the *distance* to the property. The formulation carries over the sign case with $0, +1, -1$ matrix entries. We discuss and compare this exact approach, for both signed and unsigned cases, with spectral approaches based on bisection instead.

Keywords 0–1 matrices · Consecutive 1s property · Consecutive sign property · Trigraph · QAP · Hoffman–Wielandt · Gilmore–Lawler

1 Introduction

A 0–1 matrix where in, say each row, the 1s occur consecutively is said to have the consecutive 1s property (C1P). This property and its approximations have numerous applications in clustering, seriation, and at a low-level data storage for compacting sparse matrices. Testing this property holds in polynomial time; however, it scarcely happens for real-life cases, thus methods that approximate, in some sense, the property have been proposed: upto an approximation factor [6], with respect to spectral ordering [1, 18, 10], number and distance of consecutive intervals of 1s [16], within an ordered set of matrices [17], etc. Most approaches lead quickly to non-deterministic polynomial time (NP)-hard optimization problems. In this chapter, we give a raw formulation leading to a quadratic assignment problem (QAP) an NP-hard problem too; unlike previous spectral approximations that relate the C1P

D. Fortin (✉)
INRIA, Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex, France
e-mail: Dominique.Fortin@inria.fr

I. Tseveendorj
Laboratoire PRiSM, UMR 8144, Université de Versailles 45, avenue des États-Unis 78035 Versailles Cedex, France
e-mail: Ider.Tseveendorj@prism.uvsq.fr

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_2

optimization problem to bisection (through the Laplacian of underlied correlation and its second (Fiedler) eigenvector to force convexity), the QAP formulation is exact and does not enforce convexity.

The QAP approach allows a straigthforward generalization to consecutive $-1, 0, +1$ entries (C±P for consecutive sign property), no matter the ordering among consecutive values. It brings up new connections with the sign change counting function [12], as a signed generalization of the counting function [14, 20], both playing a prominent role for improving a local solution in global optimization.

Since there is no methodological difference with maximizing the number of transitions instead, it relates our study to studies, to name a few, in mathematical physics: the six vertex model under boundary wall condition and the alternating sign matrices [9, 22]; in graph theory: a *trigraph* [2, 7, 8] has a signed vertex–vertex adjacency matrix. Bearing in mind the consecutive sign property, we mainly follow the lead of minimization, despite it deserves studying the six vertex model or the trigraph case, under further constraints on the transitions within a row.

The chapter is organized as follows: Sect. 2 provides the basic formulation for minimizing the number of transitions between 1s and 0s along either dimension of matrices; in Sect. 3, we review bounding schemes for an enumerative approach in both actual and spectral domains; in Sect. 4, we suggest a way to deal with the signed case by recoursing to the average in spite of Schur convexity. The remaining sections provide a thorough discussion of NP-hardness (Sect. 5), of experiments for moderate-size matrices either signed or unsigned (Sect. 6) and of the comparison between the circular and the standard shift case (Sect. 7).

2 Consecutive Ones Approximation

2.1 Minimizing Transitions

For a sparse $n \times m$ matrix A, let us consider the problem of minimizing the number of transitions between valid and void entries in a per row basis; values do not matter so that we assume the matrix binary with a 1 for valid entry and 0 otherwise. Denote the identity matrix I and the circular shift by one column matrix S_c, using Toeplitz matrix:

$$I - S_c = \begin{bmatrix} 1 & 0 & \dots & -1 \\ -1 & 1 & 0 & \dots \\ & & \dots & \\ 0 & \dots & -1 & 1 \end{bmatrix},$$

then the total number of circular column transitions is $\|A(I - S_c)\|^2$ using standard dot product for matrices, namely, $(A, B) = (\text{Vect}\,(A), \text{Vect}\,(B))$. The squared norm

accounts for both ± 1 differences between consecutive columns. Let $Y \in \Pi_m$ be an unknown column permutation matrix minimizing the total number of circular column transitions; define $\#C(A, Y) = \|AY(I - S_c)\|^2$, then we have to optimize $\min \#C(A, Y)$ over column permutations $Y \in \Pi_m$. It is straightforward to set it as a QAP since

$$\begin{aligned} \#C(A, Y) &= (AY(I - S_c), AY(I - S_c)) = (A^t AY, Y(I - S_c)(I - S_c)^t) \\ &= \text{QAP}(Y; A^t A, T_m) \end{aligned}$$

one of the most difficult problem in combinatorial optimization whose traveling salesperson problem (TSP) is a special case. By analogy, we could minimize the number of circular row transitions over row permutations $X \in \Pi_n$,

$$\begin{aligned} \#R(A, X) &= ((I - S_c)^t XA, (I - S_c)^t XA) = ((I - S_c)(I - S_c)^t X, XAA^t) \\ &= \text{QAP}(X; T_n, AA^t) \end{aligned}$$

where

$$T_n = \begin{bmatrix} 2 & -1 & \dots & -1 \\ -1 & 2 & -1 & \dots \\ & & \dots & \\ -1 & \dots & -1 & 2 \end{bmatrix},$$

a Toeplitz matrix with neat eigenvalues $2(1 - \cos \frac{2k\pi}{n})$ for $k = 0, n - 1$. Minimizing in both dimensions at the same time is twice involved since

$$\begin{aligned} \#R(A, X) + \#C(A, Y) = &((I - S_c)^t XAY, (I - S_c)^t XAY) \\ &+ (XAY(I - S_c), XAY(I - S_c)) \\ &\text{s.t. } X \in \Pi_n, \quad Y \in \Pi_m \end{aligned}$$

since $XX^t = I_n$ and $YY^t = I_m$, respectively.

Solving QAP is hard in general, only a few cases are known to be polynomially solvable see Sect. 5, therefore, we have to recourse to a Branch and Bound (B&B) scheme in order to prove optimality of relaxed problem:

$$\begin{aligned} &\text{minQAP}(X; T_n, AA^t) \\ &\text{s.t. } X \in \mathcal{E}_n \end{aligned}$$

where $\mathcal{E}_n$ stands for doubly-stochastic matrices, a nice domain described by linear constraints $Xe = e$ and $X^t e = e$ for the all 1s vector e. Notice the harness involved with the squaring in the n^2 unknown entries in X while the dimension is merely n. See Sect. 6 for small-sized experiments.

Table 1 Eigenvalues of Toeplitz matrices for different shifts

Type	Shift	Eigenvalues	
Circular	$s \geq 1$	$2(1-\cos\frac{2sk\pi}{n})$	$k = 0 \cdots n-1$
Standard	1	$2(1-\cos\frac{k\pi}{n})$	$k = 0 \cdots n-1$
Standard n even	2	$2(1-\cos\frac{k\pi}{n}), 2(1-\cos\frac{(k+1)\pi}{n+1})$	$k = 0, 2, \cdots n-2$
Standard n odd	2	$0, 2(1-\cos\frac{(k-1)\pi}{n}), 2(1-\cos\frac{k\pi}{n+1})$	$k = 2, 4 \cdots n-1$

If we count transitions without wrapping the matrix then circular shift is replaced by standard shift S and T_n has four corners modified:

$$I - S = \begin{bmatrix} 1 & 0 & \dots & 0 \\ -1 & 1 & 0 & \dots \\ & & \cdots & \\ \cdots & & & 1 \\ 0 & \dots & & -1 \end{bmatrix}$$

$$T_n = \begin{bmatrix} 1 & -1 & \dots & 0 \\ -1 & 2 & -1 & \dots \\ & & \cdots & \\ \cdots & -1 & 2 & -1 \\ 0 & \dots & -1 & 1 \end{bmatrix},$$

a Toeplitz matrix whose eigenvectors and eigenvalues are analytically known as a function of the dimension n [21]. Both circular and standard shifts extends to transitions between columns/rows at distances > 1 leading to symmetric semidefinite programming (SDP) Toeplitz matrices with off-diagonals at the same distance apart (whose eigenvectors and eigenvalues are analytically known as a function of the dimension n too [11]) (Table 1).

On the other hand, eigenvalues of AA^t and A^tA are derived from singular values of A since $\sigma = UAV$ for orthogonal matrices U, V yields diagonalization $\sigma^2 = UAVV^tA^tU^t = UAA^tU^t$ while $\sigma^2 = V^tA^tU^tUAV = V^tA^tAV$.

Let $\bar{A} = E_{nm} - A$, the complement of A for the $n \times m$ all 1s matrix E_{nm}; since the role played by 0s is the same as the 1s. It suggests to add the number of transitions in the complement $\#C(\bar{A}, Y) = \|\bar{A}Y(I - S_c)\|^2$. Using properties of permutation matrices $E_{nm}^t E_{nm} = nE_{mm} = E_m$, where a single index is a shortcut for square case, $E_m Y = E_m$ and $Y^t E_m = E_m$ then,

$$\begin{aligned} \#C(A, Y) + \#C(\bar{A}, Y) &= 2(A^tAY, YT) - ((\Delta + \Delta^t)Y, YT) + n(E_m, T_m) \\ &= 2\mathrm{QAP}(Y; A^tA, T_m) - \mathrm{QAP}(Y; \Delta + \Delta^t, T_m) + n(E_m, T_m), \end{aligned}$$

a kind of Laplacian for the *column degrees* $(\Delta + \Delta^t)/2 = (E_{nm}^t A + A^t E_{nm})/2$.

2.2 Maximizing Correlation

Define the in correlation as the product A^tA, another approach aims at maximizing the cumulated correlations after permutation, i.e., $((AY)^t(AY)U_m) = ((AY)^t(AY)L_m)$ for the all 1s $m \times m$ upper (respectively lower) triangular matrix U_m (respectively L_m). Using symmetry and properties of permutation matrices, we get

$$\begin{aligned}\mathrm{QAP}(Y; A^tA, U_m) + \mathrm{QAP}(Y; A^tA, L_m) &= (A^tAY, YE_m) + (A^tAY, YI_m)\\ &= (A^tAE_m, Y) + (A^tA, I_m)\\ &= \mathrm{LAP}(Y; A^tAE_m) + \mathrm{diag}(\Delta),\end{aligned}$$

a linear assignment problem with m equivalent solutions. Therefore, the intuitive correlation maximizing does not lead to a valid formulation.

3 Branch and Bound

For either correlation matrix F (AA^t and A^tA), let us consider a partial assignment, w.l.o.g. $X = \begin{bmatrix} X^{11} & 0 \\ 0 & X^{22} \end{bmatrix}$, where X^{22} stands for the unassigned indices, then the problem rewrites:

$$\begin{aligned}\mathrm{QAP}(X; T_n, F) = \mathrm{QAP}(X^{22}; T_n^{22}, F^{22}) + (T^{11}X^{11}, X^{11}F^{11})\\ + 2(T_n^{21}X^{11}F^{12}, X^{22}),\end{aligned}$$

where we used symmetry of both block submatrices $F^{21} = F^{12^t}$ and $T_n^{21} = T_n^{12^t}$. The second term is a constant and the last term is a linear assignment problem $\mathrm{LAP}(X^{22}; T_n^{21}X^{11}F^{12})$ since X^{11} is fully specified. The subproblem in first term may be bounded in various ways.

3.1 Spectral Bound

In Sect. 2.1, we gave the spectral bound for $\mathrm{QAP}(X^{22}; T_n^{22}, F^{22})$ at the root of the enumeration tree; however, deeper in the tree when some columns are assigned, the Toeplitz shift matrix shrinks to unassigned indices. For standard shift, since a fixed index symmetrically cancels a -1, it splits in three different patterns according to the diagonal corners $\mathrm{diag}([1, \ldots, 1])$, $\mathrm{diag}([1, \ldots, 2])$, $\mathrm{diag}([2, \ldots, 2])$ with possibly isolated eigenvalues (of value 1 or 2). The kernel method in [11, 21]) directly applies on corresponding diagonal discrepancies, namely, for size n, eigenvalues are $2(1 - \cos(\theta))$ with necessary conditions:

$$\text{diag}([1,\ldots,1]) : \sin(\theta)(\sin((n-1)\theta) - 2\sin(n\theta) + \sin((n+1)\theta)) = 0$$

$$\theta = \frac{k\pi}{n}, \quad k = 0..n-1,$$

$$\text{diag}([2,\ldots,2]) : \sin((n+1)\theta) = 0$$

$$\theta = \frac{k\pi}{n+1}, \quad k = 1..n,$$

$$\text{diag}([1,\ldots,2]) : \sin(\theta)(\sin((n+1)\theta - \sin(n\theta)) = 0$$

$$\theta = \frac{k\pi}{2n+1}, \quad k = 1,3,5\ldots 2n-1$$

For circular shift, however, the number of patterns increases unless the antidiagonals -1s corners are fixed, in which case it reduces to the standard shift patterns. Despite tractable, an analytical expression for circular eigenvalues is more involved.

The overall time complexity is $O(n^3)$ since singular value decomposition (SVD) requires very few sweeps of n^3 complexity each, the same complexity as linear assignment by, say hungarian method. Define $\lambda(T_n^{22}, F^{22}) \leq \text{QAP}(X^{22}; T_n^{22}, F^{22})$ as the corresponding spectral bound using Hoffman–Wielandt (Lidskii–Mirsky–Wielandt) inequalities:

$$(\lambda \nearrow (A), \lambda \searrow (B)) \leq (AU, UB) \leq (\lambda \nearrow (A), \lambda \nearrow (B))$$

for, respectively, ascending $\nearrow$, descending $\searrow$ orderings, then the whole bound simplifies to

$$\lambda(T_n^{22}, F^{22}) = (\lambda \nearrow (T_n^{22}), \lambda \searrow (F^{22})) + (T_n^{11} X^{11}, X^{11} F^{11})$$
$$\lambda(T_n^{22}, F^{22}) + 2\text{minLAP}(X^{22}; T_n^{21} X^{11} F^{12}) \leq \text{QAP}(X; T_n, F).$$

3.2 Gilmore–Lawler Bound

Every candidate assignment $x_{ij} = 1 \in X^{22}$ leads to a linear relaxation (see [19] and references therein) of $\text{QAP}(X^{22}; T_n^{22}, F^{22})$

$$\text{LAP}^{ij} \equiv (A^{ij} X^{22})$$
$$A_{kl}^{ij} = T_n^{22}{}_{ik} F^{22}{}_{jl}, \quad \text{for all } k \neq i, \text{for all } l \neq j.$$

By virtue of Hardy–Littlewood–Pólya (H.L.P. for short [13]), values of $l_{ij} = \text{minLAP}^{ij}$ for all i, j are easily retrieved by sorting rows in both matrices (negative entries in T_n^{22} are not actually a deal for which we can add E then subtract the constant to get the result):

$$l_{ij} = (T_{i.}^{22} \searrow, F_{j.}^{22} \nearrow).$$

Let $\mathcal{L}$ be the result for all such LAPs, then the whole bounding becomes:

$$\mathrm{LAP}(X^{22}; T_n, F, \mathcal{L}) = (\mathcal{L} + \mathrm{diag}(T_n^{22} \otimes F^{22}) + 2(T_n^{21} X^{11} F^{12}, X^{22}),$$
$$(T_n X^{11}, X^{11} F^{11}) + \mathrm{minLAP}(X^{22}; T_n, F, \mathcal{L}) \leq \mathrm{QAP}(X; T_n, F),$$

where $\mathrm{diag}(T_n^{22} \otimes F^{22})$ is assumed matrixified conformably with $\mathcal{L}$ (precisely in $n \times n$ row major order).

The overall time complexity is $O(n^2 \log n)$ for building $\mathcal{L}$ and $O(n^3)$ for final LAP. Notice that sorting, all but diagonal entries, simplifies to extracting maximum and second maximum to give $l_{ij} = -\max F_{i.}^{22} - \max_2 F_{i.}^{22}$ or $l_{ij} = -\max F_{i.}^{22}$ depending on either row in T_n^{22}, so that complexity shrinks to $O(n^2)$ for building $\mathcal{L}$ indeed.

4 Consecutive Signs Approximation

The QAP formulation carries over the signed version provided transitions between opposite signs never occurs contiguously; otherwise, the transition accounts for 2 instead of 1 as required. To circumvent this defect, we may apply to the ground set $\{0, +1, -1\}$ the three permutations, identity, left shift $\{+1, -1, 0\}$, and right shift $\{-1, 0, +1\}$ to yield a multiobjective problem. As usual for multiobjective problems, taking the mean of all three QAPs (Schur convexity) helps in searching the optimal solution, despite the average is accurate only at optimum. The sum of all three QAPs amounts to a single QAP so that the formulation for C1P directly applies at the expense of an actual counting at each integral node in the B&B enumeration.

5 About Problem Hardness

Testing consecutive ones property is known to be polynomial and a PC-tree storing all permutations fulfilling the property results when true; however, if the property fails then we only have a tree having prime nodes to deal with QAP formulation.

On the other hand, if the C1P is fulfilled then correlation matrix (either AA^t or $A^t A$) may be reordered through overlapping sets such that it is monotone anti-Monge (see Robinson property [3]). Since, Toeplitz matrices T are clearly benevolent [4] for the underlied Toeplitz function $f(1) = -1 <= 0 = f(j)$ for all $j \neq 0, 1$, then under the C1P, the monotone anti-Monge-benevolent Toeplitz QAP polynomially solvable case applies. Otherwise, either correlation matrix fails to fulfill anti-Monge monotone property and despite Toeplitz is benevolent, minimizing the number of transitions is NP-hard by reduction from the Hamming TSP [18] as soon as there is more than one row. It completes the panel of negative result for QAP easy solvable cases to the case of non-anti-Monge monotone-Toeplitz benevolent pair [4]. Moreover, it yields another direct hardness proof by reduction from the NP-complete even–odd partition problem, after converting each 1 in the matrix by a sequence 10 and each 0

by a sequence 01 to guarantee evenness while the number of transitions is doubled after unshuffling the overall sequence. Most generalizations like $(k-\delta)$C1P, for instance, leads to NP-hard problems [5].

Unlike other spectral approaches, ours does not force any convexity and straightforwardly generalizes to 3D dissimilarity data [17]: in real applications, very oftenly the objects indexing a dissimilarity $\mathcal{D}$ are measured with respect to some human predefined criterion so that there is a set of dissimilarities $\mathcal{D} = \{D_1, \ldots, D_K\}$ associated with each *measurable* criterion in $[1, k]$. Correspondingly, the minimization of the weighted number of transitions rewrites as a weighted maximization of correlations:

$$\#R(\mathcal{A}, X) = \sum_{k=1}^{K} w_k \mathrm{QAP}(X; T_n, A_k A_k^t).$$

6 Experiments

For a maximum of 1000 B&B nodes, the history of enumeration for Gilmore–Lawler bounds and Hoffman–Wielandt (spectral) bounds, showed in minimizing case, that the spectral lower bound happens to be far below than Gilmore–Lawler lower bound, despite the Toeplitz structure gives eigenvalues for free. Many subproblems in the history have a negative Hoffman–Wielandt lower spectral bound due to the remaining assignment part while the number of transitions is obviously nonnegative!

As for eigenvalues of subproblems, the corresponding Toeplitz matrix decomposes into a principal diagonal matrix fulfilling the standard case from upper left corner $\begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$ to lower right corner $\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ while the remaining matrix simplifies to a smaller circular Toeplitz case along one pair of 0s on symmetric entries at upper and lower diagonals. W.l.o.g. assume they occur at entries $(1, 2)$, $(2, 1)$, then applying the row and column permutations $[2, 3, \ldots, n, 1]$ yields a standard shift case with diagonal corners equal to 2, like in standard subproblems. Notice that, due to this observation, we need no facet defining inequalities for sum of hermitian matrices [15] and directly retrieve the eigenvalues of subproblems instead.

Due to these spectral observations, we do not report the history of enumeration for bounds and the examples below were rerun with the Gilmore–Lawler bound solely.

It is worth noticing that the standard count associated with the circular count is not monotonic; it may happen in the enumeration that standard count may be better than the next standard count associated to the next circular incumbent (upto 1 or 2 transitions); however, we always obtain the best standard incumbent from the best circular incumbent. Finally, for all experiments we never succeed in closing the gap within the node limit; however, the index of the best incumbent remains far from the node limit so that we are in the common situation where the optimum is very likely but the proof of optimality is hopeless due to the huge number of nodes remaining.

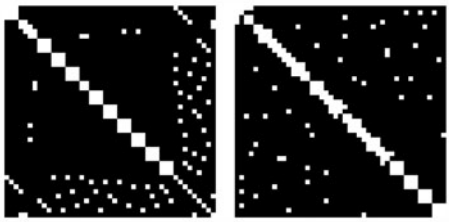


Fig. 1 MANN_a9 quadratic assignment problem (QAP; min, start, max) circular (respectively standard) rowwise transitions at (400,0,378)th Branch and Bound (B&B) nodes and value (184,234,378; respectively (176,225,293))

Fig. 2 Johnson8-2-4 quadratic assignment problem (QAP; min, start, max) circular (respectively standard) rowwise transitions at (13,0,8)th Branch and Bound (B&B) nodes and value (280,328,504; respectively (270,310,480))

6.1 C1P on Symmetric Matrices

We borrow from maximum clique Dimacs benchmark, two easy examples for the maximum clique problem that exhibit two opposite behaviors w.r.t. the enumerative procedure.

6.2 C±P on Random Matrix

The lack of dedicated benchmarks for the consecutive sign problem leads us to test the formulation against random matrices. As in Sect. 6.1, we draw the (best min incumbent, original, best max incumbent) in black, white, and gray pixels instead, according to the trivaluation. The enumeration is applied on both a single valuation (Fig. 3) and the averaging approximation (Fig. 4). Clearly, the averaging formulation yields better results for both the circular count and the standard count but unlike the C1P case there is no visual evidence that the best incumbent improves over the original data despite the relative improvements are comparable (see Table 2); however, the visual evidence looks stronger between the best min and max incumbents. Our random examples have a dense number ($>60\,\%$) of transitions among the complete

Fig. 3 Random 45 × 45quadratic assignment problem (QAP; min, start, max) circular (respectively standard) rowwise transitions at (565,0,0)th Branch and Bound (B&B) nodes and value (1174,1348,1343; respectively (1150,1313,1321))

Fig. 4 Random 45 × 45 average quadratic assignment problem (QAP; min, start, max) circular (resp. standard) rowwise transitions at (141,0,36)th Branch and Bound (B&B) nodes and value (1141,1348,1568) (resp. (1118,1313,1531))

Table 2 Relative improvement of rowwise transitions between original and best incumbent

	Minimize		Maximize	
Example	Circular (%)	Standard (%)	Circular (%)	Standard (%)
C1P (Fig. 1)	21	22	38	23
C1P (Fig. 2)	15	13	44	35
C±P (Fig. 3)	13	12	0	0
C±P (Fig. 4)	15	15	14	14

alternate case, so the best incumbent in standard max case is mostly found at the root of the B&B tree unlike the averaging formulation.

7 Circular Shift versus Standard Shift

Though most real-life applications require standard shift formulation, counting circular transitions makes sense for it is clearly an upper bound for C1P; since the optimization problem is more constrained, the B&B enumeration seems faster when there is a big gap between bounds and slower when the gap is narrow. It always lead in our experiments, to better standard incumbents than the standard formulation.

8 Concluding Remarks

Unlike Vuokko [18] who stresses, after Atkins et al. [1], the efficiency of spectral ordering by studying the Laplacian of an underlied graph, we study in this paper a QAP formulation for optimizing the number of transitions with respect to C1P. The formulation easily carries over the signed version of the problem C±P through an averaging technique; to our knowledge, the recognition of the consecutive signs Property has not been studied yet and opens up a challenging question about its polynomiality.

Experiments for both standard and signed versions, happen to be effective within the B&B framework even though the computation load is more heavy than the Fiedler vector of the Laplacian approach. Despite the spectral bounding remains deceptive compared with the standard Gilmore–Lawler bounding, it is worth noticing that the Toeplitz structure of one matrix in the QAP allows the same $O(n^3)$ worse time complexity for both bounds.

As a side effect, we find another negative result for the QAP pair of matrices [4], a not anti-Monge monotone and Toeplitz benevolent pair.

References

1. Atkins, J.E., Boman, E.G., Hendrickson, B.: A spectral algorithm for seriation and the consecutive ones problem. SIAM J. Comput., **28**(1), 297–310 (1999) (electronic)
2. Bondy, J.A.: Trigraphs. Discrete Math. **75**(1–3), 69–79 (1989). (Graph theory and combinatorics, Cambridge, 1988)
3. Brusco M.J.: A branch-and-bound algorithm for fitting anti-Robinson structures to symmetric dissimilarity matrices. Psychometrika **67**(3), 459–471 (2002)
4. Burkard, R.E., Çela, E., Rote, G., Woeginger, G.J.: The quadratic assignment problem with a monotone anti-Monge and a symmetric Toeplitz matrix: easy and hard cases. Math. Programming **82**(1–2, Ser. B), 125–158 (1998). (Networks and matroids; Sequencing and scheduling)
5. Chauve, C., Maňuch, J., Patterson, M.: On the gapped consecutive-ones property. European Conference on Combinatorics, Graph Theory and Applications (EuroComb 2009). Electronic Notes Discrete Math., vol. 34, pp. 121–125. Elsevier Sci. B. V., Amsterdam (2009)
6. Chepoi, V., Seston, M.: Seriation in the presence of errors: A factor 16 approximation algorithm for l_∞-fitting Robinson structures to distances. Algorithmica **59**(4), 521–568 (2011)
7. Chudnovsky, M.: The structure of bull-free graphs I—three-edge-paths with centers and anticenters. J. Combin. Theory Ser. B **102**(1), 233–251 (2012)
8. Feder, T., Hell, P., Kim, T.-N.: Digraph matrix partitions and trigraph homomorphisms. Discrete Appl. Math. **154**(17), 2458–2469 (2006)
9. Fiedler, M., Hall, F.J., Stroev, M.: Dense alternating sign matrices and extensions. Linear Algebra Appl. **444**, 219–226, (2014)
10. Fogel, F., Jenatton, R., Bach, F., d'Aspremont, A.: Convex relaxations for permutation problems. http://arXiv:1306.4805[math.OC] (2013)
11. D. Fortin. Eigenvectors of Toeplitz matrices under higher order three term recurrence and circulant perturbations. Int. J. Pure Appl. Math., **60**(2), 217–228 (2010)
12. Fortin, D., Tseveendorj, I.: Generalized subdifferentials of the sign change counting function. http://hal.inria.fr/hal-00915606 (2013)(submitted)

13. Hardy, G.H., Littlewood, J.E., Pólya, G.: Inequalities. Cambridge Mathematical Library. Cambridge University Press, Cambridge (1988). (Reprint of the 1952 edition)
14. Hiriart-Urruty, J.-B., Yen Le, H.: Convexifying the set of matrices of bounded rank: Applications to the quasiconvexification and convexification of the rank function. Optim. Lett. **6**(5), 841–849 (2012)
15. Knutson, A., Tao, T., Woodward, C.: The honeycomb model of $GL_n(\mathbf{C})$ tensor products. II. Puzzles determine facets of the Littlewood-Richardson cone. J. Amer. Math. Soc. **17**(1), 19–48 (2004) (electronic)
16. Maňuch, J., Patterson, M., Chauve, C.: Hardness results on the gapped consecutive-ones property problem. Discrete Appl. Math. **160**(18), 2760–2768 (2012)
17. Vicari, D., Vichi, M.: Structural classification analysis of three-way dissimilarity data. J. Classification **26**(2), 121–154 (2009)
18. Vuokko, N.: Consecutive ones property and spectral ordering. In: SIAM International Conference on Data MINING, Colombus, Ohio, April 2010. SDM10, pp. 350–360 (2010)
19. Xia, Y., Gilmore-Lawler bound of quadratic assignment problem. Front. Math. China **3**(1), 109–118 (2008)
20. Yen Le, H.: Confexifying the counting function on $\mathbf{R}^p$ for convexifying the rank function on $\mathcal{M}_{m,n}(\mathbf{R})$. J. Convex Anal. **19**(2), 519–524 (2012)
21. Yueh, W.-C., Cheng, S.S.. Explicit eigenvalues and inverses of tridiagonal Toeplitz matrices with four perturbed corners. ANZIAM J. **49**(3), 361–387 (2008)
22. Zeilberger, D.: Proof of the alternating sign matrix conjecture. Electron. J. Combin. **3**(2), 84 (1996). (Research Paper 13, electronic, 1996. The Foata Festschrift)

Variational and Hemivariational Inequalities in Mechanics of Elastoplastic, Granular Media, and Quasibrittle Cracks

Boris D. Annin, Victor A. Kovtunenko and Vladimir M. Sadovskii

Abstract This contribution is devoted to the mathematical theory of elastoplastic and granular solids as well as the quasibrittle fracture of nonlinear cracks. Basic variational and hemivariational inequalities describing nonlinear phenomena due to plasticity, internal friction, interfacial interaction, and alike dissipative physics are outlined from the point of view of nonsmooth and nonconvex optimization. Primary results of the nonlinear theory and its application to solid mechanics are surveyed.

Keywords Plasticity · Granular solid · Quasibrittle crack · (Hemi)variational inequality · Set-valued optimization · Constrained optimization · Nonsmooth optimization · Nonconvex optimization

1 Introduction

The mathematical theory of elastoplastic and granular solids as well as their fracture is originated in the engineering sciences related to materials, geophysics, and biophysics. As it is marked in the literature, modern materials developed in the recent past exhibit essentially nonlinear properties. In particular, when the materials are undergoing critical deformations. This motivates the actual research of nonlinear

V. A. Kovtunenko (✉)
Lavrentyev Institute of Hydrodynamics, Siberian Branch of the Russian Academy of Sciences, 630090 Novosibirsk, Russia

Institute for Mathematics and Scientific Computing, Karl-Franzens University of Graz, NAWI Graz, Heinrichstr. 36, 8010 Graz, Austria
e-mail: victor.kovtunenko@uni-graz.at

B. D. Annin
Lavrentyev Institute of Hydrodynamics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk State University, 630090 Novosibirsk, Russia
e-mail: annin@hydro.nsc.ru

V. M. Sadovskii
Institute of Computational Modeling, Siberian Branch of the Russian Academy of Sciences, Akademgorodok, 660036 Krasnoyarsk, Russia
e-mail: sadov@icm.krasn.ru

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_3

phenomena due to plasticity, internal friction, interfacial interaction, and other dissipative physics. For its mathematical modeling, variational and hemivariational inequalities are well suitable. In the present contribution this chapter, we survey the principal issues of modeling in this respect.

Variational inequalities were applied to the problems of mechanics yet in nineteenth century by V. M. Ostrogradsky for the description of constrained motion of a material point. Significant development of the methods using inequalities in mechanics was contributed recently by (alphabetically) B. D. Annin, G. Fichera, A. Haar, J. Haslinger, T. Karman, A. S. Kravchuk, J.-L. Lions, P. D. Panagiotopoulos, A. Signorini, R. Temam, and others.

From the mathematical point of view, variational and hemivariational inequalities appear in the governing relations as the consequence of fundamental thermodynamics principles subject to one-sided restrictions. In fact, inequality constraints imposed on geometric displacements lead to contact conditions, virtual stress that does not exceed the yield limit implies plasticity, while the material strength is expressed by restrictions on strains.

In the following sections, we outline, respectively, the modeling of governing inequalities for elastoplastic and granular media, and in the theory of quasibrittle cracks.

2 Variational Inequalities in Elastoplastic Theory

We start with the notation. For a reference solid occupying the domain Ω, spatial points $x \in \Omega$, and time $t \geq 0$, we refer the displacement vector $u(t, x)$, the strain $\varepsilon(u)$ which relies on the symmetric tensor $\varepsilon(u) = \frac{1}{2}(\nabla u + \nabla^\top u)$ of linear deformations, and the stress tensor σ. In elastoplasticity, there is assumed a plastic strain ε^p.

Within the *Hencky law,* it holds (see the books [1, 13]) the constitutive equation:

$$\varepsilon(u) = \tfrac{\partial W}{\partial \sigma} + \varepsilon^p \tag{1}$$

supported with the following variational inequalities:

$$f(\sigma) \leq 0, \quad (\bar{\sigma} - \sigma) : \varepsilon^p \leq 0 \quad \text{for all } \bar{\sigma} \colon f(\bar{\sigma}) \leq 0. \tag{2}$$

In (1), the notation $W(\sigma)$ stands for the strain energy potential. In linear elasticity, it is quadratic, $W(\sigma) = \frac{1}{2}\sigma : A : \sigma$ with the symmetric compliance tensor A, hence $\frac{\partial W}{\partial \sigma} = A : \sigma$. Inequality (2) imply that the true stress σ lies inside the given yield surface $f(\sigma) \leq 0$, and the plastic deformation ε^p is orthogonal to this surface. Typical yield surfaces are the Tresca and von Mises ones. In the simplest case of scalar σ, the yield surface is determined by $f(\sigma) = |\sigma| - \sigma^0$ with the yield limit σ^0. Together with the (quasi)static equilibrium equation:

$$-\nabla \cdot \sigma = F, \tag{3}$$

where F is the external force, governing relations (1)–(3) form a complete system.

Note that, if the one-sided constraint $f(\sigma) \le 0$ was skipped, then inequalities (2) would turn into the equality $\varepsilon^p = 0$. In this case, there is no plastic deformations, and governing relations (1)–(3) correspond to common (generally nonlinear) elasticity.

From the nonlinear optimization viewpoint, the variational inequality (2) implies that ε^p is a (nonunique) element of the Clarke subdifferential:

$$\varepsilon^p \in \underline{\partial}\chi_K(\sigma) := \{\varepsilon : \chi_K(\bar{\sigma}) - \chi_K(\sigma) \ge (\bar{\sigma} - \sigma) : \varepsilon \text{ for all } \bar{\sigma}\}. \tag{4}$$

In (4), the nonsmooth potential χ_K implies the indicator function of the convex set:

$$K = \{\bar{\sigma} : f(\bar{\sigma}) \le 0\} \tag{5}$$

that is $\chi_K(\sigma) = 0$ for $\sigma \in K$ if $f(\sigma) \le 0$, otherwise $\chi_K(\sigma) = +\infty$. Details of this formalism are presented in [29] and endowed with dual arguments using the Legendre–Fenchel–Young transformation.

The *Prandtl–Reuss law* provides the following flow model (see [2, 30]):

$$\varepsilon(v) = \tfrac{\partial}{\partial t}(\tfrac{\partial W}{\partial \sigma}) + e^p \tag{6}$$

with the velocities $v := \dot{u}$ and $e^p := \dot{\varepsilon}^p$, the variational inequality:

$$f(\sigma) \le 0, \quad (\bar{\sigma} - \sigma) : e^p \le 0 \quad \text{for all } \bar{\sigma}\colon f(\bar{\sigma}) \le 0, \tag{7}$$

and the dynamic equation of motion:

$$\rho\dot{v} - \nabla \cdot \sigma = F. \tag{8}$$

Following the approach of J. Mandel, discontinuous solutions to the dynamic elastoplastic problems (6)–(8), which are of the shock wave type, are analyzed in the monograph [30]. In this case, as far as in more general case of hardening materials, the system of governing equations can be written in the unified form:

$$\Big(\tfrac{\partial \varphi(U)}{\partial t} - \sum_{k=1}^{n} \tfrac{\partial \psi_k(U)}{\partial x_k}\Big) \cdot (\bar{U} - U) \ge 0 \tag{9}$$

where n is the spacial dimension. In (9), the true variable U and its variation $\bar{U}$ are admissible vectors composing all unknown functions: velocities, stresses, and parameters of hardening. The vector functions $\varphi = \frac{\partial \Phi}{\partial U}$ and $\psi_k = \frac{\partial \Psi_k}{\partial U}$ are expressed in the terms of scalar generating potentials $\Phi(U)$ and $\Psi_k(U)$, $k = 1, \ldots, n$, which are quadratic functions in the case of physically linear processes.

The formulation in form (9), in particular, is numerically advantageous for the construction of algorithms of the Wilkins type, see [2]. This theoretical result is strengthened by the engineering applications presented in [3, 4].

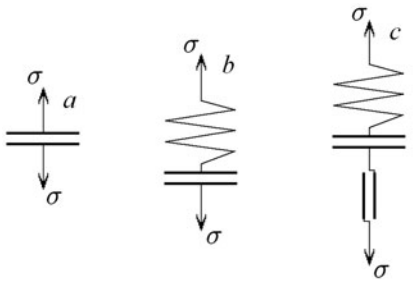


Fig. 1 Rheological schemes of granular material with rigid (**a**), elastic (**b**) and elastic–plastic (**c**) particles

3 Variational Inequalities in Theory of Granular Media

Granular solids exhibit complex behavior, in particular, different resistance under compression and tension. To take it into account, the classical rheological method is supplemented by a new element, namely, the rigid contact which can be expressed schematically as two plates being in contact. Combining this element with the traditional rheological elements; the elastic spring, viscous dashpot, and plastic hinge models with a suitable level of complexity can be derived (see the examples in Fig. 1).

The constitutive relationships expressing the rigid contact can be represented as the linear complementarity problem:

$$\sigma \leq 0, \quad \varepsilon \geq 0, \quad \sigma\,\varepsilon = 0 \tag{10}$$

for the scalar stress σ and strain ε. Indeed, the inequalities in (10) exclude arising tensile stresses and compressive strains in a perfect granular material composed of rigid particles. From the complementarity condition, it follows that one of the quantities being considered (either stress or strain) must be zero. Therefore, (10) can be reduced to two variational inequalities:

$$\sigma \leq 0, \quad (\sigma - \bar{\sigma})\varepsilon \leq 0 \quad \text{for all } \bar{\sigma} \leq 0, \tag{11}$$

$$\varepsilon \geq 0, \quad \sigma(\bar{\varepsilon} - \varepsilon) \leq 0 \quad \text{for all } \bar{\varepsilon} \geq 0, \tag{12}$$

which are equivalent. This consideration admits extension to more complicated rheological models and higher spacial dimensions.

Within the rheological approach, phenomenological models of granular solids are generalized in the book [29]. In [28], the model of granular materials under finite strains is considered. The generic model of materials with different compressive and tensile strengths is analyzed in [23] where the modeling result is supported with the existence theorems, analysis of mechanical properties, and estimation of the critical equilibrium.

4 Hemivariational Inequalities in Nonlinear Theory of Cracks

Macro- and microcracks appear in a wide range of real-world applications related to fracture in the material science, faults in the earth causing earthquakes and subsequent tsunami, modern biomedical methodologies, and alike. The actual problems of tribology and fracture need nonlinear modeling of cracking phenomena taking into account dissipative interaction between the crack faces. This results in quasibrittle models of fracture in contrast to classic brittle hypothesis due to Griffith–Irwin–Rice.

The fundamentals of quasibrittle and dynamic fracture were developed by (alphabetically) G. I. Barenblatt, G. P. Cherepanov, S. A. Christianovich, D. S. Dugdale, R. V. Goldstein, M. Ya. Leonov, N. F. Morozov, V. V. Novozhilov, V. V. Panasyuk, Yu. N. Rabotnov, L. Truskinovsky, and others. From a physical point of view, the dissipative work of interaction phenomena due to contact with cohesion or friction at the crack is closely related to elastoplastic physics. From a mathematical viewpoint, its modeling results in hemivariational inequalities within set-valued and nonconvex optimization context.

The basics of mathematical theory describing quasibrittle cracks are outlined below by following the results obtained in [11, 17, 19, 20]. In the nonlinear optimization framework, we suggest a class of hemivariational inequalities introduced as follows.

Let $\Gamma \subset \Omega$ be an interface. In the equilibrium equation (3), the total stress σ distribution admits the following representation (compare with (1)):

$$\sigma = \frac{\partial W}{\partial \varepsilon} + F^i \chi_\Gamma \tag{13}$$

with the stress energy potential $W(\varepsilon)$ and the interfacial traction F^i. It is added in Ω with the help of characteristic function χ_Γ of the interface Γ. In linear elasticity where W is quadratic, $W(\varepsilon) = \frac{1}{2}\varepsilon : C : \varepsilon$ with the symmetric tensor C of elastic stiffness, hence $\frac{\partial W}{\partial \varepsilon} = C : \varepsilon$ in (13).

At Γ, we suggest complementary contact conditions (compare with (10)):

$$\mathrm{tr}_\Gamma(u) \geq 0, \quad F^c \geq 0, \quad F^c \,\mathrm{tr}_\Gamma(u) = 0 \tag{14}$$

where the contact force F^c admits, generally, the decomposition as:

$$F^c = -F^i + F^d. \tag{15}$$

In (15), the dissipative force F^d represents irreversible work caused by cohesion as well as friction at the interface Γ. In the context of cracks, F^d describes interaction force between two crack faces being in contact, and $\mathrm{tr}_\Gamma(u)$ implies the jump of the normal traces of u across the crack [18]. For more issues of the modeling of nonpenetration conditions $\mathrm{tr}_\Gamma(u) \geq 0$, see [13, 16].

With a generating potential g, which is typically a concave function, the cohesion force F^d in (15) can be expressed as a (nonunique) element of the superdifferential:

$$F^d \in \overline{\partial} g(u) := \{F : g(\mathrm{tr}_\Gamma(\bar{u})) - g(\mathrm{tr}_\Gamma(u)) \leq F \,\mathrm{tr}_\Gamma(\bar{u} - u) \text{ for all } \bar{u}\}. \tag{16}$$

Examples of the generating potential g are given in [17]. For the reference setup:

$$g(\mathrm{tr}_\Gamma(u)) = \tfrac{\sigma^0}{l}\min(l, \mathrm{tr}_\Gamma(u)),$$

with the yield limit σ^0 and the characteristic length l.

After integration over the domain $\Omega \setminus \Gamma$, relations (13)–(15) can be summarized as the hemivariational inequality:

$$\mathrm{tr}_\Gamma(u) \geq 0, \quad \int_{\Omega\setminus\Gamma} (\tfrac{\partial W}{\partial \varepsilon} - \sigma) : \varepsilon(\bar{u} - u)\,\mathrm{d}x + \int_\Gamma F^d\,\mathrm{tr}_\Gamma(\bar{u} - u)\,\mathrm{d}S_x \geq 0, \qquad (17)$$

$$\text{for all } \bar{u}\text{: } \mathrm{tr}_\Gamma(\bar{u}) \geq 0.$$

In fact, inclusion (16) argues that (17) is the necessary optimality condition for minimization over admissible $\bar{u}$ of the nonsmooth functional of energy:

$$\text{minimize} \left\{ \int_{\Omega\setminus\Gamma} (W - \sigma) : \varepsilon(\bar{u})\,\mathrm{d}x + \int_\Gamma g(\mathrm{tr}_\Gamma(\bar{u}))\,\mathrm{d}S_x \right\} \text{ subject to } \mathrm{tr}_\Gamma(\bar{u}) \geq 0. \qquad (18)$$

Moreover, the energy functional is nonconvex since g is concave.

The issues of nonsmoothness and nonconvexity are the principal difficulties for analysis of the constrained minimization problem (18) which is presented in the cited works.

While the hemivariational inequality (17) is necessary to (18), its sufficient optimality condition implies a saddle point minimax problem with respect to the pair of the primal variable u and the dual variable F^c (the Lagrange multiplier) associated to the constraint $\mathrm{tr}_\Gamma(u) \geq 0$ in accordance with (14). The saddle-point problem reads

$$\min_u \max_{F^c} \left\{ \int_{\Omega\setminus\Gamma} (W - \sigma) : \varepsilon(\bar{u})\,\mathrm{d}x + \int_\Gamma g(\mathrm{tr}_\Gamma(\bar{u}))\,\mathrm{d}S_x - \int_\Gamma F^c\,\mathrm{tr}_\Gamma(\bar{u})\,\mathrm{d}S_x \right\} \qquad (19)$$

$$\text{subject to } F^c \geq 0.$$

For nonunique solutions, the sufficient and necessary conditions do not coincide with each other. This fact is in contrast to the case of convex minimization.

For the numerical solution of (19), hence (18) and (17), a primal dual active set (PDAS)-based strategy is suggested in [11, 19]. The PDAS strategy is associated to generalized Newton methods obeying locally superlinear as well as globally monotone convergence properties.

5 Conclusion

Here, we outline the further development of the subject directed to variational modeling of fractional, damage, and geometrically singular phenomena in mechanics.

Starting with frictional contact due to the Coulomb law [7, 12], typically, tangential components of the shear are subject to restriction. Its generalization is developed in [6] for nonmonotone friction laws, and further in [21] for a cohesive–frictional interaction restricting both the tangential as well as the normal shear components. The resulting hemivariational inequalities are argued as pseudo-monotone variational inequalities by the authors of [25].

The actual task concerns singular geometries, see [10, 14], arising in practice. Motivated by fracture of composites (and used also in inverse problems for stratified media [24]), geometrically heterogeneous models with nonlinear inclusions subject to cracks and anticracks were developed in [15]. This study aims at the shape-topological control to force either shielding or amplification of an incipient cracking. For the variational analysis, Γ-convergence techniques are useful [26].

The other development consists in constituting variational models of damaged elastic, elastoplastic, and cracked materials. The damaged models are treated by using Γ-limits in [8] and within hysteresis formalism and rate-independent systems in [22].

In respect to numerical theory of the underlying optimization problems, we refer to [5] for saddle-point algorithms within nonconvex programming, to [9] for globalization strategies, and to [27] for parametric and dynamic optimization.

Acknowledgement B. D. Annin is supported by the Russian Foundation for Basic Research (grant no. 12-01-00507). V. A. Kovtunenko is supported by the Austrian Science Fund (FWF), project P26147-N26. He thanks O. Vasilieva and J. R. Gonzalez for his visit to the Universidad Tecnologica de Pereira and ICAMI 2013 with the support of the Colombian Department for Science (COLCIENCIAS). V. M. Sadovskii is supported by the Complex Fundamental Research Program no. 18 "Algorithms and Software for Computational Systems of Superhigh Productivity" of the Presidium of Russian Academy of Sciences and by the Russian Foundation for Basic Research (grant no. 14–01–00130).

References

1. Annin, B.D., Cherepanov, G.P.: Elastic–Plastic Problems. ASME, New York (1998)
2. Annin, B.D., Sadovskii, V.M.: On the numerical realization of a variational inequality in problems of the dynamics of elastoplastic bodies. Comput. Math. Math. Phys. **36**, 1313–1324 (1996)
3. Annin, B.D., Sadovsky, V.M.: A numerical analysis of laminated elastic-plastic plates under dynamic loading. Compos. Sci. Technol. **45**, 241–246 (1992)
4. Annin, B.D., Sadovskaya, O.V., Sadovskii, V.M.: Dynamic contact problems of elastoplasticity. Problems Mater. Sci. **33**, 426–434 (2003)
5. Antipin, A.: Splitting of the gradient approach for solving extremal inclusions. Comput. Math. Math. Phys. **38**, 1069–1082 (1998)
6. Baniotopoulos, C.C., Haslinger, J., Moravkova, Z.: Mathematical modeling of delamination and nonmonotone friction problems by hemivariational inequalities. Appl. Math. **50**, 1–25 (2005)
7. Bychek, O.V., Sadovskii, V.M.: On the investigation of the dynamic contact interaction of deformable bodies. J. Appl. Mech. Tech. Phys. **39**, 628–633 (1998)

8. Dal Maso, G., Iurlano, F.: Fracture models as Γ-limits of damage models. Commun. Pure Appl. Anal. **12**, 1657–1686 (2013).
9. Fortin, D., Tseveendorj, I.: Q-subdifferential and Q-conjugate for global optimality. Comput. Math. Math. Phys. **54**, 265–274 (2014)
10. Fremiot, G., Horn, W., Laurain, A., Rao, M., Sokolowski, J.: On the analysis of boundary value problems in nonsmooth domains. Dissertationes Mathematicae, Inst. Math. Polish Acad. Sci., vol. 462, Warsaw (2009)
11. Hintermüller, M., Kovtunenko, V.A., Kunisch, K.: Obstacle problems with cohesion: A hemivariational inequality approach and its efficient numerical solution. SIAM J. Optim. **21**, 491–516 (2011)
12. Itou, H., Kovtunenko, V.A., Tani, A.: The interface crack with Coulomb friction between two bonded dissimilar elastic media. Appl. Math. **56**, 69–97 (2011)
13. Khludnev, A.M., Kovtunenko, V.A.: Analysis of Cracks in Solids. WIT-Press, Southampton (2000)
14. Khludnev, A.M., Kovtunenko, V.A., Tani, A.: On the topological derivative due to kink of a crack with non-penetration. Anti-plane model. J. Math. Pures Appl. **94**, 571–596 (2010)
15. Khludnev, A.M., Leugering, G.: Optimal control of cracks in elastic bodies with thin rigid inclusions. ZAMM Z. Angew. Math. Mech. **91**, 125–137 (2011)
16. Khludnev, A.M., Sokolowski, J.: Modelling and Control in Solid Mechanics. Birkhäuser, Basel (1997)
17. Kovtunenko, V.A.: Nonconvex problem for crack with nonpenetration. Z. Angew. Math. Mech. **85**, 242–251 (2005)
18. Kovtunenko, V.A.: Primal-dual methods of shape sensitivity analysis for curvilinear cracks with non-penetration. IMA J. Appl. Math. **71**, 635–657 (2006)
19. Kovtunenko, V.A.: A hemivariational inequality in crack problems. Optimization **60**, 1071–1089 (2011)
20. Kovtunenko, V.A., Sukhorukov, I.V.: Optimization formulation of the evolutionary problem of crack propagation under quasibrittle fracture. Appl. Mech. Tech. Phys. **47**, 704–713 (2006)
21. Leugering, G., Prechtel, M., Steinmann, P., Stingl, M.: A cohesive crack propagation model: Mathematical theory and numerical solution. Commun. Pure Appl. Anal. **12**, 1705–1729 (2013)
22. Mielke, A., Truskinovsky, L.: From discrete visco-elasticity to continuum rate-independent plasticity: Rigorous results. Arch. Ration. Mech. Anal. **203**, 577–619 (2012)
23. Myasnikov, V.P., Sadovskii, V.M.: Variational principles of the theory of limit equilibrium of media with different strengths. J. Appl. Math. Mech. **68**, 437–446 (2004)
24. Oralbekova, Zh.O., Iskakov, K.T., Karchevsky, A.L.: Existence of the residual functional derivative with respect to a coordinate of gap point of medium. Appl. Comput. Math. **12**, 222–233 (2013)
25. Ovcharova, N., Gwinner, J.: A study of regularization techniques of nondifferentiable optimization in view of application to hemivariational inequalities. J. Optim. Theory Appl. **162**, 754–778 (2014).
26. Plotnikov, P.I., Rudoy, E.M.: Shape sensitivity analysis of energy integrals for bodies with rigid inclusions and cracks. Dokl. Math. **84**, 681–684 (2011).
27. Radwan, A., Vasilieva, O., Enkhbat, R., Griewank, A., Guddat, J.: Parametric approach to optimal control. Optim. Lett. **6**, 1303–1316 (2012)
28. Sadovskaya, O.V., Sadovskii, V.M.: On the theory of finite deformations of a granular medium. J. Appl. Math. Mech. **71**, 93–110 (2007)
29. Sadovskaya, O., Sadovskii, V.: Mathematical Modeling in Mechanics of Granular Materials. Springer, Berlin (2012)
30. Sadovskii, V.M.: Discontinuous Solutions in Dynamic Elastic-Plastic Problems. Fizmatlit, Moscow (1997) (in Russian)

Effects of a Discrete Time Delay on an HIV Pandemic

Ibrahim Diakite and Benito M. Chen-Charpentier

Abstract We investigate the effects of a discrete time delay on the disease progression of a human immunodeficiency virus (HIV) pandemic. We consider a model of the cell-free viral spread of HIV in a well-mixed compartment such as the bloodstream. A discrete time delay is introduced to take into account the time between the infection of a $CD4^+$ T cell and the emission of viral particles at the cellular level. We first investigate the effects of the delay on the virulence of the HIV strains. We derive an analytical expression of the evolutionary stable strategy (ESS), and characterize how changes in the delay could alter that evolutionary optimum. Our analysis will show that the ESS of the HIV strains does not depend on the delay; however, the virulence of the HIV strains may increase as a consequence of increasing the delay time. We then present an analytic stability analysis of the endemically infected equilibrium.

We also present a *novel* numerical analysis of the stability and bifurcation process of the same equilibrium using numerical tools. With the numerical methods, we are able to reach the same conclusion as reached analytically.

Keywords Delay differential equations · Evolutionary stable strategy · Hopf bifurcation

1 Introduction

All processes take time to complete. While some physical processes happen very fast, biological process times such as gestation periods and maturation times can be substantial when compared to the data collection times in most population studies [9]. Therefore, to have more realistic models of many biological processes it is imperative to explicitly incorporate these process times into the mathematical models. Such

I. Diakite (✉) · B. M. Chen-Charpentier
University of Texas at Arlington, Arlington, TX 76019-0408, USA
e-mail: ibrahim.diakite@mavs.uta.edu

B. M. Chen-Charpentier
e-mail: bmchen@uta.edu

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_4

delay models are referred as delay differential equation (DDE) models. Previous works have shown how crucial a time delay can be for the study of stability of dynamical systems. In general, DDEs exhibit much more complicated dynamics than ordinary differential equations since a time delay could cause equilibrium states to change stability and produce a Hopf bifurcation near that equilibrium. Therefore, an analytic investigation of the stability of such models can be a very complicated task, especially when one has multiple time delays in the model. Since, most biological events have more than one time delay, numerical approaches to the stability analysis of such delay models is of utmost importance.

We first present a DDE model of the human immunodeficiency virus (HIV) [3]. As stated in this chapter, HIV attacks the immune system by focusing on the $CD4^+$ T lymphocytes. The virions bind to the membrane of the $CD4^+$ T cells and injects its own genetic material. After a time delay, this genetic material is replicated and many new virions are released. These new virions can infect susceptible $CD4^+$ T cells.

In this chapter, we first simplify the DDE model proposed in [3] by dropping the logistic growth factor and then investigate the effects of the delay on the virulence of the HIV strains. We derive an analytical expression of the ESS, and characterize how changes in delay could alter that evolutionary optimum. The existence and stability of the infected steady state are presented explicitly. We then investigate numerically the dynamics of the system by using numerical tools such that DDE-BITFOL and DDE23. The stability and bifurcation analysis of the steady state is again presented numerically by using DDE-BITFOL.

2 Model Equations

Here, we consider the well-known standard ordinary differential equation for HIV [3, 11], assuming that all the infected cells are capable of producing virus:

$$\frac{\mathrm{d}T}{\mathrm{d}t} = s - \mu_{\mathrm{T}} T - k_1 V T, \tag{1}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = k_2 V T - \mu_{\mathrm{I}} I, \tag{2}$$

$$\frac{\mathrm{d}V}{\mathrm{d}t} = N\mu_{\mathrm{b}} I - k_1 V T - \mu_{\mathrm{V}} V, \tag{3}$$

where $T(t)$ is the concentration of healthy $CD4^+$ T cells, $I(t)$ is the concentration infected $CD4^+$ T cells and $V(t)$ is the concentration of free HIV. All parameter descriptions and values are given in Table 1.

The corresponding DDEs, where a discrete time delay is introduced to represent the viral eclipse phase, is given as follows:

$$\frac{\mathrm{d}T}{\mathrm{d}t} = s - \mu_{\mathrm{T}} T - k_1 V T, \tag{4}$$

Table 1 Parameters for viral spread

Parameters	Description	Values
μ_{T}	Natural death rate of $CD4^+$ T cells	0.02 day^{-1}
μ_{I}	Blanket death rate of infected $CD4^+$ T cells	0.26 day^{-1}
μ_{b}	Lytic death rate for infected cells	0.24 day^{-1}
μ_{V}	Death rate of free virus	2.4 day^{-1}
k_1	Rate of $CD4^+$ T cell to become infected by virus	2.4×10^{-5} mm^{3}day^{-1}
k_2	Rate infected cells become active	2×10^{-5} mm^{3}day^{-1}
N	Number of virions produced by infected $CD4^+$ T cells	500
s	Source term for uninfected $CD4^+$ T cells	10 day^{-1}mm^{-3}
τ	Discrete time delay due to viral eclipse phase	Varies

$$\frac{\mathrm{d}I}{\mathrm{d}t} = k_2 V(t-\tau)T(t-\tau) - \mu_1 I, \tag{5}$$

$$\frac{\mathrm{d}V}{\mathrm{d}t} = N\mu_{\mathrm{b}} I - k_1 VT - \mu_{\mathrm{V}} V, \tag{6}$$

with the initial values

$$T(\theta) = T_0 = 1000 \text{ mm}^{-3},\ I(\theta) = 0 \text{ mm}^{-3},$$
$$V(\theta) = V_0 = 1000 \text{ mm}^{-3}, \quad \theta \in [-\tau, 0].$$

The description of the parameters and their values are given in Table 1.

3 Effects of Discrete Time Delay on the Virulence

3.1 The Basic Reproduction Number

The basic reproduction number, R_0, is defined as the expected number of hary cases produced by a single (typical) infection in a completely susceptible population. It is important to note that R_0 is a dimensionless number and not a rate, which would have units of time^{-1}. Some authors incorrectly call R_0 the basic reproductive rate. We can use the fact that R_0 is a dimensionless number to help us in calculating it, see [8].

$$R_0 \propto \left(\frac{\text{infection}}{\text{contact}}\right) \times \left(\frac{\text{contact}}{\text{time}}\right) \times \left(\frac{\text{time}}{\text{infection}}\right).$$

Note that R_0 is a dimensionless quantity. More specifically:

$$R_0 = \gamma \times \overline{c} \times d,$$

where γ is the transmissibility (i.e., probability of infection given contact between a susceptible and infected individual), $\overline{c}$ is the average rate of contact between susceptible and infected individuals, and d is the duration of infectiousness. If $R_0 > 1$, then

the disease will propagate, otherwise the disease will eventually die and a fraction of the population will escape infection.

3.2 Next-Generation Matrix Method

There are several methods of computing R_0. The most formal and most widely used approach is the next generation matrix approach. Many papers such as [6] and [8] provide a nice introduction for calculating R_0 using this method. The notation we use here follows their usage. Consider the next generation matrix G. It is composed of two parts: F and V^{-1}, where

$$F = \left[\frac{\partial F_i(x_0)}{\partial x_j} \right]$$

and

$$V = \left[\frac{\partial V_i(x_0)}{\partial x_j} \right].$$

The F_i are the new infections, while the V_i transfers of infections from one compartment to another. x_0 is the disease-free equilibrium state. R_0 is the dominant eigenvalue of the matrix $G = FV^{-1}$.

3.3 Evolutionary Stable Strategy of Virulence

It is instructive to think about epidemics from the pathogen's perspective. Pathogens bear biological information in their nucleic acids. This information varies from one copy of a pathogen to another, and the ability of a pathogen to persist and multiply can be a function of this variability [1, 8], known as virulence. In other word, virulence is the ability of the pathogen to transmit disease to a host. There is a trade-off between virulence and transmissibility. An increase in virulence will first lead to an increase in transmissibility, which consequently will lead to a faster weakening and may be death of the host. So, there are fewer possible contacts and shorter times to transmit. But more virulent strains may have higher instantaneous transmissibility rates that compensate for the fewer contacts. So, for a pathogen to invade a susceptible population there has to be a balance between transmissibility and virulence.

An evolutionary stable strategy (ESS) is a phenotype that cannot be invaded by a rare mutant. Loosely speaking, it represents the optimal phenotype. The ESS virulence occurs where the fitness gradient equals zero [8], meaning:

$$\frac{\mathrm{d}R_0}{\mathrm{d}x} = 0,$$

where x denotes the virulence.

3.4 Change in Selective Pressures

Previous work [2, 10] has proved that the direction of virulence evolution around an ESS as selective pressures change will be determined by the sign of the derivative of the fitness gradient with respect to the parameter that is changing. In other words, the virulence will increase (decrease) when we increase (decrease) a selected parameter n if:

$$\frac{\partial}{\partial n}\left[\frac{\partial R_0(x,n)}{\partial x}\right] > 0. \quad (< 0).$$

3.5 Effects of Discrete Time Delay on Virulence

To investigate the effects of the discrete time delay τ on the virulence, we compute the basic reproduction number R_0 (see Appendix):

$$R_0(x,\tau) = \frac{\log r_0(x)\tau}{\tau},$$

where

$$r_0(x) = \frac{N\mu_b k_2(x)}{\mu_1 \mu_V(x)}$$

is the basic reproduction number when there is no delay. x denotes the virulence, and $k_2(x)$ and $\mu_V(x)$ are the rate at which the infected cells become active and the death rate of free virions, respectively.

Theorem 1 *The ESS of the virulence is independent of the discrete time delay.*

Proof The fitness gradient of the system is given by:

$$\frac{\partial R_0(x,\tau)}{\partial x} = \frac{\frac{\mathrm{d}k_2(x)}{k_2(x)} - \frac{\mathrm{d}\mu_v(x)}{\mu_v(x)}}{\tau}. \tag{7}$$

The ESS virulence occurs where

$$\frac{\mathrm{d}R_0}{\mathrm{d}x} = 0,$$

that is, if and only if:

$$\frac{\mathrm{d}k_2(x)}{\mathrm{d}\mu_v(x)} = \frac{k_2(x^*)}{\mu_v(x^*)}, \tag{8}$$

where x^* denoted the ESS of x (virulence).
When there is no delay ($\tau = 0$) the fitness gradient is given by:

$$\frac{\mathrm{d}r_0(x)}{\mathrm{d}x} = \frac{N\mu_b\left[\mu_v k_2'(x) - k_2\mu_v'(x)\right]}{\mu_1\mu_v}, \tag{9}$$

and therefore the ESS occurs when

$$\mu_v k_2'(x) - k_2 \mu_v'(x) = 0,$$

which is equivalent to Eq. 8. □

Equation 8 has a nice geometric interpretation. The ESS virulence occurs where a line (L1) is tangent to the curve that relates k_2 to μ_v. This result is known as the marginal value theorem and has applications in economics and ecology as well as epidemiology.

Theorem 2 *The virulence of the HIV strain of the system (4)–(6) increases when we increase the discrete time delay τ due to the viral eclipse if and only if*

$$k_2 < \mu_v.$$

Proof The derivative of the fitness gradient (Eq. 7) with respect to τ is given as:

$$\frac{\partial}{\partial \tau}\left[\frac{\partial R_0(x,\tau)}{\partial x}\right] = -\frac{\frac{dk_2(x)}{k_2(x)} - \frac{d\mu_v(x)}{\mu_v(x)}}{\tau^2}$$

and

$$\frac{\partial}{\partial \tau}\left[\frac{\partial R_0(x,\tau)}{\partial x}\right] > 0,$$

if and only

$$\frac{dk_2(x)}{k_2(x)} < \frac{d\mu_v(x)}{\mu_v(x)}.$$

Take the integral of both sides and notice that $k_2(0) = \mu_v(0) = 0$, to obtain

$$k_2 < \mu_v.$$ □

3.6 Results

To illustrate the effects of the delay on the virulence of the infection, we compute numerically the solution of system (4)–(6) using MATLAB package DDE23 [13]. The parameter values are given in Table 1. The discrete time delay introduces a time shift, but has a minimal effect on the number of copies of uninfected $CD4^+$ T cell as shown in Fig. 1.

As we increase the delay, the virulence of the disease has a large increase and therefore the decrease in the number of copies of infected $CD4^+$ T cell, see Fig. 2, and the number of copies of free virions, see Fig. 3, which shows the results for $\tau = 5$ days.

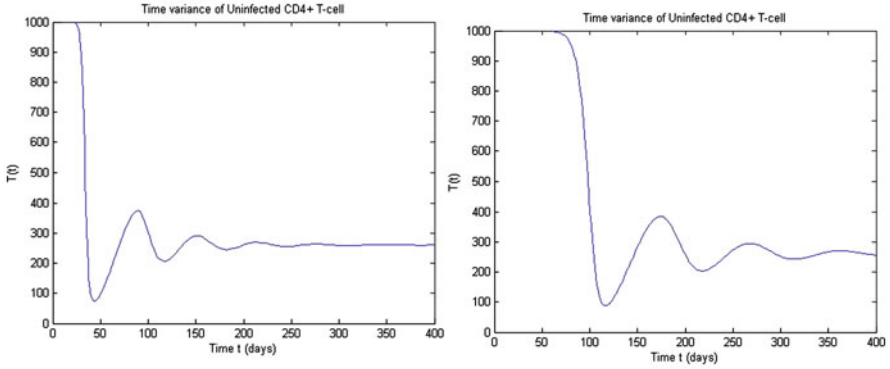


Fig. 1 Time variation of uninfected $CD4^+$ T cells without delay (*left*) and with delay $\tau = 5$ days (*right*)

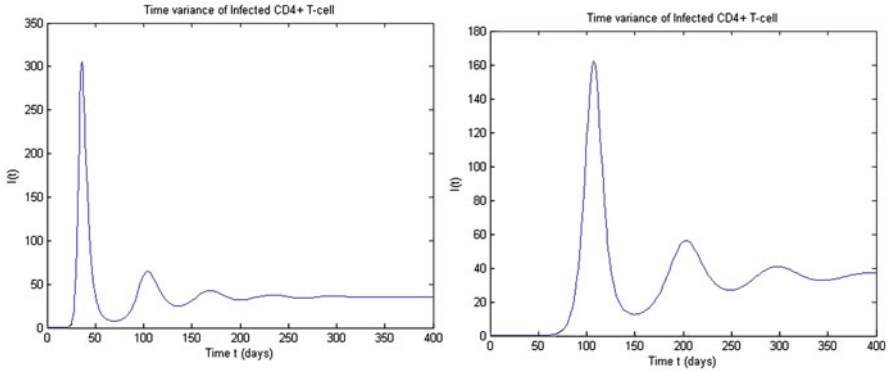


Fig. 2 Time variation of infected $CD4^+$ T cells without delay (*left*) and with delay $\tau = 5$ days (*right*)

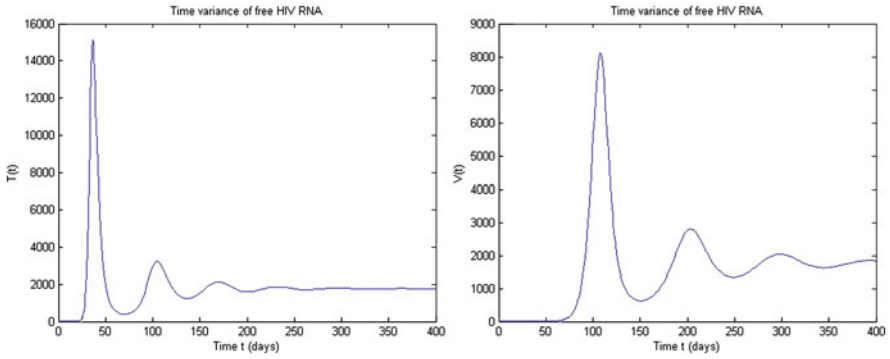


Fig. 3 Time variation of free HIV virions without delay(*left*) and with delay $\tau = 5$ days (*right*)

4 Effects of Discrete Time Delay on Stability

In this section, we investigate the effects of a discrete time delay on the stability of a steady state of a given dynamical system. To do so, we consider the same DDEs for a HIV infection (system (4)–(6) as in Sect. 3). We first describe the stability of the steady states of the system when there is no delay, i.e., $\tau = 0$, and then investigate the changes in stability as we introduce the delay. We also point out necessary and sufficient conditions for the delay to affect the stability of the steady states, and to introduce possible Hopf bifurcations.

4.1 Stability of the Ordinary Differential Model

In this subsection, we present necessary and sufficient conditions for the existence and stability of the steady states of the standard system (1)–(3).

Proposition 1

i) If $R \leq 1$, then the nonnegative steady state of the system (1)–(3) is $(T_0^*, I_0^*, V_0^*) = (\frac{s}{\mu_T}, 0, 0)$.

ii) If $R > 1$ and $\beta > 0$ (i.e., $N > N_{\text{crit}} = \frac{\mu_I(k_1 s + \mu_v \mu_T)}{\mu_b k_1 s}$), then the nonnegative steady states are: $(T_0^*, I_0^*, V_0^*) = (\frac{s}{\mu_T}, 0, 0)$, $(T_1^*, I_1^*, V_1^*) = \left(\frac{\mu_v \mu_I^2}{k_1(R-1)}, \frac{\beta \mu_I}{k_1(R-1)}, \frac{\beta}{k_1 \mu_v \mu_I}\right)$, where

$$R = \frac{N\mu_b}{\mu_I}, \quad \beta = (N\mu_b - \mu_I)k_1 s - \mu_v \mu_I \mu_T.$$

Notice that the threshold parameter R can be interpreted as the basic reproductive number. If $R > 1$, then the disease will spread into the system, otherwise for $R \leq 1$ the disease will eventually die. The parameter N is clearly a bifurcation parameter

Stability Analysis of the Equilibria
The Jacobian matrix of the model system evaluated at (T^*, I^*, V^*) is

$$J = \begin{bmatrix} -\mu_T - k_1 V^* & 0 & -k_1 T^* \\ k_2 V^* & -\mu_I & k_2 T^* \\ -k_1 V^* & N\mu_b & -k_1 T^* - \mu_v. \end{bmatrix}$$

We will study the stability of our model based on the eigenvalues of the Jacobian matrix.

Proposition 2

(i) If $R \leq 1$ *and* $\beta < 0$ *(i.e.,* $N < N_{\text{crit}} = \frac{\mu_{\text{I}}(k_1 s + \mu_{\text{v}}\mu_{\text{T}})}{\mu_{\text{b}} k_1 s}$*), then the steady state* $(T_0^*, I_0^*, V_0^*) = (\frac{s}{\mu_{\text{T}}}, 0, 0)$ *is stable.*

(ii) If $R \leq 1$ *and* $\beta > 0$*, then the steady state* (T_0^*, I_0^*, V_0^*) *is unstable.*

(iii) If $R > 1$*,* $\beta > 0$*,* $a_1 > 0$*,* $a_4 + a_5 > 0$ *and* $a_1(a_2 + a_3) - (a_4 + a_5) > 0$*, then the steady state* $(T_1^*, I_1^*, V_1^*) = \left(\frac{\mu_{\text{v}}\mu_{\text{I}}^2}{k_1(R-1)}, \frac{\beta\mu_{\text{I}}}{k_1(R-1)}, \frac{\beta}{k_1\mu_{\text{v}}\mu_{\text{I}}}\right)$ *is stable.*

Here,

$$a_1 := k_1(T^* + V^*) + \mu_{\text{v}} + \mu_{\text{I}} - \mu_{\text{T}} \tag{10}$$

$$a_2 := (\mu_{\text{I}} + k_1 V^*)k_1 T^* + \mu_{\text{I}}\mu_{\text{v}} \tag{11}$$

$$a_3 := N\mu_{\text{b}} k_2 T^* \tag{12}$$

$$a_4 := (\mu_{\text{T}} + k_1 V^*)N\mu_{\text{b}} k_1 k_2 T^* V^* \tag{13}$$

$$a_5 := (\mu_{\text{T}} + k_1 V^*)[(N\mu_{\text{b}} - k_2 V^* - \mu_{\text{I}})k_1 T^* - \mu_{\text{I}}\mu_{\text{v}}]. \tag{14}$$

Proof The characteristic equation of matrix J is given by:

$$\lambda^3 + a_1\lambda^2 + (a_2 + a_3)\lambda + (a_4 + a_5) = 0. \tag{15}$$

(i) In this case, we substitute the steady state (T_0^*, I_0^*, V_0^*) into Eq. 15 and find:

$$(\lambda + \mu_{\text{T}})(\lambda^2 + b_1\lambda + b_2) = 0,$$

with

$$b_1 := \frac{\mu_{\text{I}}\mu_{\text{T}} + k_1 s + \mu_{\text{v}}\mu_{\text{T}}}{\mu_{\text{T}}} \quad \textit{and} \quad b_2 := \frac{-\beta}{\mu_{\text{T}}}$$

- If $\beta < 0$, then $\zeta = b_1^2 - 4b_2 \leq 0$ and therefore the eigenvalues of J are: $\lambda_1 = -\mu_{\text{T}}$, $\lambda_2 = \frac{-b_1}{2} - \frac{\sqrt{\zeta}}{2}i$ and $\lambda_3 = \frac{-b_1}{2} + \frac{\sqrt{\zeta}}{2}i$. Thus, the steady state is stable.
- If $\beta > 0$, then $\zeta = b_1^2 - 4b_2 > 0$ and therefore the eigenvalues of J are: $\lambda_1 = -\mu_{\text{T}} < 0$, $\lambda_2 = \frac{-b_1}{2} - \frac{\sqrt{\zeta}}{2} < 0$ and $\lambda_3 = \frac{-b_1}{2} + \frac{\sqrt{\zeta}}{2} > 0$. Thus, the steady state is unstable.

(ii) Since $R > 1$ and $\beta > 0$, the steady state (T_1^*, I_1^*, V_1^*) exists. By the Routh–Hurwitz criterion [7, 12], it follows that all roots of the characteristic equation have negative real parts, if and only if, $a_1 > 0$, $a_4 + a_5 > 0$ and $a_1(a_2 + a_3) - (a_4 + a_5) > 0$. □

4.2 Stability of the DDE Model

Now, we consider the system (4)–(6), which is the delay model of HIV.

Notice that the delay system has the same steady states as the ordinary differential equation (ODE) model. To study the stability of those steady states, let us define solutions of the delay system of the form:

$$\begin{bmatrix} cT' \\ I' \\ V' \end{bmatrix} = e^{-\lambda\tau} \begin{bmatrix} cT \\ I \\ V \end{bmatrix}.$$

Then the Jacobian of the system is given by:

$$M = \begin{bmatrix} -\mu_{\mathrm{T}} - k_1 V^* & 0 & -k_1 T^* \\ e^{-\lambda\tau} k_2 V^* & -\mu_{\mathrm{I}} & e^{-\lambda\tau} k_2 T^* \\ -k_1 V^* & N\mu_{\mathrm{b}} & -k_1 T^* - \mu_{\mathrm{v}} \end{bmatrix}.$$

The characteristic equation of the DDE model is given by:

$$\lambda^3 + a_1\lambda^2 + a_2\lambda + a_3 e^{-\lambda\tau}\lambda + a_4 e^{-\lambda\tau} + a_5 = 0, \tag{16}$$

where $a_i, i = 1, ..., 5$ are defined in Eqs. 10–14.

Proposition 3 *The stability of the noninfected steady state does not depend on the delay. Therefore, the stability conditions of the steady state* (T_0^*, I_0^*, V_0^*) *remain the same as those given in Proposition 2*

Proof Notice that when we consider (T_0^*, I_0^*, V_0^*), then the coefficients $a_3 = a_4 = 0$ and then the characteristic equation of the DDE system becomes $(\lambda + \mu_T)(\lambda^2 + b_1\lambda + b_2) = 0$, for all $\tau > 0$. □

Recall that for the ODE model the steady state (T_1^*, I_1^*, V_1^*) is stable for the parameter values satisfying conditions in Proposition 3.2.1(ii). Here, we are interested in determining whether there exists a critical delay $\tau_c > 0$ so that $Re(\lambda) > 0$ for $\tau > \tau_c$. In other words, τ_c is the value of τ such that $Re(\lambda) = 0$, at which the transition from stability to instability occurs. For the steady state (T_1^*, I_1^*, V_1^*), if we let $\lambda(\tau) = \alpha(\tau) + \mathrm{i}\omega(\tau)$, where α and ω are real, then we have $\alpha(0) < 0$. Suppose $\alpha(\tau_c) = 0$ for some $\tau_c > 0$, then by the continuity in τ of α, $\alpha(\tau) < 0$ for values of τ such that $0 \leq \tau < \tau_c$. Therefore, the steady state remains stable for these values of τ.

If such $\tau_c > 0$ exists, with $\alpha(\tau_c) = 0$ and $\alpha(\tau) < 0$ for $0 \leq \tau < \tau_c$, then by Rouche's theorem (Dieudonne [4], Theorem 9.17.4), the steady state will lose stability at $\tau = \tau_c$. In fact such τ_c exists if and only there exists $\omega(\tau_c) > 0$ such that

$\lambda(\tau_c) = i\omega(\tau_c) = i\omega_c$ is a root of the characteristic equation 16. That is

$$-i\omega_c^3 - a_1\omega_c^2 + a_2 i\omega_c + a_5 + (a_4 + a_3 i\omega_c)(\cos\omega_c\tau_c - i\sin\omega_c\tau_c) = 0.$$

Equating real parts and imaginary parts of the equation to zero, one obtains:

$$a_1\omega_c^2 - a_5 = a_4\cos\omega_c\tau_c + a_3\omega_c\sin\omega_c\tau_c, \tag{17}$$

$$-\omega_c^3 + a_2\omega_c = a_4\sin\omega_c\tau_c - a_3\omega_c\cos\omega_c\tau_c. \tag{18}$$

Adding up the squares of equations 17 and 18, one obtains

$$u(\omega_c) := \omega_c^6 + (a_1^2 - 2a_2)\omega_c^4 + (a_2^2 - 2a_1a_5 - a_3^2)\omega_c^2 + a_5^2 - a_4^2 = 0. \tag{19}$$

For simplicity, we introduce the quantities

$$z := \omega_c^2, \quad p := a_1^2 - 2a_2, \quad q := a_2^2 - 2a_1a_5 - a_3^2, \quad r := a_5^2 - a_4^2.$$

Then Eq. 19 reduces to

$$K(z) := z^3 + pz^2 + qz + r = 0. \tag{20}$$

□

Lemma 1 *Suppose that Eq. 20 has no positive roots. Then all the roots of the characteristic equation have negative real parts for all* $\tau > 0$.

We present conditions that ensure that Eq. 20 has a positive root or has no positive roots.

$$K'(z) = 3z^2 + 2pz + q = 0$$

has the roots:

$$Z_0 := \frac{-p + \sqrt{p^2 - 3q}}{3}, \quad Z_1 := \frac{-p - \sqrt{p^2 - 3q}}{3}.$$

Lemma 2

(i) If either (a) $r < 0$, *or (b)* $r \geq 0, \quad p^2 - 3q > 0, \quad p < 0$ *and* $K(Z_0) < 0$, *then Eq. 20 has a positive root.*

(ii) If $r \geq 0$ *and* $p^2 - 3q < 0$, *then Eq. 20 has no positive roots.*

Proof

(i) Suppose that condition (a) holds, that is, $r < 0$. Then we have $K(0) = r < 0$. On the other hand, since

$$\lim_{z\to+\infty} K(z) = \infty,$$

by the intermediate value theorem Eq. 20 must have a positive root z_0, that is, $K(z_0) = 0$. Now suppose that condition (b) holds. Since $r \geq 0$, $p < 0$, and

$p^2 - 3q > 0$, we find that Z_0 is real and $Z_0 > 0$. Since $K(0) = r \geq 0$ and $k(Z_0) < 0$, again by the intermediate value theorem, K has a zero between the origin and Z_0.

(ii) Since $p^2 - 3q < 0$, both zeros Z_0 and Z_1 are not real. That is, $K'(z) = 0$ has no real root. Noting that

$$K'(0) = q > \frac{p^2}{3} \geq 0$$

we conclude that the quadratic polynomial K' is strictly positive on the real numbers. This implies that K is increasing on the real numbers. Moreover, since $K(0) = r \geq 0$, we observe that $K(z)$ does not vanish for $z > 0$ and thus Eq. 20 has no positive roots.

Notice that Lemma 2(ii) implies that there is no positive ω such that $i\omega$ is a solution of the characteristic equation 16. Therefore, the real parts of all the eigenvalues of (16) are negative for all delay $\tau \geq 0$. □

Next, we will provide the conditions on the parameters to ensure that Hopf bifurcation occurs. Suppose conditions in Lemma 2(i) hold, then Eq. 20 has a positive root. We denote, without loss of generality the positive roots of (20) by m_j, $j \in \{0, 1, 2\}$ depending on the number of positive roots (20) has. Equation 19, therefore has at most six roots, $\pm\sqrt{m_j}$ for $j = 0, 1, 2$.

If the solution of Eq. 19 exists, it is among these $\pm\sqrt{m_j}$ for $j = 0, 1, 2$. If $\lambda = i\omega$ is a root of Eq. 16 so is $-i\omega$.

Substituting ω_j into Eqs. 17 and 18 and solving for τ, we obtain

$$\tau_j^{(n)} = \frac{1}{\omega_j} \arccos \frac{a_3\omega_j^4 + (a_1a_4 - a_2a_3)\omega_j^2 - a_4a_5}{a_4^2 + a_3^2\omega_j^2} + \frac{2n\pi}{\omega_j},$$

where $j = 0, 1, 2$ and $n = 0, 1, 2, \ldots$
Now, let $\tau_c > 0$ be the smallest of such τ for which $\alpha(\tau_c) = 0$. Thus,

$$\tau_c = \min\tau_j^{(n)} > 0, \quad 0 \leq j \leq 2, n \geq 1, \quad \omega_c = \omega_{jc} \tag{21}$$

Theorem 3 *For the time lag τ, let the critical time lag τ_c and ω_c be defined as in (21), and suppose that $(E_2E_3 - E_1E_4)\sin\omega_c\tau_c - (E_2E_4 + E_1E_3)\cos\omega_c\tau_c \neq 0$ then the system of DDEs (4)–(6) exhibits a Hopf bifurcation at the steady state (T_1^*, I_1^*, V_1^*), with*

$$E_1 := a_3\sin\omega_c\tau_c - 2a_1\omega_c, \quad E_2 := a_3\cos\omega_c\tau_c + a_2 - 3\omega_c^2,$$

$$E_3 := a_4\omega_c, \quad E_4 := a_3\omega_c^2.$$

Proof We will show that

$$\frac{d\alpha(\tau)}{d\tau}|_{\tau=\tau_c} \neq 0,$$

which guarantees that the Hopf bifurcation occurs. First, we equate real parts and imaginary parts of the characteristic equation to zero:

$$\alpha^3 - 3\alpha\omega^2 + a_1\alpha^2 - a_1\omega^2 + a_2\alpha + a_5 + e^{-\alpha\tau}[(\alpha\cos\omega\tau + \omega\sin\omega\tau)a_3 + a_4\cos\omega\tau] = 0, \tag{22}$$

$$3\alpha^2\omega - \omega^3 + 2a_1\alpha\omega + a_2\omega + e^{-\alpha\tau}[(\omega\cos\omega\tau - \alpha\sin\omega\tau)a_3 - a_4\sin\omega\tau] = 0. \tag{23}$$

We differentiate Eqs. 22 and 23 with respect to τ and evaluate at $\tau = \tau_c$ for which $\alpha(\tau_c) = 0$ and $\omega(\tau_c) = \omega_c$. We then obtain

$$E_1\frac{d\omega(\tau)}{d\tau}\Big|_{\tau=\tau_c} + E_2\frac{d\alpha(\tau)}{d\tau}\Big|_{\tau=\tau_c} = E_3\sin\omega_c\tau_c - E_4\cos\omega_c\tau_c, \tag{24}$$

$$E_2\frac{d\omega(\tau)}{d\tau}\Big|_{\tau=\tau_c} - E_1\frac{d\alpha(\tau)}{d\tau}\Big|_{\tau=\tau_c} = E_3\cos\omega_c\tau_c + E_4\sin\omega_c\tau_c. \tag{25}$$

By solving Eqs. 24 and 25, we obtain

$$\frac{d\alpha(\tau)}{d\tau}|_{\tau=\tau_c} = \frac{(E_2E_3 - E_1E_4)\sin\omega_c\tau_c - (E_2E_4 + E_1E_3)\cos\omega_c\tau_c}{E_1^2 + E_2^2} \neq 0.$$

Hence, the Hopf bifurcation occurs when τ passes through the critical value τ_c. □

5 Numerical Methods

Using DDE-BIFTOOL [5], we will examine the stability and the bifurcation process of the steady state $(T_1^*, I_1^*, V_1^*) = (\frac{\mu_v\mu_I^2}{k_1(R-1)}, \frac{\beta\mu_I}{k_1(R-1)}, \frac{\beta}{k_1\mu_v\mu_I})$. We compute the eigenvalues of the characteristic equation 16, and display their real parts versus imaginary parts as shown in Fig. 4.

All eigenvalues have negative real part, therefore the steady state (T_1^*, I_1^*, V_1^*) is stable for those values of τ. But as we increase the delay, there are eigenvalues with positive real part, see Fig. 5. So, there exists a critical delay τ_c such that the steady state is destabilized (some of the eigenvalues of the characteristic Eq. 16 have strictly positive real parts) as τ passes through τ_c.

Figure 6 shows the existence of a pair of pure imaginary eigenvalues where there is a Hopf bifurcation, and also of a real eigenvalue where we have a fold or turning point bifurcation.

One can plot the time lag τ versus the rate of infection of the CD4$^+$ T cells with free virus, and notice that as τ passes through the critical delay τ_c the steady state is destabilized through a second Hopf bifurcation branch, see Fig. 7.

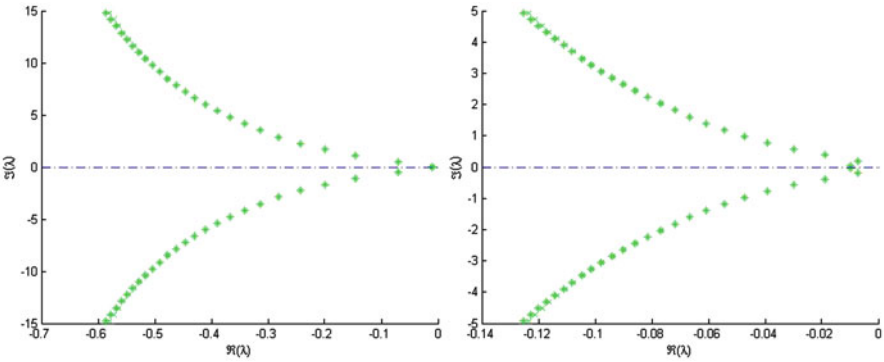


Fig. 4 Roots of the characteristic equation 16 with $\tau = 10$ days (*left*) and $\tau = 15$ days (*right*)

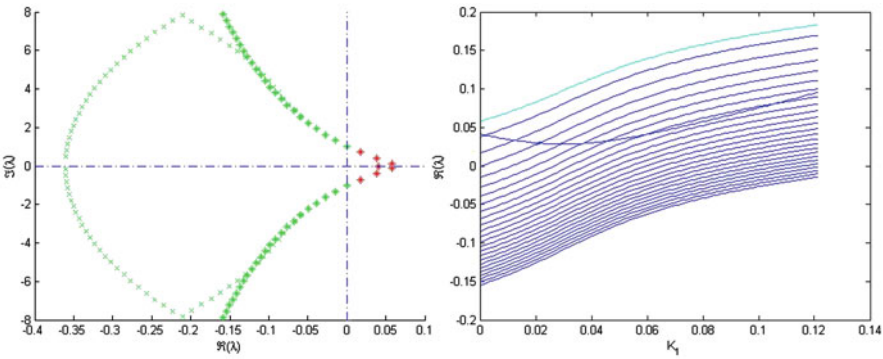


Fig. 5 Roots of the characteristic equation 16 with $\tau = 20$ days (*left*), and real part vs. k_1 (*right*)

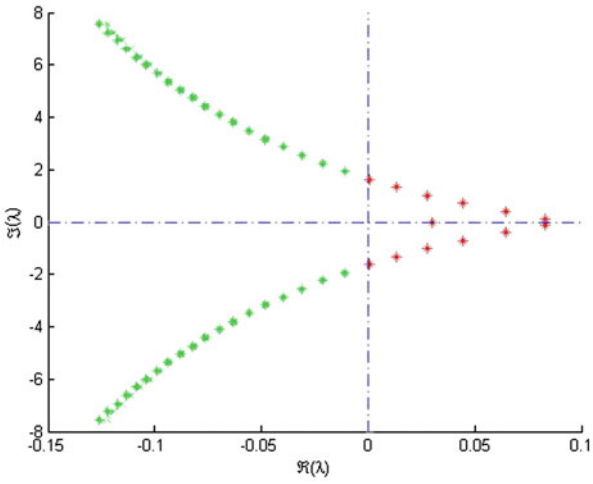


Fig. 6 A pair of pure eigenvalues is clearly visible (Hopf bifurcation) and also a real eigenvalue (turning point or fold bifurcation)

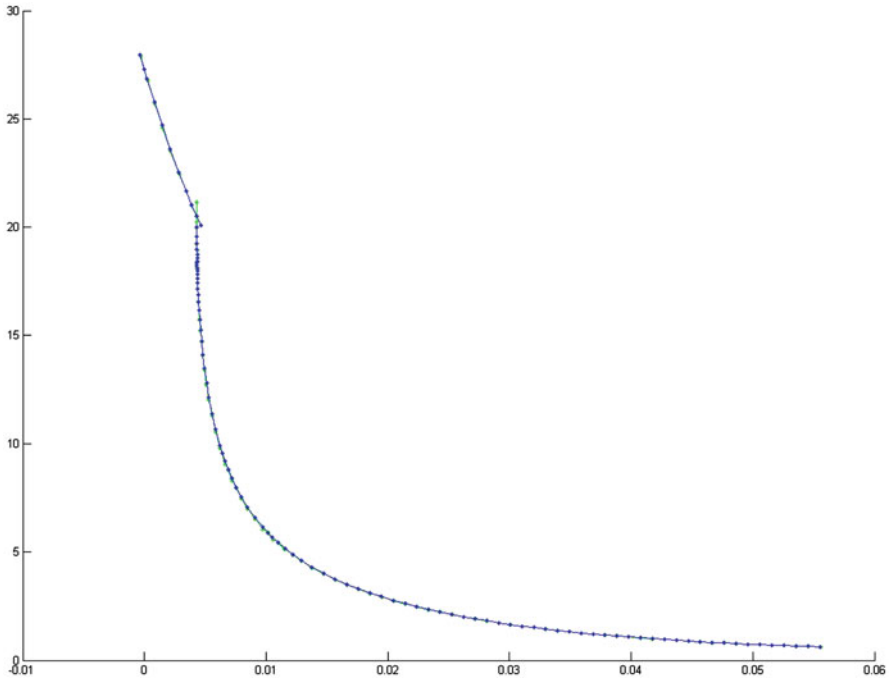


Fig. 7 Hopf bifurcation branches: τ vs. k_1

6 Conclusions

We investigated the effects of a discrete time delay on disease progression of an HIV pandemic. We first investigated the effects of the delay on the virulence of the HIV strains. We derived an analytical expression of the ESS, and characterized how changes in delay could alter that evolutionary optimum. Our analysis showed that the ESS of the HIV strains does not depend on the delay; however, the virulence of the HIV strains may increase as a consequence of increasing the delay time. We then presented an analytic stability analysis of the endemically infected equilibrium similar to the analysis done by [3]. We presented a novel numerical analysis of the stability and bifurcation process of the same equilibrium using numerical tools. With the numerical methods we were able to reach the same conclusion as the analytic version.

Appendix

In this appendix, we compute the basic reproductive number by the method of next generation matrix. System (4)–(6) has the matrix of newly raised infections:

$$F = \begin{bmatrix} 0 & k_2 e^{-\lambda\tau} \\ 0 & 0 \end{bmatrix} \tag{26}$$

and the matrix of transferred infections:

$$V = \begin{bmatrix} \mu_1 & 0 \\ -N\mu_\mathrm{b} & \mu_\mathrm{v} \end{bmatrix}. \tag{27}$$

The next generation matrix is

$$FV^{-1} = \begin{bmatrix} e^{-\lambda\tau}\frac{k_2 N\mu_\mathrm{b}}{\mu_1\mu_\mathrm{b}} & e^{-\lambda\tau}\frac{k_2}{\mu_v} \\ 0 & 0 \end{bmatrix}, \tag{28}$$

which has the characteristic equation

$$\lambda\left(\lambda - e^{-\lambda\tau}\frac{k_2 N\mu_\mathrm{b}}{\mu_1\mu_\mathrm{v}}\right) = 0. \tag{29}$$

If $\tau = 0$, then the dominant eigenvalue is

$$r_0 = \frac{k_2 N\mu_\mathrm{b}}{\mu_1\mu_\mathrm{v}} \tag{30}$$

If $\tau > 0$, then the dominant eigenvalue is

$$R_0 = \frac{\ln r_0\tau}{\tau}. \tag{31}$$

References

1. Baalen, M.V., Sabelis, M.: The dynamics of multiple infection and the evolution of virulence. Am. Nat. **146**, 881–910 (1995)
2. Basu, S., Galvani, A.P.: The evolution of tuberculosis virulence. Bul. Math. Bio. **71**, 1073–1088 (2009)
3. Culshaw, R.V., Ruan, S.: A delay-differential equation model of HIV infection of $CD4^+$ T-cells, Math. Biosci. **165**, 27–39 (2000)
4. Dieudonne, J.: Foundations of modern analysis. Academic, New York (1960)
5. Engelborghs, K., Luzyanina, T., Samaey, G.: DDE-BIFTOOL v. 2.00: a Matlab package for bifurcation analysis of delay differential equations, Report TW 330, NY (2001)

6. Hefferman, J., Smith, R., Wahl, L.: Perspectives on the basic reproduction ratio. J.R. Soc. Interface **2**, 281–293 (2000)
7. Hurwitz, A.: On the conditions under which an equation has only roots with negative real parts. Selected Papers on Mathematical Trends in Control Theory. Dover, New York (1964)
8. Jones, J.H.: Notes on R_0. Department of Anthropological Sciences, Stanford University, Stanford (2007)
9. Kuang, Y.: Delay differential equations with applications in population dynamics. Academic, San Diego (1993)
10. Otto, S.P., Day, T.: A Biologist's guide to mathematical modeling in ecology and evolution. Princeton University Press, Princeton (2007)
11. Perelson, A.S., Nelson, P.W.: Mathematical analysis of HIV-1 dynamics in vivo. SIAM Rev. **41**, 3–44 (1999)
12. Routh, E.J.: A treatise on the stability of a given state of motion, particularly steady motion. Macmillan, London (1977)
13. Shampine, L.F., Thompson, S.: Solving DDEs in MATLAB, Appl. Numer. Math. **37**, 441–458 (2001)

On the Riemann Problem for a Hyperbolic System of Temple Class

Richard A. De la cruz Guerrero and Juan C. Juajibioy

Abstract In this chapter, we study the one-dimensional Riemann problem for a hyperbolic system of three conservation laws of temple class. Under suitable generalized Rankine–Hugoniot relation and entropy condition, both existence and uniqueness of particular delta-shock type solutions are established. Moreover, we show explicitly the solution of generalized Riemann problem.

Keywords Temple class · Linearly degenerate fields · Riemann problem · Generalized Riemann problem · Delta shock solution

1 Introduction

The modeling of viscoelastic materials and fluids is important for many applications.

In [6], the authors introduced a new system of conservation laws that models shallow viscoelastic fluids. This new system is motivated in [6, Eq. (5.6)] and is written as:

$$\begin{cases} \rho_t + (\rho u)_x = 0, \\ (\rho u)_t + (\rho u^2 + \pi)_x = 0, \\ (\rho \frac{\pi}{s^2})_t + (\rho u \frac{\pi}{s^2} + u)_x = 0, \\ s_t + u s_x = 0, \\ c_t + u c_x = 0, \end{cases} \tag{A}$$

where ρ denotes the layer depth of fluid, u is the horizontal velocity, s is related to the stress tensor and it is a conserved quantity, π is the relaxed pressure, and $c > 0$ is

R. A. De la cruz Guerrero (✉)
School of Mathematics and Statistics, Universidad Pedagógica y Tecnológica de Colombia - UPTC, Tunja, Colombia
e-mail: richard.delacruz@uptc.edu.co

J. C. Juajibioy
Department of Mathematics, Universidad Nacional de Colombia - UN, Bogotá, Colombia
e-mail: jcjuajibioyo@unal.edu.co

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_5

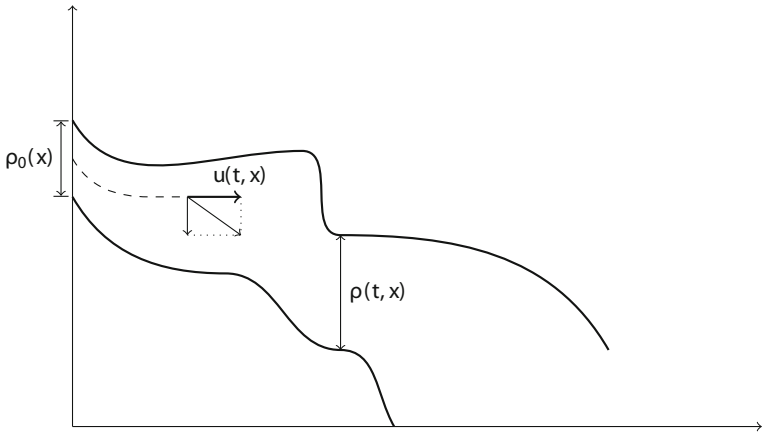


Fig. 1 Simplified viscoelastic shallow fluid model

introduced in order to parameterize the speeds. This system describes a simple model for a thin layer of non-Newtonian viscoelastic fluid over a given topography at the bottom when the movement is driven by gravitational forces such as geophysical flows (mud flows, landslides, debris avalanches).

In [19], since s is a conserved quantity, the author considers the case $s = constant > 0$. Moreover, the field c does not appear in the first four equations and we remove it. We consider $s = constant > 0$ and introducing the new variable $v = \frac{\pi}{s^2}$ to simplify the system (A) in the following:

$$\begin{cases} \rho_t + (\rho u)_x = 0, \\ (\rho u)_t + (\rho u^2 + s^2 v)_x = 0, \\ (\rho v)_t + (\rho u v + u)_x = 0. \end{cases} \tag{1}$$

We refer to the system above as the *Suliciu relaxation system* [5, 9, 24] and correspond the case homogeneous with constant stress tensor of the model proposed by Bouchut and Boyaval, i.e., it is a simplified viscoelastic shallow fluid model.

Also, this system can be considered as a relaxation for the isentropic Chaplygin gas dynamics system:

$$\begin{cases} \rho_t + (\rho u)_x = 0, \\ (\rho u)_t + (\rho u^2 + P)_x = 0, \end{cases}$$

where ρ and u, respectively, stand for the density and the velocity of the gas, while the pressure P is given by the state equation $P(\rho) = -\frac{s^2}{\rho}$ with $s = \text{constant} > 0$.

One of the main difficulties of the system (1) is to obtain existence and uniqueness of solutions of Cauchy problems in the presence of vacuum regions (where the layer deep $\rho = 0$). The existence of global weak solutions including vacuum regions,

was obtained in [19] using compensated compactness. The results on existence and uniqueness for general rich type and temple class system can be found in [2–4, 7, 13, 20, 21]. However, some of these results do not apply to (1) since it has all fields being linearly degenerate and the initial data may have oscillations.

The Riemann problem for the Suliciu relaxation system has been extensively studied, for instance in [5, 10]. However, it seems to us that they did not consider all possible arrays. Now, we propose delta wave solutions type for the Suliciu relaxation system. In this chapter, we construct the Riemann solution for the system focusing our attention on delta shock waves of certain type and the solution of the generalized Riemann problem. The existence and uniqueness of solutions involving delta shock waves can be obtained by solving the generalized Rankine–Hugoniot relation under an entropy condition [12, 16].

2 Properties of the Suliciu Relaxation System and Some Assumptions

The eigenvalues associated to the system (1) are given by,

$$\lambda_1 = u - s/\rho, \qquad \lambda_2 = u \quad \text{and} \quad \lambda_3 = u + s/\rho, \tag{2}$$

where the corresponding Riemann invariants are

$$R_1 = s^2 v - su, \qquad R_2 = v + 1/\rho, \quad \text{and} \quad R_3 = s^2 v + su. \tag{3}$$

It is easy to see that system (1) is linearly degenerate. Moreover, we have that for each $i, j, k \in \{1, 2, 3\}$ with $j \neq i, k \neq i$, it holds

$$\frac{\partial}{\partial R_j}\left(\frac{1}{\lambda_k - \lambda_i}\frac{\partial \lambda_i}{\partial R_k}\right) = \frac{\partial}{\partial R_k}\left(\frac{1}{\lambda_j - \lambda_i}\frac{\partial \lambda_i}{\partial R_j}\right). \tag{4}$$

This means that system (1) is of *rich* type [22]. In this chapter, we focus on the study of the Suliciu relaxation system of conservation laws (1) with bounded initial data:

$$\begin{aligned} &(\rho(0,x), u(0,x), v(0,x)) = (\rho_0(x), u_0(x), v_0(x)), \quad x \in \mathbb{R} \\ &\rho_0(x) \geq \underline{\rho} = \text{constant} > 0, \end{aligned} \tag{5}$$

subject to the following conditions:

H1: The functions ρ_0, u_0, and v_0 satisfy

$$\begin{aligned} &c_1 \leq u_0(x) - s v_0(x) \leq c_2, \qquad c_3 \leq u_0(x) + s v_0(x) \leq c_4 \\ &\text{and } v_0(x) + \frac{1}{\rho_0(x)} > c_5, \end{aligned}$$

where c_i, $i = 1, \ldots, 5$, are suitable constants satisfying $c_5 - \frac{c_4 - c_1}{2s} > 0$.

H2: The total variations of $u_0(x) - sv_0(x)$ and $u_0(x) + sv_0(x)$ are bounded.

The conditions H1 and H2 are somehow natural to impose since they ensure that ρ is positive giving a physical meaning to the Suliciu relaxation system (1). All entropies associated to (1) are of the form,

$$\eta(\rho, u, v) = \rho \left(F(u + sv) + G(u - sv) + H(v + 1/\rho)\right), \tag{6}$$

where F, G, H are arbitrary functions having entropy flux:

$$q(\rho, u, v) = (\rho u + s)F(u + sv) + (\rho u - s)G(u - sv) + \rho u H(v + 1/\rho). \tag{7}$$

Moreover, if the functions F, G, and H are convex, then, the entropy is also convex (see [19, Theorem 2]). Thus, from each convex pair (η, q), we have the following condition:

$$\eta_t(\rho, u, v) + q_x(\rho, u, v) = 0 \tag{8}$$

in the sense of distributions.

3 Riemann Problem

In this section, we study the solution for the Riemann problem associated with the Suliciu relaxation system with initial data:

$$(\rho, u, v)(0, x) = \begin{cases} (\rho_{\mathrm{r}}, u_{\mathrm{r}}, v_{\mathrm{r}}), & \text{if } x > 0, \\ (\rho_{\mathrm{l}}, u_{\mathrm{l}}, v_{\mathrm{l}}), & \text{if } x < 0, \end{cases} \tag{9}$$

in which left and right constant states $(\rho_{\mathrm{l}}, u_{\mathrm{l}}, v_{\mathrm{l}})$ and $(\rho_{\mathrm{r}}, u_{\mathrm{r}}, v_{\mathrm{r}})$, respectively, satisfy the conditions H1, H2 and $\lambda_1(\rho_{\mathrm{l}}, u_{\mathrm{l}}, v_{\mathrm{l}}) < \lambda_3(\rho_{\mathrm{r}}, u_{\mathrm{r}}, v_{\mathrm{r}})$.

Consider the self-similar solution $(\rho, u, v)(t, x) = (\rho, u, v)(\xi)$, $\xi = \frac{x}{t}$, for which the system (1) becomes

$$\begin{cases} -\xi\rho_\xi + (\rho u)_\xi = 0, \\ -\xi(\rho u)_\xi + (\rho u^2 + s^2 v)_\xi = 0, \\ -\xi(\rho v)_\xi + (\rho u v + u)_\xi = 0, \end{cases} \tag{10}$$

and initial data (9) changes to the boundary condition

$$(\rho, u, v)(-\infty) = (\rho_{\mathrm{l}}, u_{\mathrm{l}}, v_{\mathrm{l}}) \text{ and } (\rho, u, v)(+\infty) = (\rho_{\mathrm{r}}, u_{\mathrm{r}}, v_{\mathrm{r}}). \tag{11}$$

This is a two-point boundary value problem of first-order ordinary differential equations with the boundary values in the infinity. For smooth solution, (10) is reduced to

$$\begin{pmatrix} u - \xi & \rho & 0 \\ 0 & \rho(u - \xi) & 0 \\ 0 & 1 & \rho(u - \xi) \end{pmatrix} \begin{pmatrix} \rho \\ u \\ v \end{pmatrix}_\xi = 0. \tag{12}$$

It provides either the general solutions (constant states) $(\rho, u, v) = constant$ $(\rho > 0)$ or singular solutions

$$\begin{aligned} \xi = \lambda_1 = u - s/\rho, \quad & d\,(u - s/\rho) = 0 \text{ and } d\,(v + 1/\rho) = 0, \\ \xi = \lambda_2 = u, \quad & du = 0 \text{ and } dv = 0, \\ \xi = \lambda_3 = u + s/\rho, \quad & d\,(u + s/\rho) = 0 \text{ and } d\,(v + 1/\rho) = 0. \end{aligned} \tag{13}$$

Integrating (13) from (ρ_l, u_l, v_l) to (ρ, u, v), one can get that

$$\begin{aligned} \xi = \lambda_1 = u - s/\rho, \quad & u - s/\rho = u_l - s/\rho_l \text{ and } v + 1/\rho = v_l + 1/\rho_l, \\ \xi = \lambda_2 = u, \quad & u = u_l \text{ and } v = v_l, \\ \xi = \lambda_3 = u + s/\rho, \quad & u + s/\rho = u_l + s/\rho_l \text{ and } v + 1/\rho = v_l + 1/\rho_l. \end{aligned} \tag{14}$$

For a bounded discontinuity at $\xi = \omega$, the Rankine–Hugoniot conditions hold,

$$\begin{cases} -\omega[\rho] + [\rho u] = 0, \\ -\omega[\rho u] + [\rho u^2 + s^2 v] = 0, \\ -\omega[\rho v] + [\rho u v + u] = 0, \end{cases} \tag{15}$$

where $[q] = q_l - q$ is the jump of q across the discontinuous line and ω is the velocity of the discontinuity. From (15), we have

$$\begin{aligned} \omega = u - s/\rho, \quad & u - s/\rho = u_l - s/\rho_l \text{ and } v + 1/\rho = v_l + 1/\rho_l, \\ \omega = u, \quad & u = u_l \text{ and } v = v_l, \\ \omega = u + s/\rho, \quad & u + s/\rho = u_l + s/\rho_l \text{ and } v + 1/\rho = v_l + 1/\rho_l. \end{aligned} \tag{16}$$

From (14) and (16), we conclude that the rarefaction and shock waves are coincident [25], which correspond to contact discontinuities [8]. Namely, for a given left state (ρ_l, u_l, v_l), the contact discontinuity curves, which are the sets of states that can be connected on the right by a 1-contact discontinuity J_1, a 2-contact discontinuity J_2, or a 3-contact discontinuity J_3, are as follows:

$$\begin{aligned} J_1 : (\rho, u, v) &:= (\rho, u_l - s/\rho_l + s/\rho, v_l + 1/\rho_l - 1/\rho), \\ J_2 : (\rho, u, v) &:= (\rho, u_l, v_l), \\ J_3 : (\rho, u, v) &:= (\rho, u_l + s/\rho_l - s/\rho, v_l + 1/\rho_l - 1/\rho), \quad \rho > 0. \end{aligned} \tag{17}$$

In the space $(\rho > 0, u \in \mathbb{R}, u \in \mathbb{R})$, through the point (ρ_l, u_l, v_l), we draw curves (17) which are denoted by J_1, J_2, and J_3 respectively. So, J_1 has asymptotes $\rho = 0$ and $(\rho, u_l - s/\rho_l, v_l + 1/\rho_l)$ for $\rho \geq 0$, and J_3 has asymptotes $\rho = 0$ and $(\rho, u_l + s/\rho_l, v_l + 1/\rho_l)$.

In order to solve the Riemann problem (1)–(9), we consider left and right constant states $U_l = (\rho_l, u_l, v_l)$ and $U_r = (\rho_r, u_r, v_r)$, respectively, such that the conditions H1–H2 are satisfied and $\lambda_1(U_l) < \lambda_3(U_r)$. Then exist intermediate states,

$U_* = (\rho_*, u_*, v_*)$ and $U_{**} = (\rho_{**}, u_{**}, v_{**})$ such that $U_* = J_1(\sigma_1)(U_l)$, $U_{**} = J_2(\sigma_2)(U_*)$ and $U_r = J_3(\sigma_3)(U_{**})$, for some $\sigma_1, \sigma_2, \sigma_3$.

Furthermore, because of (17), the states U^*, U^{**} should satisfy

$$u_* = (u_l - s/\rho_l) + s/\rho_*, \quad v_* = (v_l + 1/\rho_l) - 1/\rho_*, \tag{18a}$$

$$u_* = u_{**}, \quad v_* = v_{**}, \tag{18b}$$

$$u_{**} = (u_r + s/\rho_r) - s/\rho_{**} \text{ and } v_{**} = (v_r + 1/\rho_r) - 1/\rho_{**}. \tag{18c}$$

From Eq. (19), we have

$$1/\rho_{**} - 1/\rho_* = (v_r + 1/\rho_r) - (v_r + 1/\rho_l) \tag{19}$$

and

$$1/\rho_{**} + 1/\rho_* = \{u_r + s/\rho_r) - (u_l - s/\rho_l)\}/s. \tag{20}$$

Observe that by conditions H1 and H2, we have that U_* and U_{**} also satisfies H1, H2.

This guarantees that ρ_* and ρ_{**} are positive.

Note that $\lambda_1(U_l) < \lambda_3(U_r)$, implies $|R_2(U_r) - R_2(U_l)| < \frac{1}{s}(\lambda_3(U_r) - \lambda_1(U_l))$.

Additionally, as usual, since the system is linearly degenerate, $\lambda_1(U_l) = \lambda_1(U_*)$, $\lambda_2(U_*) = \lambda_2(U_{**})$, and $\lambda_3(U_{**}) = \lambda_3(U_r)$.

The results of this section can be summarized in the following theorem.

Theorem 1 *Given left and right constant states (ρ_l, u_l, v_l) and (ρ_r, u_r, v_r), respectively, such that they satisfy conditions H1, H2 and $\lambda_1(\rho_l, u_l, v_l) < \lambda_3(\rho_r, u_r, v_r)$. Then, there is a unique global solution to the Riemann problem (1)–(9). Moreover, this solution is given by*

$$(\rho, u, v)(t, x) = \begin{cases} (\rho_l, u_l, v_l), & \text{if } x < \lambda_1(\rho_l, u_l, v_l)t, \\ (\rho_*, u_*, v_*), & \text{if } \lambda_1(\rho_l, u_l, v_l)t < x < \lambda_2(\rho_{**}, u_{**}, v_{**})t, \\ (\rho_{**}, u_{**}, v_{**}), & \text{if } \lambda_2(\rho_{**}, u_{**}, v_{**})t < x < \lambda_3(\rho_r, u_r, v_r)t, \\ (\rho_r, u_r, v_r), & \text{if } x > \lambda_3(\rho_r, u_r, v_r)t, \end{cases} \tag{21}$$

where

$$\frac{1}{\rho_*} = \frac{1}{2s}(u_r - u_l) - \frac{1}{2}(v_r - v_l) + \frac{1}{\rho_l}, \qquad \frac{1}{\rho_{**}} = \frac{1}{2s}(u_r - u_l) + \frac{1}{2}(v_r - v_l) + \frac{1}{\rho},$$

$$u_* = \frac{1}{2}\{(u_l + sv_l) + (u_r - sv_r)\} = u_{**} \text{ and } v_* = \frac{1}{2s}\{(u_l + sv_l) - (u_r - sv_r)\} = v_{**}.$$

Remark 1 (Explicit solutions) Observe that using the Euler–Lagrange (E–L) transformation [17, 20, 26, 27], $(t,x) \to (t,y) = (t, Y(t,x))$, defined by

$$dy = \rho\, dx - \rho u\, dt \quad \text{and} \quad Y(0,x) = Y_0(x) \overset{\text{def}}{=} \int_0^x \rho_0(\xi)\, d\xi,$$

the system (1) in Lagrangian coordinates becomes

$$\begin{cases} \omega_t - v_y = 0, \\ v_t + s^2 \kappa_y = 0, \\ \kappa_t + v_y = 0, \end{cases} \tag{22}$$

where $\omega(t,y)$ denotes the quantity $\frac{1}{\rho(t,x)}$ in Lagrangian coordinates, $v(t,y) = u(t,x)$ and $\kappa(t,y) = v(t,x)$. The eigenvalues associated to (22) are given by

$$\widetilde{\lambda}_1 = -s, \quad \widetilde{\lambda}_2 = 0, \quad \widetilde{\lambda}_3 = s, \tag{23}$$

and the corresponding Riemann invariants are given by $R_1 = s^2\kappa - sv$, $R_2 = v + \omega$, and $R_3 = s^2\kappa - sv$. Also, the entropy condition (8) transforms into $\widetilde{\eta}_t(\omega, v, \kappa) + \widetilde{q}_x(\omega, v, \kappa) = 0$, for each $\widetilde{\eta}$ with $\widetilde{\eta}(\omega, v, \kappa) = F(v + s\kappa) + G(v - s\kappa) + H(\omega + \kappa)$ and $\widetilde{q}(\omega, v, \kappa) = sF(v + s\kappa) - sG(v - s\kappa)$ where F, G, H are (arbitrary) convex functions. The initial conditions (5) becomes

$$\begin{cases} (\omega(0,y), v(0,y), \kappa(0,y)) = (\omega_0(y), v_0(y), \kappa_0(y)), \quad y \in \mathbb{R}, \\ \omega_0(y) \geq \underline{\omega} > 0. \end{cases} \tag{24}$$

The explicit solution of the corresponding Cauchy problem (22)–(24) is

$$\begin{aligned} \omega(t,y) &= \omega_0(y) + \kappa_0(y) - \kappa(t,y), \\ v(t,y) &= (v_0(y+st) + v_0(y-st))/2 - s(\kappa_0(y+st) - \kappa_0(y-st))/2, \\ \kappa(t,y) &= (\kappa_0(y+st) + \kappa_0(y-st))/2 - (v_0(y+st) - v_0(y-st))/2s. \end{aligned} \tag{25}$$

Moreover, by condition H1 we obtain that $c_1 \leq v(t,y) - s\kappa(t,y) \leq c_2$, $c_3 \leq v(t,y) + s\kappa(t,y) \leq c_4$ and $\omega(t,y) + \kappa(t,y) > c_5$, and since $\rho_0(x) \geq \underline{\rho} = \text{constant} > 0$ by (5), we have that $\omega(t,y) \geq \underline{\omega} > 0$, ensuring that the function $y \mapsto X(t,y)$ is invertible and bi-Lipschitzian from $\mathbb{R}$ to $\mathbb{R}$ for all $t \geq 0$.

Therefore, we consider $X_0 = Y_0^{-1}$. Then, the unique function $x = X(t,y)$ that satisfy $X(0,y) = X_0(y)$ is given by

$$\begin{aligned} X(t,y) = &\frac{1}{2s} \int_{y-st}^{y+st} u_0(X_0(\xi))\, d\xi + \int_0^y \left(v_0(X_0(\xi)) + \frac{1}{\rho_0(X_0(\xi))} \right) d\xi \\ &- \frac{1}{2} \int_0^{y+st} v_0(X_0(\xi))\, d\xi - \frac{1}{2} \int_0^{y-st} v_0(X_0(\xi))\, d\xi. \end{aligned} \tag{26}$$

From the above, we obtain the following theorem.

Theorem 2 *Assume that $\rho_0, u_0, v_0 \in L^\infty(\mathbb{R})$ with $\rho_0(x) \geq \underline{\rho} > 0$, the conditions H1, H2 hold and*

$$\inf_{x\in\mathbb{R}} \left(u_0(x) + \frac{s}{\rho_0(x)} \right) > \sup_{x\in\mathbb{R}} \left(u_0(x) - \frac{s}{\rho_0(x)} \right). \tag{27}$$

Then, the Cauchy problem (1)–(5) has an unique global solution $(\rho, u, v) \in L^\infty(\mathbb{R}^+ \times \mathbb{R})$ that satisfies the entropy condition (8) for all pair (η, q) defined in (6) and (7). Moreover, this solution is given by

$$\rho(t,x) = \frac{\rho_0(X_0(Y(t,x)))}{1 + \rho_0(X_0(Y(t,x)))\,[v_0(X_0(Y(t,x))) - v(t,x)]},$$

$$u(t,x) = \Gamma^+_{u_0}(t,x) - s\Gamma^-_{v_0}(t,x) \text{ and } v(t,x) = \Gamma^+_{v_0}(t,x) - \frac{1}{s}\Gamma^-_{u_0}(t,x),$$

where $\Gamma^\pm_{G_0}(t,x) = \frac{1}{2}\left[G_0(X_0(Y(t,x)+st)) \pm G_0(X_0(Y(t,x)-st))\right]$.

Usually, the condition (27) guarantees existence and uniqueness of solutions in linearly degenerate systems [11, 23].

4 Delta Shock Solution

Now, we discuss the solution for the Riemann problem associated with the Suliciu relaxation system, in which left and right constant states (ρ_l, u_l, v_l) and (ρ_r, u_r, v_r), respectively, satisfy the conditions H1 and H2, but unlike previous section they satisfy $\lambda_1(\rho_l, u_l, v_l) \geq \lambda_3(\rho_r, u_r, v_r)$.

Let $s \in \mathbb{R}$, the Sobolev space $H^s(\mathbb{R})$ is the collection of all temperate distributions f such that $(1+\xi^2)^{s/2}\widehat{f} \in L^2(\mathbb{R}, d\xi)$. Moreover, the Sobolev space $H^s(\mathbb{R})$ is a Hilbert space with respect to the inner product

$$(f,g)_s = \int_{\mathbb{R}} (1+\xi^2)^s \widehat{f}(\xi)\overline{\widehat{g}(\xi)}\, d\xi.$$

Properties on Sobolev spaces can be found in [1, 14, 18].

Denote by $BM(\mathbb{R})$ the space of bounded Borel measures on $\mathbb{R}$. Then, the definition of a measure solution of Suliciu relaxation system in $BM(\mathbb{R})$ can be given as follows.

Definition 1 A triple (ρ, u, v) constitutes a *measure solution* to the Suliciu relaxation system, if it holds that

1. $\rho \in L^\infty((0,\infty), BM(\mathbb{R})) \cap C((0,\infty), H^{-s}(\mathbb{R}))$,
2. $u \in L^\infty((0,\infty), L^\infty(\mathbb{R})) \cap C((0,\infty), H^{-s}(\mathbb{R}))$,
3. $v \in L^\infty_{loc}((0,\infty), L^\infty_{loc}(\mathbb{R})) \cap C((0,\infty), H^{-s}(\mathbb{R}))$, $s > 0$,
4. u and v are measurable with respect to ρ at almost for all $t \in (0,\infty)$,

and

$$\begin{cases} I_1 = \int_0^\infty \int_{\mathbb{R}} (\phi_t + u\phi_x)\, d\rho dt = 0, \\ I_2 = \int_0^\infty \int_{\mathbb{R}} u(\phi_t + u\phi_x)\, d\rho dt + \int_0^\infty \int_{\mathbb{R}} s^2 v\phi_x\, dx dt = 0, \\ I_3 = \int_0^\infty \int_{\mathbb{R}} v(\phi_t + u\phi_x)\, d\rho dt + \int_0^\infty \int_{\mathbb{R}} u\phi_x\, dx dt = 0, \end{cases} \tag{28}$$

for all test function $\phi \in C_0^\infty(\mathbb{R}^+ \times \mathbb{R})$.

Definition 2 A two-dimensional weighted delta function $w(s)\delta_L$ supported on a smooth curve L parameterized as $t = t(s)$, $x = x(s)$ $(c \le s \le d)$ is defined by

$$\langle w(s)\delta_L, \phi(t,x)\rangle = \int_c^d w(s)\phi(t(s), x(s))\, ds \tag{29}$$

for all $\phi \in C_0^\infty(\mathbb{R}^2)$.

Definition 3 A triple distribution (ρ, u, v) is called a *delta shock wave* if it is represented in the form

$$(\rho,u,v)(t,x) = \begin{cases} (\rho_1, u_1, v_1)(t,x), & x < x(t), \\ (w(t)\delta(x - x(t)), u_\delta(t), g(t)), & x = x(t), \\ (\rho_r, u_r, v_r)(t,x), & x > x(t), \end{cases} \tag{30}$$

and satisfies Definition 1, where $(\rho_1, u_1, v_1)(t,x)$ and $(\rho_r, u_r, v_r)(t,x)$ are piecewise smooth bounded solutions of the Suliciu relaxation system (1).

We set $\frac{dx}{dt} = u_\delta(t)$ since the concentration in ρ needs to travel at the speed of discontinuity. Hence, we say that a delta shock wave (30) is a measure solution to the Suliciu relaxation system (1), if and only if, the following relation holds:

$$\begin{cases} \frac{dx(t)}{dt} = u_\delta(t), \\ \frac{dw(t)}{dt} = -[\rho]u_\delta(t) + [\rho u], \\ \frac{dw(t)u_\delta(t)}{dt} = -[\rho u]u_\delta(t) + [\rho u^2 + s^2 v], \\ \frac{dw(t)g(t)}{dt} = -[\rho v]u_\delta(t) + [\rho uv + u]. \end{cases} \tag{31}$$

In fact, for any test function $\phi \in C_0^\infty(\mathbb{R}^+ \times \mathbb{R})$, from (28), we obtain

$$I_1 = \int_0^\infty \int_{\mathbb{R}} (\phi_t + u\phi_x)\, d\rho dt = \int_0^\infty \left\{ -u_\delta(t)[\rho] + [\rho u] - \frac{dw(t)}{dt} \right\} \phi\, dt,$$

$$I_2 = \int_0^\infty \left\{ -u_\delta(t)[\rho u] + [\rho u^2 + s^2 v] - \frac{dw(t)u_\delta(t)}{dt} \right\} dt, \quad \text{and}$$

$$I_3 = \int_0^\infty \left\{ -u_\delta(t)[\rho v] + [\rho uv + u] - \frac{dw(t)g(t)}{dt} \right\} \phi\, dt.$$

The relation (31) are called the generalized Rankine–Hugoniot relation.

In addition, to guarantee uniqueness, the delta shock wave should satisfy the admissibility (entropy) condition:

$$\lambda_3(\rho_r, u_r, v_r) \leq u_\delta(t) \leq \lambda_1(\rho_l, u_l, v_l). \tag{32}$$

Now, the generalized Rankine–Hugoniot relation is applied to the Riemann problem (1)–(9) with left and right constant states $U_l = (\rho_l, u_l, v_l)$ and $U_r = (\rho_r, u_r, v_r)$, respectively, satisfying the conditions H1, H2, the fact $\lambda_3(\rho_r, u_r, v_r) \leq \lambda_1(\rho_l, u_l, v_l)$ and

$$\begin{cases} u_l - u_r \geq \max\{-\frac{s^2}{\rho_l}(v_l - v_r), \frac{s^2}{\rho_r}(v_l - v_r)\}, & \text{if } u_l - u_r \geq 1, \\ (u_l - u_r)^2 \geq \max\{-\frac{s^2}{\rho_l}(v_l - v_r), \frac{s^2}{\rho_r}(v_l - v_r)\}, & \text{if } u_l - u_r \leq 1. \end{cases} \tag{33}$$

Thereby, the Riemann problem is reduced to solving (31) with initial data $x(0) = 0$, $w(0) = 0$, $g(0) = 0$, under entropy condition $u_r + \frac{s}{\rho_r} \leq u_\delta(t) \leq u_l - \frac{s}{\rho_l}$. From it follows that

$$\begin{aligned} w(t) &= -[\rho]x(t) + [\rho u]t, \\ w(t)u_\delta(t) &= -[\rho u]x(t) + [\rho u^2 + s^2 v]t, \text{ and} \\ w(t)g(t) &= -[\rho v]x(t) + [\rho uv + u]t. \end{aligned} \tag{34}$$

Multiplying the first equation in (34) by $u_\delta(t)$ and then subtracting it from the second one, we obtain that $[\rho]x^2(t) - 2[\rho u]x(t)t + [\rho u^2 + s^2 v]t^2 = 0$.

One can find $u_\delta(t) := u_\delta$ is a constant and $x(t) = u_\delta t$. So it can be rewritten in

$$[\rho]u_\delta^2 - 2[\rho u]u_\delta + [\rho u^2 + s^2 v] = 0. \tag{35}$$

When $[\rho] = \rho_l - \rho_r = 0$, the situation is very simple and one can easily calculate the solution

$$\begin{cases} u_\delta = \frac{u_l + u_r}{2} + s^2 \frac{[v]}{2\rho_l[u]}, \\ x(t) = u_\delta t, \\ w(t) = \rho_l(u_l - u_r)t, \\ g(t) = \frac{[\rho uv + u] - u_\delta}{[\rho u]}, \end{cases} \tag{36}$$

which obviously satisfies the entropy condition (32).

When $[\rho] = \rho_l - \rho_r \neq 0$, the discriminant of the quadratic equation (35) is

$$\Delta = 4[\rho u]^2 - 4[\rho][\rho u^2 + s^2 v] = \rho_l \rho_r [u]^2 - s^2[\rho][v] > 0 \tag{37}$$

and with the help of the entropy condition (32), we can find the admissible solution is

$$\begin{cases} u_\delta = \frac{[\rho u] - \sqrt{[\rho u]^2 - [\rho][\rho u^2 + s^2 v]}}{[\rho]}, \\ x(t) = \frac{[\rho u] - \sqrt{[\rho u]^2 - [\rho][\rho u^2 + s^2 v]}}{[\rho]} t, \\ w(t) = \sqrt{[\rho u]^2 - [\rho][\rho u^2 + s^2 v]}\, t, \\ g(t) = \frac{-[\rho u][\rho v] + [\rho v]\sqrt{[\rho u]^2 - [\rho][\rho u^2 + s^2 v]} + [\rho][\rho uv + u]}{[\rho]\sqrt{[\rho u]^2 - [\rho][\rho u^2 + s^2 v]}} t. \end{cases} \tag{38}$$

Thus, we have proved the following result.

Theorem 3 *Given left and right constant states* (ρ_l, u_l, v_l) *and* (ρ_r, u_r, v_r)*, respectively, such that satisfy the conditions H1, H2,* $\lambda_1(\rho_l, u_l, v_l) \geq \lambda_3(\rho_r, u_r, v_r)$ *and (33).*

Then, the Riemann problem (1)–(9) admits a unique entropy solution in the sense of measures. This solution is of the form

$$(\rho, u, v)(t, x) = \begin{cases} (\rho_l, u_l, v_l), & \text{if } x < u_\delta t, \\ (w(t)\delta(x - u_\delta t), u_\delta, g(t)), & \text{if } x = u_\delta t, \\ (\rho_r, u_r, v_r), & \text{if } x > u_\delta t, \end{cases} \tag{39}$$

where u_δ*,* $w(t)$*, and* $g(t)$ *are show in (36) for* $[\rho] = 0$ *or 38 for* $[\rho] \neq 0$.

The above result includes the array $\frac{1}{s}(\lambda_3(U_r) - \lambda_1(U_l)) - (R_2(U_r) - R_2(U_l)) = 0$ or $\frac{1}{s}(\lambda_3(U_r) - \lambda_1(U_l)) + (R_2(U_r) - R_2(U_l)) = 0$.

5 Generalized Riemann Problem

In this section, we show explicitly the solution of the generalized Riemann problem. We also calculate the first-order expansion given by Lefloch and Raviart [15]. Consider the Suliciu relaxation system in Lagrangian coordinates (22) with initial data

$$(\omega, v, \kappa)(0, y) = \begin{cases} (\omega_L, v_L, \kappa_L)(y), & \text{if } y < 0, \\ (\omega_R, v_R, \kappa_R)(y), & \text{if } y > 0, \end{cases} \tag{40}$$

where $\omega_i(y), v_i(y), \kappa_i(y)$, for $i = L, R$, are piecewise smooth functions but discontinuous at $y = 0$. Moreover, we consider that functions ω_i, v_i, κ_i, for $i = L, R$, satisfies conditions H1, H2 and $\sup_y (v(0, y) - s\omega(0, y)) < \inf_y (v(0, y) + s\omega(0, y))$.

Let $(\omega_i^0, v_i^0, \kappa_i^0) = (\omega_i, v_i, \kappa_i)(0)$ for $i = L, R$. Then by Sect. 3, the classical Riemann problem for (22) with initial data

$$(\omega, v, \kappa)(0, y) = \begin{cases} (\omega_L^0, v_L^0, \kappa_L^0), & \text{if } y < 0, \\ (\omega_R^0, v_R^0, \kappa_R^0), & \text{if } y > 0, \end{cases} \tag{41}$$

has an entropy weak solution $(\omega^0, v^0, \kappa^0)(t, y)$ which is self-similar and consists of four constant states separated by contact discontinuities,

$$(\omega^0, v^0, \kappa^0)(t, y) = \begin{cases} (\omega_L^0, v_L^0, \kappa_L^0), & \text{if } y < -st, \\ (\omega_*^0, v_*^0, \kappa_*^0), & \text{if } -st < y < 0, \\ (\omega_{**}^0, v_{**}^0, \kappa_{**}^0), & \text{if } 0 < y < st, \\ (\omega_R^0, v_R^0, \kappa_R^0), & \text{if } y > st. \end{cases} \tag{42}$$

On the other hand, the solution of the generalized Riemann problem is

$$(\omega,\nu,\kappa)(t,y)=\begin{cases}(\omega_L,\nu_L,\kappa_L)(t,y), & \text{if } y<-st,\\ (\omega_*,\nu_*,\kappa_*)(t,y), & \text{if } -st<y<0,\\ (\omega_{**},\nu_{**},\kappa_{**})(t,y), & \text{if } 0<y<st,\\ (\omega_R,\nu_R,\kappa_R)(t,y), & \text{if } y>st.\end{cases} \tag{43}$$

where

$$\begin{aligned}\omega_i(t,y)&=\omega_i(y)+\kappa_i(y)-\kappa_i(t,y),\\ \nu_i(t,y)&=(\nu_i(y+st)+\nu_i(y-st))/2-s(\kappa_i(y+st)-\kappa_i(y-st))/2,\\ \kappa_i(t,y)&=(\kappa_i(y+st)+\kappa_i(y-st))/2-(\nu_i(y+st)\\ &\quad-\nu_i(y-st))/2s,\ \text{for } i=L \text{ or } R,\end{aligned}$$

$$\begin{aligned}\omega_*(t,y)&=(\nu_R(t,y)-\nu_L(t,y))/2s-(\kappa_R(t,y)-\kappa_L(t,y))/2+\omega_L(t,y),\\ \omega_{**}(t,y)&=(\nu_R(t,y)-\nu_L(t,y))/2s+(\kappa_R(t,y)-\kappa_L(t,y))/2+\omega_R(t,y),\\ \nu_*(t,y)&=(\nu_R(t,y)+\nu_L(t,y))/2-s(\kappa_R(t,y)-\kappa_L(t,y))/2=\nu_{**}(t,y),\\ \kappa_*(t,y)&=(\kappa_R(t,y)+\kappa_L(t,y))/2-(\nu_R(t,y)-\nu_L(t,y))/2s=\kappa_{**}(t,y).\end{aligned}$$

5.1 Asymptotic Expansion of LeFloch–Raviart

For smooth solutions for the generalized Riemann problem, consider the Taylor expansions $\omega_i(y)=\omega_i^0+\sum_{j=1}^{\infty}\omega_i^j y^j$, $\nu_i(y)=\nu_i^0+\sum_{j=1}^{\infty}\nu_i^j y^j$ and $\kappa_i(y)=\kappa_i^0+\sum_{j=1}^{\infty}\kappa_i^j y^j$, $i=L$ or R. Then, by the asymptotic expansion of LeFloch–Raviart, for the first order, we obtain that

$$\begin{cases}\omega_i(t,y)\approx\omega_i^0+(y\omega_i^1+t\nu_i^1),\\ \nu_i(t,y)\approx\nu_i^0+(y\nu_i^1-s^2t\kappa_i^1),\\ \kappa_i(t,y)\approx\kappa_i^0+(y\kappa_i^1-t\nu_i^1),\ \text{for } i=L \text{ or } R,\end{cases} \tag{44}$$

$$\omega_*(t,y)\approx\omega_*^0+y(\omega_L^1+\kappa_L^1)-\Phi^-(t,y)/s, \tag{45}$$

$$\omega_{**}(t,y)\approx\omega_{**}^0+y(\omega_R^1+\kappa_R^1)-\Phi^-(t,y)/s, \tag{46}$$

$$\nu_*(t,y)=\nu_{**}(t,y)\approx\nu_*^0+\Phi^+(t,y), \tag{47}$$

$$\kappa_*(t,y)=\kappa_{**}(t,y)\approx\kappa_*^0+\Phi^-(t,y)/s, \tag{48}$$

where

$$\Phi^{\pm}(t,y) = \left[(y-st)(v_L^1 + s\kappa_L^1) \pm (y+st)(v_R^1 - s\kappa_R^1)\right]/2.$$

Note that for smooth solutions, the first order of exact solution evaluate in $y = 0$, $(\omega, v, \kappa)(t, 0)$, coincides with the expansion of Lefloch–Raviart.

6 Conclusions

In previous works, the classical Riemann problem for the Suliciu relaxation system was solved [5, 9]. In this chapter, we show the unique entropy solution of the Riemann problem associated to the Suliciu relaxation system under assumptions of the conditions H1 and H2. First, we analyze the case $\lambda_1(\rho_l, u_l, v_l) < \lambda_3(\rho_r, u_r, v_r)$ and found unique solution in L^∞. For the case $\lambda_1(\rho_l, u_l, v_l) \geq \lambda_3(\rho_r, u_r, v_r)$, we show that the existing delta shock wave solution under a entropy condition guarantees the uniqueness of the solution. Finally, we show explicitly the solution of the generalized Riemann problem and calculate the first-order expansion of Lefloch–Raviart.

Acknowledgement We would like to thank Professor Yunguang Lu for suggesting this problem, pointing out the simplification of the system proposed by Bouchut and Boyaval and also for communicating to us the preprint [19]. Also, we are grateful to Profesor Philippe G. LeFloch for bringing to our attention the references [9, 10]. Lastly, our sincere thanks to Professors Juan Galvis and Leonardo Rendon for his suggestions and various talks we had with them on the Riemann problem for general systems which included nonclassical solutions.

References

1. Adams, R.: Sobolev spaces. Academic Press (1975)
2. Baiti, P., Bressan, A.: The semigroup generated by a Temple class system with large data. Diff. Integr. Equat. **10**, 401–418 (1997)
3. Bianchini, S.: The semigroup generated by a Temple class system with non-convex flux function. Diff. Integr. Equat. **13**, 1529–1550 (2000)
4. Bianchini, S.: Stability of L^∞ solutions for hyperbolic systems with coinciding Shocks and Rarefactions. SIAM J. Math. Anal. **33**(4), 959–981 (2000)
5. Bouchut, F.: Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources (Front. Math.). Birkhäuser Verlag, Basel (2004)
6. Bouchut, F., Boyaval, S.: A new model for shallow viscoelastic fluids. Math. Models Methods Appl. Sci. **23** 1479 (2013). doi:10.1142/S0218202513500140
7. Bressan, A., Goatin, P.: Stability of L^∞ solutions of temple class systems. Diff. Integr. Equat. **13**, 1503–1528 (2000)
8. Bressan, A.: Hyperbolic systems of conservation laws: The one-dimensional Cauchy problem. Oxford Lectures Series in Mathematics and its Applications 20, Oxford University Press (2005)
9. Carbou, G., Hanouzet, B., Natalini, R.: Semilinear behavior for totally linearly degenerate hyperbolic systems with relaxation. J. Diff. Equat. **246**(1), 291–319 (2009)
10. Chalons, C., Coquel, F.: Navier-Stokes equations with several independent pressure laws and explicit predictor–corrector schemes. Numer. Math. **101**(3), 451–478 (2005)

11. Chen, G.-Q.: The method of quasidecoupling for discontinuous solutions to conservation laws. Arch. Ration. Mech. Anal. **121**, 131–185 (1992)
12. Danilov, V.G., Shelkovich, V.M.: Delta-shock wave type solution of hyperbolic systems of conservation laws. Q. Appl. Math. **63**(3), 401–427 (2005)
13. Heibig, A.: Existence and uniqueness of solutions for some hyperbolic systems of conservation laws. Arch. Ration. Mech. Anal. **126**, 79–101 (1994)
14. Kesavan, S.: Topics in functional analysis and applications. John Wiley & Sons (1989)
15. LeFloch, P., Raviart, P.A.: An asymptotic expansion for the solution of the generalized Riemann problem. Part I: General theory. Ann. Inst. H. Poincare Anal. Non Lineaire. **5**(2), 179–207 (1988)
16. Li, J.Q., Zhang, T.: Generalized Rankine–Hugoniot relations of delta-shocks in solutions of transportation equations. In: Advance in nonlinear PDE and related areas, pp. 219–232. World Scientific, Singapore (1998)
17. Li, T., Peng, Y.-J., Ruiz, J.: Entropy solutions for linearly degenerate hyperbolic systems of rich type. J. Math. Pures Appl. **91**, 553–568 (2009)
18. Linares, F., Ponce, G.: Introduction to nonlinear dispersive equations. Universitext. Springer (2009)
19. Lu, Y-G., Klingenberg, C., Rendón, L. and Zheng, D-Y.: Global Solutions for a Simplified Shallow Elastic Fluids Model. Abstract and Applied Analysis, p 5, (2014) doi: 10.1155/2014/920248. (Vol. 2014, Article ID 920248)
20. Peng, Y.-J.: Euler-Lagrange change of variables in conservation laws. Nonlinearity. **20**, 1927–1953 (2007)
21. Serre, D.: Solutions à variation bornée pour certains systèmes hyperboliques de lois de conservation. J. Diff. Equat. **67**, 137–168 (1983)
22. Serre, D.: Richness and the classification of quasilinear hyperbolic systems. IMA Preprint Series # 597, 137–168 (1989)
23. Serre, D.: Intégrabilité d'une classe de système de lois de conservation. Forum Math. **4**, 607–623 (1992)
24. Suliciu, I.: On modelling phase transition by means of rate-type constitutive equations. Shock wave structure. Int. J. Eng. Sci. **28**(8), 829–841 (1990)
25. Temple, B.: System of conservation laws with invariant submanifolds. Trans. A.M.S. **280**, 781–795 (1983)
26. Wagner, D. H.: Equivalence of the Euler and Lagrangian equations of gas dynamics for weak solutions. J. Diff. Equat. **68**, 118–136 (1987)
27. Weinan, E., Kohn, R. V.: The initial-value problem for measure-valued solutions of a canonical 2×2 system with linearly degenerate fields. Comm. Pure Appl. Math. **44**, 981–1000 (1991)

Consequences of Weak Allee Effect in a Leslie–Gower-Type Predator–Prey Model with a Generalized Holling Type III Functional Response

Paulo C. Tintinago-Ruíz, Leonardo D. Restrepo-Alape and Eduardo González-Olivares

Abstract In this work, we analyze a predator–prey model derived from the Leslie–Gower type model considering two modifications: a generalized Holling type III functional response and a weak Allee 430054755 effect on prey, which is described by an autonomous bidimensional ordinary differential equation system. Conditions for the existence of the equilibrium points or singularities and their nature are determined. The existence of separatrix curves on the phase plane dividing the behavior of the trajectories are also shown. Thus, two closed solutions but in different sides of this separatrix curve can have different ω-limit sets; therefore, there exist trajectories highly sensitive to initial conditions. The existence of constraints on the parameter values for which the unique equilibrium point at the first quadrant is unstable and surrounded by a unique limit cycle in the phase plane is also proven. Computer simulations are also given in order to support our conclusions.

Keywords Limit cycles · Stability · Separatrix curve · Predator–prey model · Allee effect · Functional response · AMS classification (2000): 34C07, 37B25, 92D25

1 Introduction

In this work, a predator–prey model described by an autonomous bidimensional differential equation system is analyzed, considering the following aspects in the interaction:

P. C. Tintinago-Ruíz (✉)
Universidad del Quindío, Quindío, Colombia
e-mail: tinti27@gmail.com

L. D. Restrepo-Alape
Grupo Gedes, Universidad del Quindío, Quindío, Colombia
e-mail: ldrestrepo@uniquindio.edu.co

E. González-Olivares
Grupo de Ecología Matemática, Pontificia Universidad Católica de Valparaiso, Valparaiso, Chile
e-mail: ejgonzal@ucv.cl

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_6

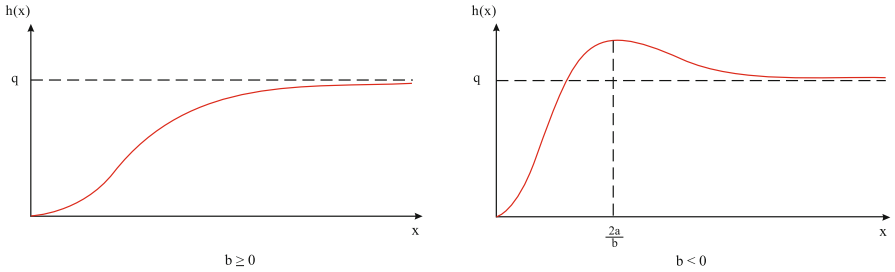


Fig. 1 Generalized Holling type III functional response when $b \geq 0$ (*left* poster) and $b < 0$ (*right* poster)

1. The predators growth function is of logistic type [30].
2. The functional response or predator consumption rate is a generalized Holling III type [18, 19].
3. The prey population is affected by the Allee effect.

The first aspect characterizes the Leslie-type predator–prey models, also named logistic predator–prey model [30] or Leslie–Gower model [3], in which, the predator environmental carrying capacity K_y is a function of prey population size x, i.e., it depends on the available resources. Here, we consider that $K_y = K(x) = nx$, proportional to prey abundance as in the May–Holling–Tanner model [2, 27]. Leslie model can leads to anomalies in their predictions [30], because it predicts that even at very low prey density, predator population can nevertheless increase, when predator prey ratio is very small and the consumption rate by individual predator is essentially zero [30]. However, these models are recently employed to study some predator–prey interactions [12, 16].

The predator functional response or consumption function refers to the change in attacked prey per unit of time per predator when the prey population size changes [9, 23, 26]. In this work, we have considered that the predator functional response is expressed by the function $h(x) = \frac{qx^2}{x^2+bx+a}$, presented in [19, 20]. This functional response can have different behaviors according to the sign of the parameter b, as is shown in Fig. 1.

We note when $b < 0$, the function $h(x)$ is a nonmonotonic functional response [29] representing the phenomenon of group defence formation [19, 26, 29] with a maximum at $x = \frac{2a}{b}$; however, in this work, we consider only the case when $b > 0$, being the function $h(x)$ sigmoid asymptotically monotonic increasing . This type of functional response has been recently used in an interesting work, considering a Leslie–Gower model in which the Allee effect is absent [18].

Sigmoid functional responses may arise from a variety of mechanisms, one of which is switching to alternative food sources [22] on predator–prey interaction [30], but this is only verifiable in the Volterra model [25] that sigmoid may stabilize an unstable equilibrium [29], but we show this does not happen in the Leslie–Gower model [28]. Many marine mammals appear to be generalist predators, and theory

would predict that they have a sigmoid functional response tending to stabilize prey populations [12].

On the other hand, any mechanism leading to a positive relationship between a component of individual fitness and the number or density of conspecifics can be regarded as a mechanism of the *Allee effect* [7]. This phenomenon has been also called *depensation* in fisheries sciences [21], or *negative competition effect, inverse density dependence, positive density dependence,* and various other names are used in population dynamics [21].

Populations can exhibit Allee dynamics due to a wide range of biological phenomenon, such as reduced antipredator vigilance, social thermoregulation, genetic drift, mating difficulty, reduced defense antipredator, deficient feeding to low densities, (see Table 2.1 in [7]).

A simplest form to the growth rate of a population affected by the Allee effect which will be used in this work is described by the cubic polynomial differential equation:

$$\frac{dx}{dt} = r\left(1 - \frac{x}{K}\right)(x - m)\,x$$

where $-K < m \ll K$ and $r, K > 0$. When $m > 0$, it has a *strong Allee effect* and the population growth rate decreases if the population size is below the threshold level m and the population goes to extinction. If $m \leq 0$, it is said that the population is affected by a *weak Allee effect* [13]. When $m < -K$, the above equation does not represent an Allee effect.

Many algebraic forms have been used to describe the Allee effect as are shown in [4], although most of them are topologically equivalent as is proved in [11]. Nonetheless, different forms may produce a change in the quantity of limit cycles surrounding a positive equilibrium point in predator–prey models as is shown (demonstrated) in [14].

The problem of determining conditions, which guarantees the uniqueness of a limit cycle or the global stability of the unique positive equilibrium in predator–prey systems, has been extensively studied over the past three decades [17], starting with the work by Cheng [5], who was the first to prove the uniqueness of a limit cycle for a specific predator–prey model with a Holling type II functional response, using the symmetry of the prey isocline.

The latter is related to the unsolved problem proposed by the mathematician David Hilbert in 1900 [10], and refers to finding the maximum number of limit cycles of a bidimensional polynomial differential equation system, whose degree must be equal to $p \in \mathbb{N}$. However, it is not an easy task to study the quantity of limit cycles that can be generated throughout the bifurcation of a center focus [6].

This chapter is organized as follows: The modified Leslie–Gower model is presented in the next section; in Sect. 2; properties of model are established in Sect. 3; Sect. 4 shows some simulations and the last section presents a discussion of results.

2 The Model

The predator–prey model to be analyzed is described by the autonomous differential equation system of Kolmogorov type given by:

$$X_\mu : \begin{cases} \frac{\mathrm{d}x}{\mathrm{d}t} = \left(r \left(1 - \frac{x}{K}\right)(x - m) - \frac{qx\ y}{x^2 + bx + a} \right) x \\ \frac{\mathrm{d}y}{\mathrm{d}t} = s \left(1 - \frac{y}{n\ x}\right) y \end{cases} \tag{1}$$

where $x(t)$ and $y(t)$ denote the prey and predator population size, respectively, for $t \geq 0$, measured as the number of individuals or biomass or density by area or volume unit. All parameters are positive, i.e., $\mu = (r, q, a, s, K, n, b, m) \in \mathbb{R}^7_+ \times]-K, K[$, having the following biological meanings: r represents the intrinsic growth rate of the prey, K is the prey environmental carrying capacity, $m > 0$ is the minimum viable population, that is, the threshold below which the population goes to extinction, q is the per capita attack rate of predators, or the maximal per capita consumption rate, i.e., the maximum number of prey that can be eaten by a predator in each time unit (when $b \geq 0$); when $b < 0$, q is the saturation predation, a and b are fitting parameters [19]. When $b = 0$, $\sqrt{a}$ is the half-saturation constant. s represents the intrinsic growth rate of predators, n is a measure of the food quality. In this work, only the case where $b > 0$ will be analyzed.

We note that system (1) is not defined at the y-axis, particularly at the point $(0, 0)$, which is a point of particular interest; the system (1) or vector field X_μ is defined on the set:

$$\Omega = \left\{(x, y) \in \mathbb{R}^2 / x > 0, y \geq 0\right\} = \mathbb{R}^+ \times \mathbb{R}^+_0.$$

The equilibrium point of system (1) or singularities of vector field X_μ are: $P_K = (K, 0)$, $P_\mathrm{e} = (x_\mathrm{e}, y_\mathrm{e})$, where P_e is the positive equilibrium point, satisfying the equations of the isoclines $y = nx$ and $y = \frac{r}{qx}\left(1 - \frac{x}{K}\right)(x - m)\left(x^2 + bx + a\right)$.

In order to reduce the number of parameters and make an adequate description of behavior of the system (1), we follow the methodology used in [1, 25, 27], making a change of variables and time rescaling given by the function:

$$\varphi : \bar{\Omega} \times \mathbb{R} \to \Omega \times \mathbb{R}$$

such as,

$$\varphi(u, v, \tau) = \left(Ku, Knv, \frac{u}{rK}\left(u^2 + \frac{b}{K}u + \frac{a}{K^2}\right)\tau \right) = (x, y, t)$$

with

$$\bar{\Omega} = \left\{(u, v) \in \mathbb{R}^2 / u \geq 0, v \geq 0\right\} = \mathbb{R}^+_0 \times \mathbb{R}^+_0.$$

We have that $\det D\varphi(u, v, \tau) > 0$, that is, φ is a diffeomorphism preserving the orientation of the time [6, 8], for which the vector field X_μ or system (1) in the new

system of coordinates, is topologically equivalent to the vector field $Y_\delta = \varphi \circ X_\mu$; it takes the form $Y_\eta = P(u,v)\frac{\partial}{\partial u} + Q(u,v)\frac{\partial}{\partial v}$ [8] and the associated differential equations is given by a sixth-order polynomial system:

$$Y_\delta : \begin{cases} \frac{\mathrm{d}u}{\mathrm{d}\tau} = \left((1-u)(u-M)\left(u^2+Bu+A\right) - Quv\right)u^2 \\ \frac{\mathrm{d}v}{\mathrm{d}\tau} = S(u-v)\left(u^2+Bu+A\right)v \end{cases} \tag{2}$$

with $A = \frac{a}{K^2}$, $Q = \frac{qn}{rK}$, $S = \frac{s}{rK}$, $B = \frac{b}{K}$; $M = \frac{m}{K}$ where, $\delta = (M, A, Q, B, S) \in]1,1[\times \mathbb{R}^4_+$. In the following, we only consider in the model the weak Allee effect, that is, when $M = 0$; the model with strong Allee effect for $M > 0$ will be analyzed in a future work. So, system (2) takes the form:

$$Y_\eta : \begin{cases} \frac{\mathrm{d}u}{\mathrm{d}\tau} = \left((1-u)\left(u^2+Bu+A\right) - Qv\right)u^3 \\ \frac{\mathrm{d}v}{\mathrm{d}\tau} = S(u-v)\left(u^2+Bu+A\right)v \end{cases} \tag{3}$$

with $\eta = (A, B, Q, S) \in \mathbb{R}^4_+$. System (3) is also defined at $\bar{\Omega}$. The equilibrium points of system (3) or singularities of vector field X_μ are $(0,0)$, $(1,0)$ and the positive equilibrium points satisfying the equations of the isoclines $v = u$ and $v = \frac{1}{Q}(1-u)\left(u^2+Bu+A\right)$.

The abscise of this point at $\bar{\Omega}$ satisfies the third-degree equation:

$$P(u) = u^3 - (1-B)u^2 + (A-B+Q)u - A = 0. \tag{4}$$

By Descartes' rule of sign, the polynomial $P(u)$ may have a real positive root or three different real positive roots or two different real positive roots, one of them with multiplicity two, depending on the sign of the coefficients $(1-B)$ and $(A-B+Q)$. We denote by H the real positive root that always exists.

If $1-B > 0$ and $A-B+Q > 0$, there exists at least one positive real root that we denote as $u = H$.

If $1-B > 0$ and $A-B+Q \leq 0$, there exists a unique positive real root.

If $1-B \leq 0$ and any be the sign of $A-B+Q$, there exists a unique positive real root.

To determine the local nature of equilibrium points we need the Jacobian matrix of system (3) given by

$$DY_\eta(u,v) = \begin{pmatrix} DY_\eta(u,v)_{11} & -Qu^3 \\ Sv\left(A+2Bu-Bv-2uv+3u^2\right) & (u-2v)\left(u^2+Bu+A\right) \end{pmatrix}$$

with $DY_\eta(u,v)_{11} = -u^2\left(6u^3+(5B-5)u^2+(4A-4B)u+(3Qv-3A)\right)$

3 Main Results

For system (3) or vector field Y_η, we have the following results:

Lemma 1 *The set* $\bar{\Gamma} = \{(u, v) \in \mathbb{R}^2 / u \geq 0, v \geq 0\}$ *is an invariant region.*

Proof As system (3) is of Kolmogorov type, then, the coordinates axis are invariant sets [6].

Let $u = 1$; we have that $\frac{\mathrm{d}u}{\mathrm{d}\tau} = -Qv < 0$, and for any sign of $\mathrm{d}v/\mathrm{d}\tau = S(1-v)(A+B+1)v$ the trajectories enter to the region $\bar{\Gamma}$. □

We note that in system (1) the set

$$\Gamma = \{(x, y) \in \mathbb{R}^2 / x > 0, y \geq 0\}$$

is an invariant region.

Lemma 2 *The solutions are bounded.*

Proof Using Poincaré compactification [6].

Let be $X = \frac{u}{v}$ and $Y = \frac{1}{v}$, then,

$$\frac{dX}{d\tau} = \frac{1}{v^2}\left(v\frac{du}{d\tau} - u\frac{dv}{d\tau}\right), \frac{dY}{d\tau} = -\frac{1}{v^2}\frac{dv}{d\tau};$$

then, the system takes the form:

$$\hat{Y}_\eta : \begin{cases} \frac{\mathrm{d}X}{\mathrm{d}\tau} = \frac{1}{Y^4}\begin{pmatrix} X^4Y - X^5 - BX^4Y - QX^3Y + SX^3Y - SX^4Y + AX^2Y^3 \\ -AX^3Y^2 + BX^3Y^2 + ASXY^3 - ASX^2Y^3 + BSX^2Y^2 - BSX^3Y^2 \end{pmatrix}. \\ \frac{\mathrm{d}Y}{\mathrm{d}\tau} = -\frac{S}{Y^2}(X-1)\left(AY^2 + X^2 + BXY\right) \end{cases}$$

To simplify the calculus, we make a time rescaling given by $T = \frac{1}{Y^4}\tau$ then,

$$\widetilde{Y}_\eta : \begin{cases} \frac{dX}{d\tau} = \begin{pmatrix} X^4Y - X^5 - BX^4Y - QX^3Y + SX^3Y - SX^4Y + AX^2Y^3 \\ -AX^3Y^2 + BX^3Y^2 + ASXY^3 - ASX^2Y^3 + BSX^2Y^2 - BSX^3Y^2 \end{pmatrix} \\ \frac{dY}{d\tau} = -SY^2(X-1)\left(AY^2 + X^2 + BXY\right) \end{cases}$$

then,

$$D\widetilde{Y}_\eta(0,0) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

For desingularizing the origin, we consider the blowing-up directional method [8], making $X = r$ and $Y = r^2 s$; then, we have:

$$V_\eta : \begin{cases} \frac{dr}{dT} = \frac{dr}{dT} \\ \frac{ds}{dT} = \frac{1}{r^2}\left(\frac{dY}{dT} - 2rs\frac{dr}{dT}\right). \end{cases}$$

So,

$$V_\eta : \begin{cases} \frac{dr}{dT} = -r^5 \begin{pmatrix} Qs - Ss - rs + Ar^2s^2 - Ar^3s^3 - Br^2s^2 + Brs \\ +Srs - BSrs^2 - ASr^2s^3 + ASr^3s^3 + BSr^2s^2 + 1 \end{pmatrix} \\ \frac{ds}{dT} = r^4 s \begin{pmatrix} 2Qs - Ss - 2rs + 2Ar^2s^2 - 2Ar^3s^3 - 2Br^2s^2 + 2Brs \\ +Srs - BSrs^2 - ASr^2s^3 + ASr^3s^3 + BSr^2s^2 + 2 \end{pmatrix} \end{cases}$$

Once again, making a time rescaling given by: $\lambda = r^4 T$, a new rescaled vector field is obtained:

$$\bar{V}_\eta : \begin{cases} \frac{dr}{d\lambda} = -r \begin{pmatrix} Qs - Ss - rs + Ar^2s^2 - Ar^3s^3 - Br^2s^2 + Brs \\ +Srs - BSrs^2 - ASr^2s^3 + ASr^3s^3 + BSr^2s^2 + 1 \end{pmatrix} \\ \frac{ds}{d\lambda} = s \begin{pmatrix} 2Qs - Ss - 2rs + 2Ar^2s^2 - 2Ar^3s^3 - 2Br^2s^2 + 2Brs \\ +Srs - BSrs^2 - ASr^2s^3 + ASr^3s^3 + BSr^2s^2 + 2 \end{pmatrix} \end{cases}.$$

Evaluating the Jacobian matrix of $\bar{V}_\eta$ in $(0,0)$, we obtain

$$D\bar{V}_\eta(0,0) = \begin{pmatrix} -1 & 0 \\ 0 & 2 \end{pmatrix}$$

Thus, $(0,0)$ is a hyperbolic saddle point of vector field $\bar{V}_\eta$ since $\det D\bar{V}_\eta(0,0)$; so, $(0,0)$ is a nonhyperbolic saddle point of vector field $\overline{Y}_\eta$ and $\widetilde{Y}_\eta$, which is repelling over the positive s-axis; hence, $(0,\infty)$ is a nonhyperbolic saddle point of vector field Y_η, repelling negatively over the $v-$ axis, Therefore, the solutions of the system (3) are bounded. □

For the following lemma, we define $\Delta = (1-B-H)^2 - 4\frac{A}{H}$.

Lemma 3

1. *For Eq.* (4)*, we have:*
   - 1.1. *There is one positive real root , if and only if,* $\Delta < 0$.
   - 1.2. *Three different real positive roots, if and only if,* $\Delta > 0$.
   - 1.3. *Two real positive roots, one of them having multiplicity two, if and only if,* $\Delta = 0$*; they are*

$$H \text{ and } E^* = \frac{1-H-B}{2}$$

2. *For system* (3) *or vector field* Y_η*, we have:*
   - 2.1. *If* $\Delta < 0$*, there is a unique equilibrium point* $P_e = (H,H)$ *at the interior of* $\bar{\Omega}$.

2.2. If $\Delta = 0$*, there exist two equilibrium points at the interior of* $\bar{\Omega}$*, which are*

$$P_{\mathrm{e}} = (H, H) \text{ and } P^* = \left(\frac{1 - H - B}{2}, \frac{1 - H - B}{2} \right)$$

2.3. If $\Delta > 0$*, there exist three equilibrium points at the interior of* $\bar{\Omega}$*, which are* $P_{\mathrm{e}} = (H, H)$, $P_2 = (E_2, E_2)$*, and* $P_3 = (E_3, E_3)$ *with*

$$E_2 = \frac{(1 - H - B) - \sqrt{\Delta}}{2} \text{ and } E_3 = \frac{(1 - H - B) + \sqrt{\Delta}}{2}$$

Proof 1. Let $u_{\mathrm{e}} = H$ be the positive real root that always exists for Eq. (4) and $P_{\mathrm{e}} = (H, H)$ the equilibrium point that always exists in $\bar{\Omega}$.

Dividing the polynomial $P(u)$ by $(u - H)$, the polynomial

$$P_1(u) = u^2 - (1 - H - B)\, u + A - B + Q + H\,(B + H - 1)$$

is obtained as factor of $P(u)$ and the rest is

$$R(H) = H^3 - (1 - B)\, H^2 + (A - B + Q)\, H - A = 0.$$

Then,

$$Q = \tfrac{1}{H}\,(1 - H)\left(H^2 + BH + A\right).$$

Replacing Q in $P_1(u)$, we have that:

$$P_1(u) = u^2 - (1 - H - B)\, u + \tfrac{A}{H}.$$

Considering the sign of Δ, for $P_1(u)$ we have $\Delta = (1 - H - B)^2 - 4\frac{A}{H}$:

(1.1) Has no real root, if and only if, $\Delta < 0$.
(1.2) Has two different real positive root, if and only if, $\Delta > 0$, which are:

$E_2 = \frac{(1-H-B)-\sqrt{\Delta}}{2}$ and $E_3 = \frac{(1-H-B)+\sqrt{\Delta}}{2}$.
Clearly, $E_2 < E_3$.

(1.3) Has one positive roots of multiplicity two, if and only if, $\Delta = 0$.
$E^* = \frac{1-H-B}{2}$
2. It is immediate. □

Lemma 4 *The singularity or equilibrium point* $(1, 0)$ *is a saddle point for all parameter values.*

Proof Evaluating the Jacobian matrix at equilibrium point $(1, 0)$

$$DY_\eta\,(1, 0) = \begin{pmatrix} -(A + B + 1) & -Q \\ 0 & S\,(A + B + 1) \end{pmatrix}$$

Clearly, $\det DY_\eta(1,0) = -S(A+B+1)^2 < 0$, thus, the point $(1,0)$ is a hyperbolic saddle point. □

Lemma 5 *The point* $(0,0)$ *is a nonhyperbolic singularity of the vector field* Y_η*, which has a hyperbolic and a parabolic sector [24]. Thus, there exists a separatrix curve* Σ *dividing the behavior of trajectories in the phase plane.*

Proof Evaluating the Jacobian matrix at the point $(0,0)$ we have that

$$DY_\eta(0,0) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Here, the origin is a nonhyperbolic singularity [8, 24]. To desingularize the origin, we consider the vertical blowing-up method [24], that is, we consider the function given by $\Psi(p,q) = (pq,q) = (u,v)$.
We have that $\frac{dp}{d\tau} = \frac{1}{q}\left(\frac{du}{d\tau} - p\frac{dq}{d\tau}\right)$ and $\frac{dq}{d\tau} = \frac{dv}{d\tau}$; rescaling the time by $T = q\tau$, it becomes,

$$\overline{Y}_\eta : \begin{cases} \frac{dp}{d\tau} = p\begin{pmatrix} p^4q^3 - p^5q^4 + AS + Ap^2q - Ap^3q^2 + Bp^3q^2 - Bp^4q^3 - Qp^2q^2 + \\ Sp^2q^2 - Sp^3q^2 - ASp - BSp^2q + BSpq \end{pmatrix} \\ \frac{dq}{d\tau} = Sq\left(p^3q^2 - p^2q^2 + Bp^2q - Bpq + Ap - A\right). \end{cases}$$

If $q = 0$, then $\frac{dq}{d\tau} = 0$. Moreover, $\frac{dp}{d\tau} = p(AS - ASp)$, thus the singularities are: $(0,0)$ and $(0,1)$, then

$$D\overline{Y}_\eta(0,0) = \begin{pmatrix} AS & 0 \\ 0 & -SA \end{pmatrix},$$

then, $\det D\overline{Y}_\eta(0,0) = -A^2S^2 < 0$; $(0,0)$ is saddle point. Repeller by the p-axis and attractor by the q-axis.

$$D\overline{Y}_\eta(0,1) = \begin{pmatrix} 0 & A \\ -BS + AS & -AS \end{pmatrix}$$

then, $\det D\overline{Y}_\eta(0,1) = -AS(A-B)$. Thus, the point $(0,1)$ of vector field $\overline{Y}$ is:

a. A nonhyperbolic saddle point , if and only if, $A > B$.
b. An attractor equilibrium point, if and only if, $A < B$.

Then, by the blowing down, the point $(0,0)$ is a nonhyperbolic saddle point in the system (3). □

We note the point $(0,0)$ of the vector field Y_η is a nonhyperbolic attractor point. The trajectories above the separatrix Σ have the point $(0,0)$ as their ω-limit. The trajectories below this separatrix Σ have different ω-limit as will be shown later.

Corollary 1 *The stable manifold* $W^s\,(0,0)$ *of the nonhyperbolic equilibrium point* $(0,0)$ *divides the behavior of trajectories; so, the point* $(0,0)$ *is an attractor and* ω*-limit of all solutions which initial conditions lie above* $W^s\,(0,0)$.

Proof By Lemma 5 above the point $(0,0)$ is a nonhyperbolic saddle point with a hyperbolic sector; the stable manifold $W^s\,(0,0)$ determined by the separatrix curve divides the behavior of trajectories in the phase plane; any solutions above the manifold $W^s\,(0,0)$ have $(0,0)$ as its ω-limit. Those trajectories with initial conditions below the separatrix curve can have different ω-limits. □

Theorem 1 *Let* $W^s\,(0,0)$ *and* $W^u\,(1,0)$ *be the stable and unstable manifolds of* $(0,0)$ *and* $(1,0)$*; then there exists a subset of parameters for which the intersection of* $W^s\,(0,0)$ *and* $W^u\,(1,0)$ *is not empty, giving rise to the heteroclinic curve* γ *joining the point* $(0,0)$ *and* $(1,0)$.

Proof By Lemma 5, the point $(0,0)$ has a separatrix and by Lemma 4 the point $(1,0)$ is saddle.

Let $W^s\,(0,0)$ and $W^u\,(1,0)$ be the stable and unstable manifolds of $(0,0)$ and $(1,0)$; it is clear that the α-limit of $W^s\,(0,0)$ and the ω-limit of $W^u\,(1,0)$ are not at infinity on the direction of v-axis; then, there are points $(u^\star, v^s) \in W^s\,(0,0)$ and $(u^\star, v^u) \in W^u\,(1,0)$ where v^s and v^u are functions of the parameters A, B, Q, and S, i.e., $v^s = f_1\,(A, B, Q, S)$ and $v^u = f_2\,(A, B, Q, S)$.

It is clear that if $0 < u \ll 1$, then $v^s < v^u$ and if $0 \ll u < 1$, then $v^s > v^u$. Since the vector field Y_η is continuous with respect to the parameter values, then the unstable manifold $W^s\,(0,0)$ intersects the unstable manifold $W^u\,(1,0)$; therefore, there exist $\left(u_s^\star, v_s^\star\right) \in \Gamma$ (invariant region), such as $v^\star = v^\star$. This equation defines a surface in the parameter space for which the heteroclinic curve γ exists. □

The separatrix curve Σ, the straight line $u = 1$ and the u-axis determine a subregion $\bar{\Gamma}$, which is closed and bounded, that is,

$$\bar{\Gamma} = \left\{(u,v) \in (\mathbb{R}_+)^2 / 0 \le u \le 1, 0 \le v \le v^s \text{ and } v^s \in \Sigma\right\}$$

it is a compact region, where it is possible to apply the Poincaré–Bendixon theorem. To study the nature of the equilibrium point (H, H) with $H < 1$, by Lemma 3, we have that $Q = \frac{1}{H}\,(1-H)\left(H^2 + BH + A\right)$; then, the vector field X_μ or system (3) takes the form:

$$X_\theta : \begin{cases} \frac{du}{d\tau} = \left((1-u)\left(u^2 + Bu + A\right) - \frac{(1-H)\left(H^2+BH+A\right)}{H} v\right) u^3 \\ \frac{dv}{d\tau} = S\,(u - v)\left(u^2 + Bu + A\right) v \end{cases} \tag{5}$$

with $\theta = (A, B, Q, H) \in (]0,1[)^2 \times \mathbb{R}^2$. The Jacobian matrix is:

$$DX_\theta\,(u,v) = \begin{pmatrix} -H^3\left(A - B - 2H + 2BH + 3H^2\right) & -H^2\,(1-H)\left(A + BH + H^2\right) \\ SH\left(A + BH + H^2\right) & -SH\left(A + BH + H^2\right) \end{pmatrix}$$

Then,

$$\det DX_\theta\,(H,H) = SH^3\left(H^2 + BH + A\right)\left(A + BH^2 - H^2 + 2H^3\right) > 0$$

and the trace is given by:

$$\mathrm{tr}DX_\theta(H,H) = -H^3\left(A - B - 2H + 2BH + 3H^2\right) - HS\left(H^2 + BH + A\right).$$

If $\mathrm{tr}DX_\theta(H,H) = 0$, then $S = \frac{H^2\left(-3H^2+2(1-B)H+(B-A)\right)}{H^2+BH+A}$ Let

$$\begin{aligned} P &= \left(\mathrm{tr}DY_\theta(H,H)\right)^2 - 4\det DY_\theta(H,H) \\ &= H^2\left(A + BH + H^2\right)^2 S^2 + 2H^3\left(A + BH + H^2\right)\left(-2A + AH - BH - H^3\right) S \\ &\quad + H^6\left(A - B - 2H + 2BH + 3H^2\right)^2 \end{aligned}$$

System (5) has the following properties:

Theorem 2 *Let (H,H) be the unique positive equilibrium point at the first quadrant; then,*

1. *It is an attractor, if and only if, $trDX_\theta(H,H) < 0$; then, $S > \frac{H^2\left(-3H^2+2(1-B)H+(B-A)\right)}{H^2+BH+A}$. Moreover,*
   a) It is an attractor node, if and only if, $P > 0$.
   b) It is an attractor focus, if and only if, $P < 0$.
2. *It is a repeller, if and only if, $trDX_\theta(H,H) > 0$; thus, $S < \frac{H^2\left(-3H^2+2(1-B)H+(B-A)\right)}{H^2+BH+A}$. Moreover,*
   a) It is a repeller focus, surrounded by a limit cycle, if and only if, $P < 0$.
   b) It is a repeller node, if and only if, $P > 0$.
3. *It is a weak focus, if and only if, $S = \frac{H^2\left(-3H^2+2(1-B)H+(B-A)\right)}{H^2+BH+A}$.*

Proof

1. (H,H) is an attractor, if and only if, $S > \frac{H^2\left(-3H^2+2(1-B)H+(B-A)\right)}{H^2+BH+A}$; moreover,
   a) If $P > 0$ then, the point (H,H) is an attractor node.
   b) If $P < 0$ then, the point (H,H) is an attractor focus.
2. If $S < \frac{H^2\left(-3H^2+2(1-B)H+(B-A)\right)}{H^2+BH+A}$ if and only if $trDX_\theta(H,H) > 0$ and (H,H) is an repeller; moreover,
   a) If $P < 0$, then, the point is a repeller focus. By the Poincaré–Bendixon theorem [6, 15, 24] in the subregion $\bar{\Gamma}$ determined by the line $u = 1$, the u-axis and the stable manifold $W^s(0,0)$, the point (H,H) is surrounded by at least one limit cycle.
   b) When the parameters change, the limit cycle increases until it coincides with the heteroclinic γ; when the heteroclinic is broken this limit cycle disappears.

Then, $P > 0$ and $S < \frac{H^2\left(-3H^2+2(1-B)H+(B-A)\right)}{H^2+BH+A}$;
therefore, (H,H) becomes a repeller node. □

Lemma 6 *A Hopf bifurcation at equilibrium point (H,H) occurs in the system (3) for the bifurcation value $S = \frac{H^2\left(-3H^2+2(1-B)H+(B-A)\right)}{H^2+BH+A}$.*

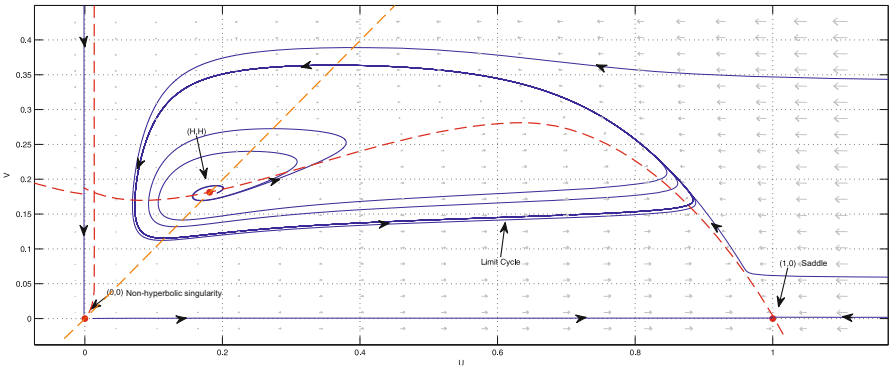


Fig. 2 For $A = 0.1$, $B = 0.0002$, $S = 0.024$, and $Q = 0.6$, the point (H, H) is a repeller focus surrounded by a limit cycle and a wide set of trajectories going to this ω-limit. The point $(0, 0)$ is a nonhyperbolic local attractor, the point $(1, 0)$ is a hyperbolic saddle; there exists a separatrix curve Σ dividing the behavior of trajectories

Proof The proof follows from the above theorem since the determinant is always positive and the trace changes sign. In addition, the transversality condition [15] is verified, since we have that

$$\frac{\mathrm{d}\left(\mathrm{tr}DY_{\eta}(H,H)\right)}{\mathrm{d}S} = -H\left(A + BH + H^2\right) < 0.$$

□

4 Simulations

In order to reinforce the obtained results, some computer simulations are also given (Figs. 2, 3 and 4), presenting different behaviors of system (3). The diverse natures of the positive equilibrium point (H, H) are shown for different parameters values.

5 Conclusions

In this work, we analyze a Leslie–Gower predator–prey model where a generalized Holling type III functional response and a weak Allee effect on prey are assumed. In order to facilitate the calculations we make a reparameterization and a time rescaling to obtain a topologically equivalent polynomial. We observe that the incorporation of this ecological phenomenon into a Leslie–Gower predator–prey model can increase the number of equilibrium points at interior of the first quadrant [19].

However, we analyzed only one case assuming the existence of a unique positive equilibrium point; the other cases must be studied in future works to complete the description of the properties of the model described by system (1).

Using the method of blowing up, we demonstrate the existence of a separatrix determined by the stable manifold of nonhyperbolic singularity $(0, 0)$ dividing the

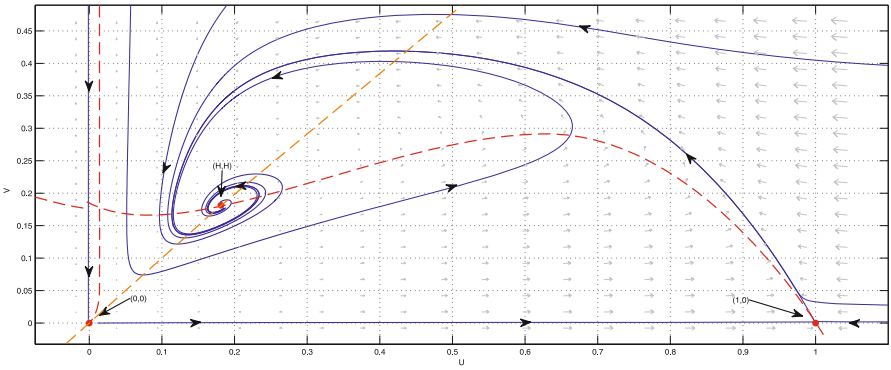


Fig. 3 For $A = 0.1$, $B = 0.0002$, $Q = 0.6$, and $S = 0.089$ the point (H, H) is an attractor focus. The point $(0, 0)$ is a nonhyperbolic saddle-node and $(1, 0)$ is a saddle point

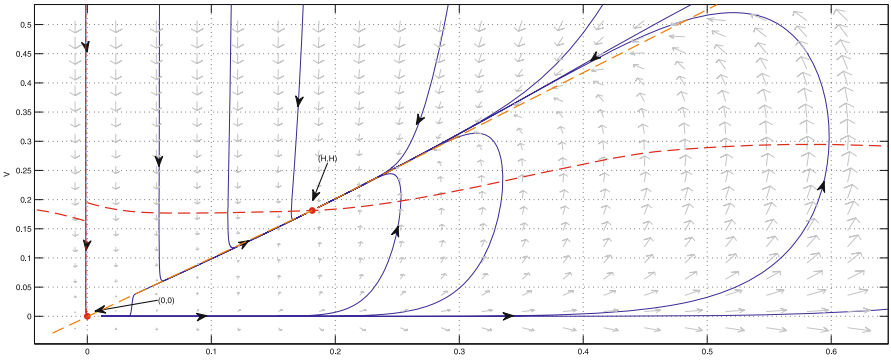


Fig. 4 For $A = 0.1$, $B = 0.0002$, $Q = 0.6$, and $S = 2.23$, the point (H, H) is an attractor node. The point $(1, 0)$ is a hyperbolic saddle and the point $(0, 0)$ is a nonhyperbolic attractor

behavior of trajectories in the phase plane, which can have different ω-limit. Then, some solutions are highly sensitive to initial conditions, which means that those trajectories with initial condition below this curve, i.e., in region $\bar{\Gamma}$, have as ω-limit either a limit cycle or a stable positive equilibrium point. Those with initial conditions above this curve have the origin $(0, 0)$ as their ω-limit.

Ecologically this implies that small perturbations due to environmental changes caused by pollution or other causes, could provoke the extinction of both populations. Then, populations can coexist oscillating around the positive equilibrium either tending to this equilibrium or the extinction of both populations can occur, yet, if the ratio prey–predator is high.

The existence of parameter constraints for which the positive equilibrium point is an attractor or is a repeller surrounded by at least one limit cycle is proved. Moreover, it is proved that there exists a heteroclinic curve joining the equilibrium $(1, 0)$ and the singularity $(0, 0)$.

We conclude that the weak Allee effect causes significant changes with respect to the system where this phenomenon is absent [18], since it can change the quantity of equilibrium points and limit cycles surrounding a positive equilibrium point.

Acknowledgement The authors would like to thank the members of the Mathematics Ecology Group from the Mathematics Institute from the Pontificia Universidad Católica de Valparaíso for their valuable comments and suggestions. This work was partially financed by Project Fondecyt N^o 1120218 and DIEA-PUCV 124.730/2012.

References

1. Aguilera-Moya, A., González-Yañez, B., González-Olivares, E.: Existence of multiple limit cycles on a predator-prey with generalized nonmonotonic functional response, In: Mondaini, R. (ed.) Proceedings of the Fourth Brazilian Symposium on Mathematical and Computational Biology, E-Papers Serviços Editoriais, Ltda., vol. 2, 196–210. Rio de Janeiro (2005)
2. Arrowsmith, D.K., Place, C.M.: Dynamical system. Differential equations, maps and chaotic behaviour. Chapman and Hall. London (1992)
3. Aziz-Alaoui, M.A., Daher-Okiye, M.: Boundedness and global stability for a predator-prey model with modied Leslie-Gower and Holling-type II schemes. Appl. Math. Lett. **16**, 1069–1075 (2003)
4. Boukal, D.S., Berec, L.: Single-species models and the Allee effect: Extinction boundaries, sex ratios and mate encounters. J. Theoret. Biol. **218**, 375–394 (2002)
5. Cheng, K.S.: Uniqueness of a limit cycle for a predator-prey system. SIAM J. Math. Anal. **12**, 541–548 (1981)
6. Chicone, C.: Ordinary differential equations with applications. Texts in Applied Mathematics, vol. 34, 2nd edn. Springer. New York (2006)
7. Courchamp, F., Berec, L., Gascoigne, J.: Allee effects in ecology and conservation. Oxford University Press. Oxford 2008
8. Dumortier, F., Llibre, J., Artés, J.C.: Qualitative theory of planar differential systems. Springer. Berlin Heidelberg (2006)
9. Freedman, H.I.: Deterministic mathematical model in population ecology. Marcel Dekker. New York (1980)
10. Gaiko, V.A.: Global bifurcation theory and Hilbert sixteenth problem. Mathematics and its applications, vol. 559. Kluwer Academic Publishers (2003)
11. González-Olivares, E., González-Yañez, B., Mena-Lorca, J., Ramos-Jiliberto, R.: Modelling the Allee effect: Are the different mathematical forms proposed equivalents?. In: Mondaini, R. (ed.) Proceedings of the International Symposium on Mathematical and Computational Biology BIOMAT 2006, E-papers Serviços Editoriais Ltda, pp. 53–71. Rio de Janeiro (2007)
12. González-Olivares, E., Rojas-Palma, A.: Multiple limit cycles in a Gause type predator prey model with Holling type III functional response and Allee effect on prey. Bull. Math. Biol. **35**, 366–381 (2011)
13. González-Olivares, E., Rojas-Palma, A: Limit cycles in a Gause-type predator-prey model with sigmoid functional response and weak Allee effect on prey. Math. Methods Appl. Sci. **35**, 963–975 (2012)
14. González-Olivares, E., Rojas-Palma, A.: Allee effect in Gause type predator-prey models: Existence of multiple attractors, limit cycles and separatrix curves. A brief review. Math. Model. Nat. Phenom. **8**(6), 143–164 (2013)
15. Guckenheimer, F., Holmes, P.: Nonlinear oscillations, dynamical systems, and bifurcations of vector fields. Springer. New York (1983)

16. Hanski, I., Hentonnen, H., Korpimaki, E., Oksanen, L., Turchin, P.: Small-rodent dynamics and predation. Ecology **82**, 1505–1520 (2001)
17. Hasik, K.: On a predator prey system of Gause type. J. Math. Biol. **60**, 59–74 (2010)
18. Huang J., Ruan S., Song, J.: Bifurcations in a predator-prey system of Leslie type with generalized Holling type III functional response. J. Diff. Equat. (2014, in press)
19. Lamontagne, Y., Coutu, C., Rousseau, C.: Bifurcation analysis of a predator-prey system with generalised Holling type III functional response. J. Dyn. Difference Equat. **20**, 535–571 (2008)
20. Li, C., Zhu, H.: Canard cycles for predator–prey systems with Holling types of functional response. J. Diff. Equat. **254**, 879–910 (2013)
21. Liermann, M., Hilborn, R.: Depensation: Evidence, models and implications. Fish Fisheries **2**, 33–58 (2001)
22. Ludwig, D., Jones, D.D., Holling, C.S.: Qualitative analysis of insect outbreak systems: The spruce budworm and forest. J. Anim. Ecol. **36**, 204–221 (1978)
23. May, R.M.: Stability and complexity in model ecosystems. Princeton University Press. Princeton (1974)
24. Perko, L.: Differential equations and dynamical systems, 3rd edn. Springer. New York (2001)
25. Rojas-Palma, A., González-Olivares, E., González-Yañez, B.: Metastability in a Gause type predator-prey models with sigmoid functional response and multiplicative Allee efect on prey. In: Mondaini, R. (ed.) Proceedings of International Symposium on Mathematical and Computational Biology, E-papers Serviços Editoriais Ltda., pp. 295–321. Rio de Janeiro (2007)
26. Ruan, S., Xiao, D.: Global analysis in a predator-prey system with nonmonotonic functional response. SIAM J. Appl. Math. **61**, 1445–1472 (2001)
27. Sáez, E., González-Olivares, E.: Dynamics on a predator-prey model. SIAM J. Appl. Math. **59**(5), 1867–1878 (1999)
28. Sugie, J., Miyamoto, K., Morino, K.: Absence of limits cycle of a predator-prey system with a sigmoid functional response. Appl. Math. Lett. **9**, 85–90 (1996)
29. Taylor, R.J.: Predation. Chapman and Hall. Springer, Dobredecht (1984)
30. Turchin, P.: Complex population dynamics: A theoretical/empirical synthesis. Princeton University Press, Princeton (2003)

Critical Points of Solutions to Elliptic Equations in Planar Domains with Corners

Jaime Arango and Jairo Delgado

Abstract We consider the solution u to a semilinear elliptic boundary value problem with Dirichlet boundary condition on an annular planar domain with corners. We prove that u possesses a finite number of critical points and at most one critical curve. For certain annular domains having a regular n–gon as an outer boundary, we rule out the existence of critical curves.

Keywords Critical points · Morse theory · Nonlinear elliptic equations · Moving planes

1 Introduction

Several applications in continuous mechanics are modeled by boundary value problems of the type:

$$\Delta u = f(u) \quad \text{in} \quad \Omega, \tag{1}$$

$$u = 0 \quad \text{on} \quad \partial\Omega, \tag{2}$$

where Δ is the Laplace operator, Ω is a planar domain, and f is a real-value function. For instance, (1) is the prototypical model for the deflection of a nonlinear membrane, fixed at the boundary, upon an external force f. When $f \equiv 1$, problem (1) reduces to the famous torsion problem.

The existence, uniqueness, and regularity questions of solution to (1) has been thoroughly investigated. However, the qualitative properties of solution to (1) seems to be less documented. In this investigation, we study the critical set of u, i.e., the points of vanishing gradient:

J. Arango (✉)
Departamento de Matemáticas, Universidad del Valle, Calle 13, 100-00, Cali, Colombia
e-mail: jaime.arango@correounivalle.edu.co

J. Delgado
Posgrado en Ciencias Matemáticas, Universidad del Valle, Calle 13, 100-00, Cali, Colombia
e-mail: jairo.delgado@correounivalle.edu.co

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_7

$$IK = \{x \in \overline{\Omega} : \nabla u(x) = 0\}.$$

Let us summarize some previous results on the description of the critical set IK. Makar–Limanov [10] considered the linear case $f \equiv 1$ of problem (1) in 1971 and proved that $z = \sqrt{-u}$ is concave provided the domain Ω is convex, and as a consequence u has exactly one critical point. Later on, Cabré and Chanillo [3] extended Makar–Limanov result; they proved that semistable solutions to problem (1) in convex domains with positive curvature have a unique nondegenerated critical point. For more recent results, we refer the reader to Finn (2008) [4], Arango and Gómez (2012) [1, 2]. We also mention Grossi and Molle [8], Gladiali and Grossi [9], and Grecco [6] who have tackled this and some other related problems. Although there are plenty results concerning the regularity of solutions to (1) in nonsmooth domains (see for instance [7]), very few is known about the critical set IK in such domains, not even when the boundary $\partial\Omega$ is made up of polygons.

The goal of this investigation is to extend to domains with corners the following result due to Arango and Gómez [1] :

Theorem 1 *(cf. [1, Theorem 3.1]) Let f be real analytic with $f(0) > 0$ and Ω be a smooth planar domain. If u is a solution to 1, then its critical set is made up of finitely many isolated points and Jordan curves. Moreover, if there is any Jordan curve contained in the critical set, this curve must be analytic and Ω cannot be simply connected.*

Assumption 1 Ω is an annular open region whose boundary $\partial\Omega$ is made up of two Jordan curves with convex interior. We assume that the outer boundary is a Lipschitz continuous piecewise analytic curve with a finite number of corners, while the inner is a smooth curve.

Let us precise the meaning of corners:

Definition 1 Assume Ω fulfills Assumption 1. An outer boundary point p is called a corner, if there exist $\alpha_j \in C^1([0,1], \partial\Omega)$, $j = 1, 2$, such that $\alpha_1([0,1]) \cap \alpha_2([0,1]) = \{p\}$ and

$$\alpha_1(1) = \alpha_2(0) = p \quad \text{and} \quad \lim_{t\to 1^-} \alpha_1'(0) \neq \lim_{t\to 0^+} \alpha_2'(0).$$

If f is real analytic with $f(0) > 0$ and Ω fulfills Assumption 1, we will extend Theorem 1 by proving that IK is made up of finitely many isolated points and at most one Jordan curve. We further prove that nut-like domains (see Definition 2) possesses no critical curves, and as a consequence, in nut-like domains IK is finite. We finish the chapter stating the following conjecture: in nut-like domain with a n-gon outer boundary, the set IK is made up of exactly $2n$ (critical) points. We also conjecture that only concentric annulus possesses a critical curve.

2 Critical Points and Moving Planes

In order to guarantee existence, uniqueness, and regularity of the solution u to (1), the following assumption on f will suffice:

Assumption 2 $f : IR \to IR$ is real analytic and nondecreasing, and $f(0) > 0$.

Proposition 1 *Let Ω be a planar domain having a Lipschitz continuous boundary $\partial\Omega$. If f fulfills Assumption 2, then there is a unique solution u to problem 1. Moreover, u is negative, real analytic in Ω, and continuous in $\overline{\Omega}$.*

The proof of the above proposition is spread through the technical literature of elliptic equations. See for instance [5, Theorems 8.15 and 12.5] regarding the existence and uniqueness results; the analyticity of the solution is proved in [11]. We also refer the reader to the existence and uniqueness results due to Krasnoselki (see [12, Theorem 1.16]).

Next, we establish some general results regarding the composition of the critical set IK.

Lemma 1 *Let $\Omega \subset IR^2$ is a bounded planar domain and suppose $f \in C^1(IR)$ is such that f is nondecreasing and $f(0) > 0$. If u is a solution to 1, then $\Delta u > 0$ on Ω. In particular, the Hessian matrix H_u vanishes nowhere in Ω*

The proof follows closely the one presented in Lemma 2 and Corollary 1 of [1]. We also refer to [2] for a slightly more general version of Lemma 1.

An immediate consequence of Lemma 1 is that all critical points of the solution u are semi-Morse. We use the term semi-Morse to refer to critical points p of a smooth function v, such that the Hessian matrix $H_v(p)$ does not vanish (see [1, 2] for more details).

Proposition 2 *Let f fulfill Assumption 2. If a Jordan critical curve exists inside the domain Ω, then Ω cannot be simply connected. Moreover, if Ω fulfills Assumption 1, then there is at most one Jordan critical curve, and in such a case, the curve circles the inner boundary of Ω.*

Proof Let Ω_γ be the subregion of Ω inside a critical curve γ. Notice that for all $\theta \in S^1$, the directional derivatives $u_\theta(x) = \nabla u(x) \cdot \theta$ satisfy

$$\begin{aligned} (\Delta - f'(u))u_\theta &= 0 \text{ in } \quad \Omega_\gamma, \\ u_\theta &= 0 \text{ on } \quad \partial\Omega_\gamma. \end{aligned}$$

Suppose by contradiction that Ω is simply connected. Since f' is nonnegative by Assumption 2, a classical maximum principle argument yields $u_\theta \equiv 0$ on Ω_γ, and this holds for all θ. Now by Proposition 1, u_θ is real analytic on Ω, therefore $u_\theta \equiv 0$ for all θ in Ω so that $u \equiv 0$. In view of Assumption 2, $u \equiv 0$ cannot be a solution to problem 1 and a contradiction is reached.

If Ω is smooth, then a straightforward application of Lemma 1 and the Hopf's boundary point lemma (see [14, Theorem 2.8.3]) yields that the solution u to problem (1) possesses no critical point at the boundary. However, if we have corners at the boundary, the matter is very different.

Example 1 The interior of an equilateral triangle of side $2\sqrt{3}$ centered at the origin can be described by

$$\Omega = \{(x_1, x_2) \in IR^2 : x_2 + 1 > 0,\ \sqrt{3}|x_1| < 2 - x_2\}.$$

Now, consider problem (1) with $f(s) \equiv 1$. The solution u is given by

$$u(x_1, x_2) = \frac{x_2 + 1}{12}\Big(3x_1^2 - (x_2 - 2)^2\Big).$$

A direct calculation shows that ∇u vanishes at $(0, 2)$, $(\sqrt{3}, -1)$, $(-\sqrt{3}, -1)$. These points are precisely the vertices of the triangle.

In what remains of this section, we will prove that critical points cannot accumulate at the boundary. This is a direct consequence of the Hopf's boundary point lemma when the boundary is smooth. However, in corner point we have to resource to the moving plane method. We refer the very influential paper of Serrin [13] to the reader for more details about the moving planes.

Proposition 3 *Suppose that f and Ω fulfill Assumptions 2 and 1, respectively. Then any critical boundary point of the solution u to (1) must be a corner. Moreover, boundary points are not limit points of IK.*

Proof Let $p \in \partial\Omega$. If p is not a corner, by Lemma 1, we can readily apply Hopf's boundary point lemma to conclude that

$$\frac{\partial u}{\partial \eta}(p) \neq 0,$$

where η stands for the unitary normal outer direction to $\partial\Omega$ at p. And as consequence $\nabla u(p) \neq 0$, therefore critical points cannot accumulate at p.

Let $p \in \partial\Omega$ be a corner and suppose by contradiction that there exists a sequence $(q_n)_n$ in IK such that $q_n \to p$. Since the outer boundary has a convex interior, then for every $q \in \Omega$, close enough to p, there exists a straight segment T such that $\Omega(T) \subset \Omega$, $\Omega'(T) \subset \Omega$, where $\Omega(T)$ is closure of the subregion of Ω, bounded by T and containing p and $\Omega'(T)$ is the reflection of $\Omega(T)$ with respect to T, see Fig. 1. Therefore, if n is big enough, we may choose T such that $q_n \in T$. In $\Omega'(T)$, we define

$$w(x) = u(x) - u(x'),$$

where x' denotes the reflection with respect to the line T of the point x. A direct calculation yields $\Delta w(x) = f'(c(x))w(x)$, where c is a smooth function, depends on $u(x)$ and $u^*(x)$. A little tough shows that w satisfies

$$\begin{aligned} \Delta w - f'(c(x))w = 0 \quad &\text{in} \quad \Omega'(T), \\ w < 0 \quad &\text{on} \quad \partial\Omega'(T) \setminus T, \\ w = 0 \quad &\text{on} \quad T \cap \partial\Omega'(T). \end{aligned}$$

Then by the maximum principle (recall that f' is nonnegative), w reaches the maximum at the boundary $\partial\Omega'(T)$, i.e., w reaches its maximum on T. By Hopf's boundary point lemma, we obtain:

$$\frac{\partial w}{\partial \eta} > 0 \quad \textit{on} \quad T \cap \partial\Omega'(T).$$

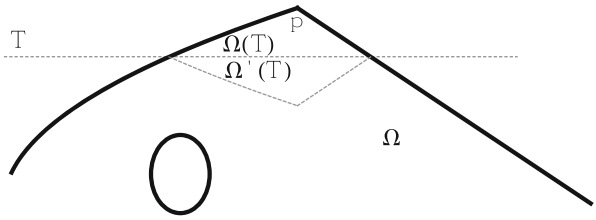


Fig. 1 The corner point p, the line T, the set $\Omega(T)$ and its reflection $\Omega^{'}(T) \subset \Omega$ with respect to T

Since $q_n \in T$, we have reached a contradiction.

Corollary 1 *Under the hypothesis of Proposition 3, there exist a compact set $Q \subset \Omega$ containing all interior critical points of the solution u to (1).*

3 Main Results

Now, we are in a position to estate the main results of our investigation.

Theorem 2 *If f and Ω fulfill Assumptions 2 and 1 respectively, then the critical set IK associated to the solution u to 1 is made up of finitely many isolated points and at most one Jordan curve.*

Proof By Corollary 1, there exists a compact set $Q \subset \Omega$ containing all the interior critical points of u.

Let $\epsilon_0 > 0$ such that $0 < f(-\epsilon_0)$ and $\sup_Q u < -\epsilon_0$. For $0 < \epsilon < \epsilon_0$, define

$$\ell = \{x \in \Omega : u(x) = -\epsilon\}.$$

Notice that ℓ is made up of exactly two smooth curves defining the boundary of an open annular region Ω_ε. Now define

$$v(x) = u(x) + \epsilon, \quad x \in \Omega_\epsilon.$$

By Proposition 2, there is at most one critical curve.

If $g(z) = f(z - \epsilon)$, then $g(0) = f(-\epsilon) > 0$ and so, g satisfies Assumption 2. Moreover, since $H_u = H_v$ in Ω_ϵ, then according to Lemma 1, H_v does not vanish by and v satisfies:

$$\begin{aligned} \Delta v &= g(v) \quad \text{in} \quad \Omega_\epsilon, \\ v &= 0 \qquad \text{on} \quad \partial\Omega_\epsilon. \end{aligned}$$

It follows from Theorem 1 that v has a finite number of isolated critical points and at most one Jordan critical curve. Finally, observe that u and v have the same critical points in Ω_ϵ. Outside Ω_ϵ, the only critical points of u, if at all exist, are corners.

Next, we rule out the existence of critical curves in certain planar domains.

Definition 2 A domain satisfying Assumption 1 is called a *nut-like domain* if the outer boundary is a convex regular polygon and its inner boundary is circle-centered at the polygon's center.

If the outer boundary of a nut-like domain Ω is a convex regular n–gon, then the rays, $e^{i\,j\pi/n}$, determine respectively axis of symmetry T_j of the domain Ω for $j = 1, \ldots, n$. Further, the symmetries of the domain are inherited by the solution u to Problem 1, in the sense, that for all $x \in \Omega$, and for all $j = 1, \ldots, n$, $u(x)$ equals $u(x^{(j)})$, where $x^{(j)}$ is the reflection of x about the axis T_j.

Theorem 3 *If f satisfies Assumption 2 and Ω is a nut-like domain, then the critical set IK corresponding to the solution u to 1 has no Jordan curves.*

Proof We carry out the proof assuming the outer boundary is a square, so that Ω can be described by:

$$\Omega = \left\{(x_1, x_2) \in IR^2 : r^2 < x_1^2 + x_2^2, \quad |x_1| < a, \quad |x_2| < a\right\},$$

r and a fixed parameters with $0 < r < a$.

Let q the $\pi/4$ rotation around the origin, $\Omega^* = q(\Omega)$ and $S = \Omega \cap \Omega^*$ see Fig. 2. We write $u^* = u \circ q^{-1}$ and notice that u^* satisfies (1) in Ω^*. Now, observe that S can be split up in eight congruent regions as it is shown in Fig. 2; the one containing the x_2 axis, lying above the x_1 axis, is called D. Notice that $\partial D = L_1 \cup L_2 \cup \alpha \cup \beta$, where L_1 and L_2 are segments of the rays $\mathrm{e}^{\mathrm{i}\,3\pi/8}$ and $\mathrm{e}^{\mathrm{i}\,5\pi/8}$, respectively, α is the arc $r\,\mathrm{e}^{\mathrm{i}\,\theta}$, $3\pi/8 < \theta < 5\pi/8$, and β is the segment of the horizontal line $x_2 = a$, bounded by the rays $\mathrm{e}^{\mathrm{i}\,3\pi/8}$ and $\mathrm{e}^{\mathrm{i}\,5\pi/8}$.

We remove the two extreme points of the segment β so that $q^{-1}(\beta) \subset \Omega$. Since $u^*(\beta) = u \circ q^{-1}(\beta)$, we obtain by Proposition 1 that $u^*(x) < 0$ for $x \in \beta$. Further, any axis of symmetry of the domain Ω is also an axis of symmetry of the solution u. Moreover, u and u^* coincide on $L_1 \cup L_2$, and due to the boundary conditions satisfying u and u^*, we have $u = u^*$ on α. As a result, it follows that

$$u - u^* \geq 0 \quad \text{on } \partial\Omega.$$

Next, a straightforward calculation shows that $u - u^*$ satisfies

$$\Delta(u - u^*) = f'(\xi(x))\left(u - u^*\right) \quad \text{in } D,$$

where $\xi(x)$ depends on $u(x)$ and $u^*(x)$. We recall that Assumption 2 guarantees that $f'(\xi(x)) \geq 0$ for all $x \in D$. Now by a standard maximum principle argument (see, for example, Theorem 2.1.1 in [14]), it follows that

$$u - u^* > 0 \quad \text{in } D.$$

To finish the proof, suppose by contradiction that there exists a critical curve γ. By Proposition 2, γ circles the inner boundary of Ω so that there exists a point

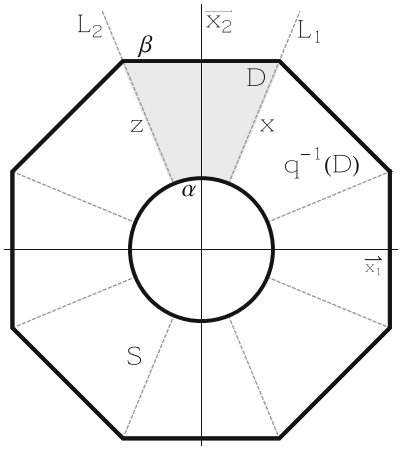


Fig. 2 The set $S = \Omega \cap \Omega^*$, the region D and its boundary $\partial D = L_1 \cup L_2 \cup \alpha \cup \beta$

$p \in \Omega \cap \gamma \cap L_1$. We observe that $\nabla(u - u^*)(p) = 0$. But the Hopf's boundary point Lemma (see [14, Theorem 2.8.4]) leads to a contradiction.

As we have shown in Theorem 3, the set IK is finite provided Ω is a nut-like domain. It is not difficult to prove that any axis of symmetries of an annular domain has at least two critical points, so we expect a nut-like domain, whose outer boundary is an n-gon, to have exactly $2n$ critical points, but a proof of this remains elusive to the authors.

Finally, it is worth noticing that the only known examples of problem (1) having a critical curve occur when Ω is a concentric annulus. Based on numerical evidence and several particular cases (see [1]), we conjecture that under Assumptions 2 and 1, there are no other examples exhibiting critical curves.

Acknowledgement The authors thank El Posgrado en Ciencias Matemáticas de la Universidad del Valle for providing the financial support and the academic environment to carry out this research. J. Delgado thanks the organizing committee of ICAMI 2013 for allowing him to present these results in the ICAMI meeting. J. Arango is greatly indebted to D. Cabarcas for her collaborations in proving Theorem 3 and for many stimulating conversations.

References

1. Arango, J., Gómez, A.: Critical points of solutions to elliptic problems in planar domains. Comm. Pure Appl. Anal. **10**, 327–338 (2011)
2. Arango, J., Gómez, A.: Critical points of solutions to quasilinear elliptic problems. Nonlinear Anal., Theory Methods Appl., Ser. A, Theory Methods **75**, 4375–4381 (2012)
3. Cabré, X., Chanillo, S.: Stable solutions of semilinear elliptic problems in convex domains. Sel. Math., New Ser. **4**, 1–10 (1998)
4. Finn, D.: Convexity of level curves for solutions to semilinear elliptic equations. Commun. Pure Appl. Anal. **7**, 1335–1343 (2008)

5. Gilbarg, G., Trudinger, N.: Elliptic Partial Differential Equations of Second Order. Springer. Berlin (1983)
6. Greco, A.: Quasi-concavity for semilinear elliptic equations with non-monotone and anisotropic nonlinearities. Bound. Value Probl. (2006). doi:10.1155/BVP/2006/80347
7. Grisvard, P.: Elliptic Problems in Nonsmooth Domains. Monographs and Studies in Mathematics. Pitman, Boston (1985)
8. Grossi, M., Molle, R.: On the shape of the solutions of some semilinear elliptic problems. Contemp. Math. **5**, 85–99 (2003)
9. Grossi, M., Gladiali, F.: Strict convexity of level sets of solutions of some nonlinear elliptic equations. Proc. R. Soc. Edinb., Sect. A, Math. **134**, 363–373 (2004)
10. Makar-Limanov, L.: Solutions for the Dirichlet's problem for the equation $\Delta u = -1$ in a convex region. Notes Acad. Sci. USSR. **9**, 52–53 (1971) (Math)
11. Müller, F.: On the continuation of solutions for elliptic equations in two variables. Ann. Inst. H. Poincaré Anal. Non Linéaire. **19**, 745–776 (2002)
12. Radulescu, V.: Qualitative Analysis of Nonlinear Elliptic Partial Differential Equations: Monotonicity, Analytic, and Variational Methods. Hindawi. New York (2008)
13. Serrin, J.: A symmetry problem in potential theory. Arch. Rational Mech. **36**, 304–318 (1971)
14. Serrin, J., Pucci, P.: The Maximum Principle. Birkhauser, Bassel-Boston-Berlin (2007)

Sub-Riemannian Geodesics in the Octonionic H-type Group

Christian Autenried and Mauricio Godoy Molina

Abstract In this chapter, we study sub-Riemannian geodesics in the octonionic H-type group G_7^1, which is a nilpotent group of step 2 and, as a manifold, diffeomorphic to $\mathbb{R}^{15}$.

The Lie group structure of G_7^1, obtained via the Cayley–Dickson construction of real division algebras, induces a natural Riemannian metric and a bracket-generating distribution $\mathcal{H}$ of rank eight and step 2 on G_7^1. Restricting the metric to $\mathcal{H}$ we obtain a sub-Riemannian structure on G_7^1.

The class of curves we are interested in are horizontal with respect to $\mathcal{H}$ and, most importantly, critical points of the natural sub-Riemannian length functional. We present a characterization of these critical points via a differential equation, similar to the geodesic equation in Riemannian geometry, which states that for critical points of the length functional the intrinsic acceleration $\nabla_{\dot\gamma}\dot\gamma$ is a linear combination with constant coefficients of some special rotations of the velocity $\dot\gamma$.

Keywords H-type group · First variation of length · Sub-Riemannian geodesics · Geodesic equation

1 Introduction

The H(eisenberg)-type Lie algebras were introduced by A. Kaplan in his foundational work [7]. Their Lie algebra structure is intimately related to the existence of a Clifford algebra representation over a certain inner product space. To make this claim more precise, recall that a composition of two positive definite real quadratic forms φ

C. Autenried (✉) · M. Godoy Molina
Department of Mathematics, University of Bergen, Bergen, Norway
e-mail: christian.autenried@math.uib.no

M. Godoy Molina
Departamento de Matemática y Estadística,
Universidad de la Frontera, Casilla 54-D, 4780000 Temuco, Chile
e-mail: mauricio.godoy@math.uib.no

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_8

and λ on two vector spaces H and U, respectively, is a bilinear map $\mu : H \times U \to H$ such that for any $h \in H$, $u \in U$,

$$\varphi(h)\lambda(u) = \varphi(\mu(h,u)).$$

One can always assume there exists a vector $u_0 \in U$ such that $\mu(h,u_0) = h$ for all $h \in H$. Setting V as the orthogonal complement of $\mathbb{R}u_0$ in U, one can introduce a Lie bracket $[\cdot,\cdot] : H \times H \to V$ that induces a Lie algebra structure of step 2 on $H \oplus V$. The Clifford algebra representation mentioned before refers to the fact that

$$\mu(\mu(h,v),v) = -\lambda(v)h,$$

i.e., the existence of μ induces an H-representation of the Clifford algebra $C\ell(V,-\lambda)$.

Among the plethora of H-type Lie algebras, one can distinguish the class of those satisfying the so-called J^2-condition, which is Clifford algebraic in its very nature. This family of algebras was introduced in [4], and has been the subject of intense study by analysts for the past 20 years. A major result, obtained in the previous reference, is the fact that the nilpotent, connected, and simply connected groups corresponding to H-type Lie algebras can be singled out as those appearing in Iwasawa decompositions of real rank one simple Lie groups, and thus, there are but a few classes of H-type Lie algebras satisfying the J^2-condition. These families of H-type algebras are the trivial Euclidean spaces $\mathbb{R}^n$, the Heisenberg Lie algebras $\mathfrak{g}_1^{2n+1}$, the quaternionic H-type algebras $\mathfrak{g}_3^{4n+3}$, and the octonionic H-type algebra $\mathfrak{g}_7^1$. Note that, although there are nontrivial H-type Lie algebras with centers of arbitrary dimension [7, Corollary 1], those that satisfy the J^2-condition are either abelian or have centers of dimension 1, 3, and 7.

There is a natural connection between H-type Lie algebras and sub-Riemannian geometry, which we proceed to explain. Recall that a sub-Riemannian manifold is a triplet $(M,\mathcal{H},\langle\cdot,\cdot\rangle)$, where $\mathcal{H} \hookrightarrow TM$ is a distribution, i.e., a subbundle of the tangent bundle of M, and $\langle\cdot,\cdot\rangle$ is a fiber inner product defined on $\mathcal{H}$ called the sub-Riemannian metric. For most applications, it is assumed that the distribution $\mathcal{H}$ is bracket generating, that is,

$$\text{Lie}\,\mathcal{H} = \text{ Lie algebra generated by sections of } \mathcal{H} = \Gamma(TM),$$

where $\Gamma(TM)$ denotes the space of vector fields on M. The step of $\mathcal{H}$ is, by convention, the minimal length of brackets needed to generate all the vector fields on M plus one. Associated to an H-type Lie algebra $\mathfrak{g} = H \oplus V$ there is a unique (up to isomorphism) connected and simply connected Lie group G with Lie algebra $\mathfrak{g}$. By left-translating the subspace H of $\mathfrak{g}$, we obtain a bracket-generating distribution $\mathcal{H} \hookrightarrow TG$ of step 2. The quadratic form φ induces a sub-Riemannian metric on $\mathcal{H}$.

From now on, we focus our attention on the sub-Riemannian octonionic H-type group, that is, the sub-Riemannian structure defined on the connected and simply connected Lie group G_7^1 with Lie algebra $\mathfrak{g}_7^1$. The main purpose of this note is to give a variational description of the critical points of the length functional:

$$L(\gamma) = \int \sqrt{\varphi(\dot\gamma(t))}\,\mathrm{d}t$$

defined for horizontal curves in G_7^1, that is, piecewise smooth curves γ whose velocity vector satisfies the constraint $\dot{\gamma}(t) \in \mathcal{H}_{\gamma(t)}$, whenever $\dot{\gamma}$ is defined. We will refer to these critical points as sub-Riemannian geodesics. An alternative description of these curves, from a Hamiltonian point of view, has been obtained in [2]. Let us stress the fact that we use their model of the group G_7^1, which is obtained from the Cayley–Dickson construction of division algebras, instead of the Clifford algebraic model defined in [4].

This chapter is organized as follows. In Sect. 2, we recall briefly the definition and main properties of the octonionic H-type group and its natural sub-Riemannian structure, following [2]. In Sect. 3, we prove the main result of this chapter, following the lines of [5, 10]. The two major difficulties to overcome when dealing with G_7^1 are the fact that as a manifold it is 15-dimensional and that underlying its structure we are using the octonions, the only normed division algebra which is nonassociative. Finally, we conclude with two appendices, where we collect all the formulas that are too large to be displayed in an aesthetically pleasing way within the main line of argumentation.

2 The Octonionic H-type Group G_7^1

In this section, we give a short introduction to the octonionic H-type Lie algebra $\mathfrak{g}_7^1$ and the sub-Riemannian geometry of its (unique connected and simply connected) Lie group G_7^1, both concretely realized in $\mathbb{R}^{15}$. For a deeper study, and some interesting facts about its horizontal curves, we recommend [2].

Let us start by giving a description of $\mathfrak{g}_7^1$ through vector fields defined on $\mathbb{R}^{15} = \mathbb{R}^8 \oplus \mathbb{R}^7$, with coordinates $x_1, \ldots, x_8, z_1, \ldots, z_7$. Consider the 8×8 matrices $\mathcal{J}_1, \ldots, \mathcal{J}_7$ with real coefficients given in Appendix 1. The horizontal space $H = \operatorname{span}\{X_1, \ldots, X_8\}$ corresponds to the distribution generated by the vector fields:

$$X_l(x,z) = \partial_{x_l} + \frac{1}{2}\sum_{m=1}^{7} (x\mathcal{J}_m)_l \partial_{z_m}, \quad l \in \{1, \ldots, 8\},$$

where $x = (x_1, \ldots, x_8)$ and $(x\mathcal{J}_m)_l$ denotes the lth coordinate of the row vector $x\mathcal{J}_m$. Explicitly, these vector fields are given by

$$X_1(x,z) = \partial x_1 + \frac{1}{2}(-x_2\partial_{z_1} - x_3\partial_{z_2} - x_4\partial_{z_3} - x_5\partial_{z_4} - x_6\partial_{z_5} - x_7\partial_{z_6} - x_8\partial_{z_7}),$$

$$X_2(x,z) = \partial x_2 + \frac{1}{2}(x_1\partial_{z_1} + x_4\partial_{z_2} - x_3\partial_{z_3} + x_6\partial_{z_4} - x_5\partial_{z_5} - x_8\partial_{z_6} + x_7\partial_{z_7}),$$

$$X_3(x,z) = \partial x_3 + \frac{1}{2}(-x_4\partial_{z_1} + x_1\partial_{z_2} + x_2\partial_{z_3} + x_7\partial_{z_4} + x_8\partial_{z_5} - x_5\partial_{z_6} - x_6\partial_{z_7}),$$

$$X_4(x,z) = \partial x_4 + \frac{1}{2}(x_3\partial_{z_1} - x_2\partial_{z_2} + x_1\partial_{z_3} + x_8\partial_{z_4} - x_7\partial_{z_5} + x_6\partial_{z_6} - x_5\partial_{z_7}),$$

$$X_5(x,z) = \partial x_5 + \frac{1}{2}(-x_6\partial_{z_1} - x_7\partial_{z_2} - x_8\partial_{z_3} + x_1\partial_{z_4} + x_2\partial_{z_5} + x_3\partial_{z_6} + x_4\partial_{z_7}),$$

$$X_6(x,z) = \partial x_6 + \frac{1}{2}(x_5\partial_{z_1} - x_8\partial_{z_2} + x_7\partial_{z_3} - x_2\partial_{z_4} + x_1\partial_{z_5} - x_4\partial_{z_6} + x_3\partial_{z_7}),$$

$$X_7(x,z) = \partial x_7 + \frac{1}{2}(x_8\partial_{z_1} + x_5\partial_{z_2} - x_6\partial_{z_3} - x_3\partial_{z_4} + x_4\partial_{z_5} + x_1\partial_{z_6} - x_2\partial_{z_7}),$$

$$X_8(x,z) = \partial x_8 + \frac{1}{2}(-x_7\partial_{z_1} + x_6\partial_{z_2} + x_5\partial_{z_3} - x_4\partial_{z_4} - x_3\partial_{z_5} + x_2\partial_{z_6} + x_1\partial_{z_7}).$$

The vertical distribution V, i.e., the center of the Lie algebra $\mathfrak{g}_7^1$, is defined by

$$V = \operatorname{span}\{Z_1, \ldots, Z_7\},$$

where $Z_i(x,z) = \partial_{z_i}$. The Lie algebra $\mathfrak{g}_7^1$ is the algebra spanned by the vector fields $X_1, \ldots, X_8, Z_1, \ldots, Z_7$ with the usual commutator of vector fields in $\mathbb{R}^{15}$, see Table 1.

Table 1 Nontrivial Lie bracket relations in $\mathfrak{g}_7^1$

[Row, col.]	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	0	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7
X_2	$-Z_1$	0	Z_3	$-Z_2$	Z_5	$-Z_4$	$-Z_7$	Z_6
X_3	$-Z_2$	$-Z_3$	0	Z_1	Z_6	Z_7	$-Z_4$	$-Z_5$
X_4	$-Z_3$	Z_2	$-Z_1$	0	Z_7	$-Z_6$	Z_5	$-Z_4$
X_5	$-Z_4$	$-Z_5$	$-Z_6$	$-Z_7$	0	Z_1	Z_2	Z_3
X_6	$-Z_5$	Z_4	$-Z_7$	Z_6	$-Z_1$	0	$-Z_3$	Z_2
X_7	$-Z_6$	Z_7	Z_4	$-Z_5$	$-Z_2$	Z_3	0	$-Z_1$
X_8	$-Z_7$	$-Z_6$	Z_5	Z_4	$-Z_3$	$-Z_2$	Z_1	0

The Lie group G_7^1 is the nilpotent Lie group structure on $\mathbb{R}^{15}$ of step 2 induced by the Lie algebra $\mathfrak{g}_7^1$ via the Baker–Campbell–Hausdorff formula. An explicit expression for the product rule can be found in [2, Eq. (3.7)].

We define an inner product $\langle \cdot, \cdot \rangle$ on $\mathfrak{g}_7^1$ such that the vector fields $X_1, \ldots, X_8$, $Z_1, \ldots, Z_7$ form an orthonormal frame. The left invariant distribution:

$$\mathcal{H} := \operatorname{span}\{X_1, \ldots, X_8\},$$

and the restriction of $\langle \cdot, \cdot \rangle$ to $\mathcal{H}$ give us the sub-Riemannian structure on G_7^1 we want to study further. The group G_7^1 with the structure introduced before is called the octonionic H-type group, since the map:

$$\operatorname{ad}_X : \ker(\operatorname{ad}_X)^\perp \subset H \to V,$$

is a surjective isometry for any $X \in H$ of norm one, see [7]. From this definition, it follows immediately that the distribution $\mathcal{H}$ is strongly bracket generating and, thus, all length-minimizing curves are normal, i.e., they all solve a natural Hamiltonian

equation, see [8, Chap. 1]. Explicit solutions to this equation in the case of the octonionic H-type group can be found in [2]. The method employed to find these solutions in [2] uses explicitly the coordinates of $\mathbb{R}^{15}$, instead our approach is entirely coordinate free.

With all these ingredients at hand, we can compute explicitly the Levi–Civita connection of the metric $\langle \cdot , \cdot \rangle$. To do this, we employ the well-known Koszul formula:

$$\begin{aligned}\langle Z, \nabla_Y X\rangle = \frac{1}{2}(&X\langle Y, Z\rangle + Y\langle Z, X\rangle - Z\langle X, Y\rangle \\ &- \langle [X, Z], Y\rangle - \langle [Y, Z], X\rangle - \langle [X, Y], Z\rangle),\end{aligned}$$

and we immediately notice that the following equations:

$$\langle X_b, \nabla_{X_a} Z_r\rangle = -\frac{1}{2}\langle [X_a, X_b], Z_r\rangle, \quad \langle Z_s, \nabla_{X_a} Z_r\rangle = 0,$$

hold, for all $a, b \in \{1, \dots, 8\}$, $r, s \in \{1, \dots, 7\}$. We conclude that $\nabla_{X_a} Z_r$ has trivial vertical part, and thus

$$\nabla_{X_a} Z_r = -\frac{1}{2}\sum_{b=1}^{8} \langle [X_a, X_b], Z_r\rangle X_b.$$

From this and the information in Table 1, we can deduce the expressions found in Appendix 2. From these, it is natural to define the operators $J_r : \mathcal{H} \to \mathcal{H}$, $r \in \{1, \dots, 7\}$, by

$$J_r(X) := 2\nabla_X Z_r, \quad r \in \{1, \dots, 7\}.$$

These are almost complex structures on $\mathcal{H}$, i.e., $J_r^2 = -Id|_{\mathcal{H}}$, with the property that

$$\langle J_r(X), Y\rangle + \langle X, J_r(Y)\rangle = 0, \tag{1}$$

for every $r \in \{1, \dots, 7\}$ and all $X, Y \in \mathcal{H}$. Furthermore, we note that this equation implies that $\langle X, J_r(X)\rangle = 0$, for all $X \in \mathcal{H}$.

3 Geodesic Equation on G_7^1

In this section, we follow the arguments in [5, 9, 10] to find an intrinsic differential equation for the sub-Riemannian geodesics of G_7^1 with respect to the sub-Riemannian structure introduced in Sect. 2. An earlier attempt to this problem can be found in [11], where the author obtained a differential equation for geodesics in CR sub-Riemannian three-manifolds using the Tanaka–Webster connection. We conclude with some examples and interpretations.

3.1 Main Result

Recall that a piecewise smooth curve $\gamma : [a,b] \to G_7^1$ is called horizontal if $\dot{\gamma}(s) \in \mathcal{H}_{\gamma(s)}$, whenever $\dot{\gamma}$ is defined. A variation of a curve $\gamma : [a,b] \to G_7^1$ is a C^2-map $\tilde{\gamma} : [a,b] \times I \to G_7^1$, where I is an open interval containing 0 and $\tilde{\gamma}(s,0) = \gamma(s)$. As customary, we will denote $\tilde{\gamma}(s,\varepsilon) = \gamma_\varepsilon(s)$. If γ is horizontal, we say that $\tilde{\gamma}$ is an admissible variation if all curves $\gamma_\varepsilon : [a,b] \to G_7^1$ are horizontal, $\gamma_\varepsilon(a) = \gamma(a)$ and $\gamma_\varepsilon(b) = \gamma(b)$. As an abuse of notation, we call γ_ε an admissible variation of γ.

Given a vector $v \in \mathfrak{g}_7^1$, we write v_H for its orthogonal projection to the horizontal space H. We will use the same notation for the horizontal components of vector fields, vector fields along curves, etc.

Lemma 1 *Let $\gamma : [a,b] \to G_7^1$ be a horizontal curve parameterized by arc length, and let W be any C^1 vector field along γ such that $W(a) = W(b) = 0$ satisfying*

$$0 = \dot{\gamma}\langle W, Z_r\rangle - 2\langle W_H, J_r(\dot{\gamma})\rangle, \quad r \in \{1,\dots,7\}. \tag{2}$$

Then there exists an admissible variation γ_ε of γ such that $\frac{\partial}{\partial\varepsilon}|_{\varepsilon=0}\gamma_\varepsilon(s) = W$.

Proof Note that there exists a vector field $\tilde{W}$ along γ, orthogonal to $\dot{\gamma}$, such that we can write $W = f\dot{\gamma} + \tilde{W}$ for some smooth function f satisfying $f(a) = f(b) = 0$. From the choice of $\tilde{W}$, the definition of the almost complex structures J_r, the arc length parameterization and horizontality of γ, we can immediately see that

$$\begin{aligned}
\langle W, \dot{\gamma}\rangle &= f\langle \dot{\gamma}, \dot{\gamma}\rangle + \langle \tilde{W}, \dot{\gamma}\rangle = f \\
\langle W, J_r(\dot{\gamma})\rangle &= f\langle \dot{\gamma}, J_r(\dot{\gamma})\rangle + \langle \tilde{W}, J_r(\dot{\gamma})\rangle = \langle \tilde{W}, J_r(\dot{\gamma})\rangle, \\
\langle W, Z_r\rangle &= f\langle \dot{\gamma}, Z_r\rangle + \langle \tilde{W}, Z_r\rangle = \langle \tilde{W}, Z_r\rangle,
\end{aligned}$$

for all $r \in \{1,\dots,7\}$.

It is easy to see that if there exists a (not necessarily admissible) variation $\gamma(s,\varepsilon)$ for which $\frac{\partial}{\partial\varepsilon}|_{\varepsilon=0}\gamma(s,\varepsilon) = \tilde{W}$, then there exists $\gamma_1(s,\varepsilon)$ satisfying $\frac{\partial}{\partial\varepsilon}|_{\varepsilon=0}\gamma_1(s,\varepsilon) = W$. This implies that, without loss of generality, we can and will assume that $W \perp \dot{\gamma}$.

We have to distinguish the cases in which the vector field W is horizontal or not. Let us first examine the case when W is horizontal on some nonempty interval $I_0 \subset [a,b]$. By definition, we have that $W = W_H$ for all $s \in I_0$, and since we are assuming that W satisfies condition (2), we have the equalities:

$$\langle W_H, J_r(\dot{\gamma})\rangle = \langle W, J_r(\dot{\gamma})\rangle = \frac{1}{2}\,\dot{\gamma}\langle W_H, Z_r\rangle = 0,$$

for all $r \in \{1,\dots,7\}$. This implies that $W_H \in \operatorname{span}\{\dot{\gamma}\}$, and since W_H is also orthogonal to $\dot{\gamma}$, we can conclude that $W_H = 0$.

The nonhorizontal case requires more care. If exp is the exponential map associated to the (Riemannian) metric $\langle\cdot,\cdot\rangle$ on G_7^1, we can define the mapping

$$F(s,\varepsilon) = \exp_{\gamma(s)}(\varepsilon W(s))$$

for sufficiently small $\varepsilon > 0$ and $s \in [a, b]$. Let us assume there exists $s_0 \in [a, b]$ such that $W(s_0) \notin \mathcal{H}_{\gamma(s_0)}$. We note that $F(s, \varepsilon)$ defines locally a surface, which is transverse to the horizontal space $\mathcal{H}_{\gamma(s_0)}$, as it contains curves in non-horizontal directions by definition. Furthermore, it is foliated by horizontal curves. These two facts together imply that there exists a function $g(s, \varepsilon)$ of class C^2 such that we can define a family of horizontal curves:

$$\gamma_\varepsilon(s) = \exp_{\gamma(s)}(g(s, \varepsilon)W(s)).$$

If we choose g such that $\frac{\partial}{\partial \varepsilon}|_{\varepsilon=0} g(s_0, \varepsilon) = 1$, it follows that γ_ε is an admissible variation of γ with associated vector field W. □

As simple computation shows that the converse of Lemma 1 also holds. For completeness, we include it here. Given an admissible variation γ_ε of a horizontal curve γ with variational vector field W, then

$$0 = \dot{\gamma}\langle W, Z_r\rangle - 2\langle W_H, J_r(\dot{\gamma})\rangle, \quad r \in \{1, \ldots, 7\}.$$

Since $\langle \dot{\gamma}_\epsilon , Z_r\rangle = 0$, for all $r \in \{1, \ldots, 7\}$, it follows trivially that $\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\big|_{\varepsilon=0} \langle \dot{\gamma}_\varepsilon , Z_r\rangle = 0$. From this equality, the fact that $\nabla_{Z_l} Z_r = 0$ for all $r, l \in \{1, \ldots, 7\}$, and Eq. 1, we deduce that

$$\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} \langle \dot{\gamma}_\varepsilon, Z_r\rangle = \langle \nabla_W \dot{\gamma}, Z_r\rangle + \langle \dot{\gamma}, \nabla_W Z_r\rangle \\
&= \langle \nabla_{\dot{\gamma}} W, Z_r\rangle + \langle \dot{\gamma}, \nabla_{W_H} Z_r\rangle \\
&= \dot{\gamma}\langle W, Z_r\rangle - \langle W, \nabla_{\dot{\gamma}} Z_r\rangle + \langle \dot{\gamma}, J_r(W_H)\rangle \\
&= \dot{\gamma}\langle W, Z_r\rangle - \langle W_H, J_r(\dot{\gamma})\rangle - \langle J_r(\dot{\gamma}), W_H\rangle \\
&= \dot{\gamma}\langle W, Z_r\rangle - 2\langle W_H, J_r(\dot{\gamma})\rangle.
\end{aligned}$$

Now, we have all tools to prove the main theorem.

Theorem 1 *Let $\gamma : [a, b] \to G_7^1$ be a horizontal curve of class C^2, parametrized by arc length. Then γ is a critical point of the length functional (with respect to admissible variations) if, and only if, there exist constants $\lambda_1, \ldots, \lambda_7 \in \mathbb{R}$ such that γ satisfies the second-order differential equation:*

$$\nabla_{\dot{\gamma}} \dot{\gamma} - 2 \sum_{r=1}^{7} \lambda_r J_r(\dot{\gamma}) = 0. \tag{3}$$

Proof Let us first assume that $\gamma : [a, b] \to G_7^1$ is a horizontal curve, parametrized by arc length, satisfying Eq. (3) for some constants $\lambda_1, \ldots, \lambda_7 \in \mathbb{R}$. We consider a C^1-smooth vector field W, vanishing at the end points of γ and satisfying

$$\dot{\gamma}\langle W, Z_r\rangle = 2\langle W_H, J_r(\dot{\gamma})\rangle, \tag{4}$$

for all $r \in \{1, \dots, 7\}$. It is well known, see [3], that the length functional L satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\bigg|_{\varepsilon=0} L(\gamma_\varepsilon) = -\int_a^b \langle \nabla_{\dot\gamma}\dot\gamma, W\rangle,$$

therefore, to prove that γ is a critical point of L with respect to admissible variations, we need to show that $\int_a^b \langle \nabla_{\dot\gamma}\dot\gamma, W\rangle = 0$. Decompose $W = W_H + W_V$ in its horizontal and vertical parts, where $W_V = \sum_{r=1}^7 g_r Z_r$ for some smooth functions $g_1, \dots, g_7$ satisfying $g_r(a) = g_r(b) = 0$. Then,

$$\begin{aligned}
\int_a^b \langle \nabla_{\dot\gamma}\dot\gamma, W\rangle &\overset{(3)}{=} 2\sum_{r=1}^7 \lambda_r \int_a^b \langle J_r(\dot\gamma), W\rangle \overset{J_r(\dot\gamma)\in\mathcal{H}}{=} 2\sum_{r=1}^7 \lambda_r \int_a^b \langle J_r(\dot\gamma), W_H\rangle \\
&\overset{(4)}{=} \sum_{r=1}^7 \lambda_r \int_a^b \dot\gamma\langle W, Z_r\rangle \overset{Z_r\in V}{=} \sum_{r=1}^7 \lambda_r \int_a^b \dot\gamma\langle W_V, Z_r\rangle \\
&= \sum_{r=1}^7 \lambda_r \int_a^b \dot\gamma \left\langle \sum_{\ell=1}^7 g_\ell Z_\ell, Z_r\right\rangle = \sum_{r=1}^7 \lambda_r \int_a^b \dot\gamma(g_r) \\
&= \sum_{r=1}^7 \lambda_r \int_a^b \frac{\mathrm{d}}{\mathrm{d}t}(g_r(t)) \overset{g_r(a)=g_r(b)=0}{=} 0.
\end{aligned}$$

For the converse, let γ be a critical point of the length functional, which is horizontal and parametrized by arc length. This implies that

$$0 = \frac{\mathrm{d}}{\mathrm{d}\epsilon}\bigg|_{\epsilon=0} L(\gamma_\epsilon) = -\int_a^b \langle \nabla_{\dot\gamma}\dot\gamma, W\rangle,$$

where W is the vector field of the variation γ_ϵ.

We know that the condition $\|\dot\gamma\|^2 = \langle\dot\gamma, \dot\gamma\rangle = 1$ implies

$$\langle \nabla_{\dot\gamma}\dot\gamma, \dot\gamma\rangle = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\langle\dot\gamma, \dot\gamma\rangle = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}1 = 0.$$

Furthermore, since γ is horizontal, then $\langle\dot\gamma, Z_r\rangle = 0$ for all $r \in \{1, \dots, 7\}$, and thus

$$\begin{aligned}
0 &\overset{\langle\dot\gamma, Z_r\rangle=0}{=} \dot\gamma\langle\dot\gamma, Z_r\rangle = \langle\nabla_{\dot\gamma}\dot\gamma, Z_r\rangle + \langle\dot\gamma, \nabla_{\dot\gamma} Z_r\rangle = \langle\nabla_{\dot\gamma}\dot\gamma, Z_r\rangle + \langle\dot\gamma, J_r(\dot\gamma)\rangle \\
&\overset{\langle X, J_r(X)\rangle=0}{=} \langle\nabla_{\dot\gamma}\dot\gamma, Z_r\rangle,
\end{aligned}$$

for all $r \in \{1, \dots, 7\}$. In summary, we have shown that $\nabla_{\dot\gamma}\dot\gamma \perp \dot\gamma$ and $\nabla_{\dot\gamma}\dot\gamma \perp Z_r$ for all $r \in \{1, \dots, 7\}$. Therefore, the vector field $\nabla_{\dot\gamma}\dot\gamma$ has to be contained in the seven dimensional subspace $\operatorname{span}\{J_1(\dot\gamma), \dots, J_7(\dot\gamma)\}$, that is

$$\nabla_{\dot\gamma}\dot\gamma = \sum_{r=1}^7 g_r J_r(\dot\gamma).$$

It remains to show that the functions g_r are in fact constant. We fix $f_r : [a, b] \to \mathbb{R}$ for $r \in \{1, \ldots, 7\}$ such that $f_r(a) = f_r(b) = 0$ and $\int_a^b f_r = 0$. Furthermore, we consider a vector field $\tilde{W}$ such that its horizontal part satisfies $\tilde{W}_H = \Sigma_{r=1}^7 f_r J_r(\dot{\gamma})$ and satisfies $\langle \tilde{W}, Z_r \rangle(s) = 2 \int_a^s f_r(t) \mathrm{d}t$.

The last condition for the vertical part of $\tilde{W}$ yields:

$$\dot{\gamma} \langle \tilde{W}, Z_r \rangle = \frac{\mathrm{d}}{\mathrm{d}s} \left(2 \int_a^s f_r(t) \mathrm{d}t \right) = 2 f_r(s),$$

for all $r \in \{1, \ldots, 7\}$. The horizontal condition and the orthonormality of the family $\{J_1(\dot{\gamma}), \ldots, J_7(\dot{\gamma})\}$, see Appendix 2, imply

$$\langle \tilde{W}_H, J_r(\dot{\gamma}) \rangle = \left\langle \sum_{l=1}^7 f_l J_l(\dot{\gamma}), J_r(\dot{\gamma}) \right\rangle = f_r(s),$$

for all $r \in \{1, \ldots, 7\}$. These two equations together imply the condition (2) of Lemma 1, which reads

$$\dot{\gamma} \langle \tilde{W}, Z_r \rangle = 2 \langle \tilde{W}_H, J_r(\dot{\gamma}) \rangle,$$

for all $r \in \{1, \ldots, 7\}$. Using Lemma 1, we conclude that $\tilde{W}$ is a vector field for an admissible variation of γ. We obtain:

$$0 = \int_a^b \langle \nabla_{\dot{\gamma}} \dot{\gamma}, \tilde{W} \rangle = \sum_{r=1}^7 \int_a^b f_r \langle \nabla_{\dot{\gamma}} \dot{\gamma}, J_r(\dot{\gamma}) \rangle,$$

which is valid for any seven functions with mean zero, which implies that $\langle \nabla_{\dot{\gamma}} \dot{\gamma}, J_r(\dot{\gamma}) \rangle$ is constant for all $r \in \{1, \ldots, 7\}$. We obtain Eq. (3) for suitable constants $\lambda_1, \ldots, \lambda_7 \in \mathbb{R}$. □

3.2 *Interpretations and Examples*

Similar equations to the one in our main theorem can be found in the literature in different guises, and with various geometric and physical interpretations.

As mentioned in [6], when studying the case of the natural CR sub-Riemannian structure on the three-dimensional sphere S^3, the admissible C^2 critical points of the length functional satisfy the equation:

$$\nabla_{\dot{\gamma}} \dot{\gamma} + 2\lambda J(\dot{\gamma}) = 0, \tag{5}$$

where J is the almost complex structure on the horizontal distribution of S^3 induced by the CR structure. In that case, the constant λ corresponds to a curvature in the following sense: if γ solves Eq. 5 with parameter λ, then the projection of γ to S^2

via the Hopf fibration produces a piece of a geodesic circle with constant geodesic curvature λ (see [6, Lemma 3.2]).

In the case of Theorem 1, after a rather tedious computation, we can show that the curves in G_7^1 starting from the origin and satisfying Eq. 3 with $\lambda_1 = \cdots = \lambda_7 = 0$ are straight lines in $\mathbb{R}^{15}$ contained in the eight-plane $z_1 = \cdots = z_7 = 0$. This fact indicates that we can again interpret the constants as curvatures. In a sense, the values of $\lambda_1, \ldots, \lambda_7$ measure how far are the curves solving (3) from being a Riemannian geodesic. We are currently working on making this claim precise and applying it to all the similar cases known to us.

Finally, it is of worth mentioning this equation has a very similar structure to the so-called Wong's equation, see [8, Chap. 12], which corresponds to a nonabelian version of Lorentz equations for the dynamics of a particle. In that case, the parameter λ corresponds to the charge of the particle which satisfies an additional restriction in the form of an evolution equation. It would be of interest to study the precise relation between Wong's equation and the general formulation of critical points of length in sub-Riemannian manifolds with transverse symmetries, see [1].

Acknowledgment The authors are partially supported by the NFR-FRINAT grants #204726/V30 and #213440/BG.

We would like to thank Professor Irina Markina for pointing out some subtle differences between the models for the octonionic H-type group in [2] and [4]. The second author would like to thank Olga Vasilieva and the rest of the organizing committee of the conference ICAMI 2013 for their warm hospitality and kindness while visiting Colombia. Also, we thank the anonymous referees for their comments and remarks.

Finally, we would like to mention that, during the refereeing process for this chapter, Prof. Fabrice Baudoin informed us that our main theorem is contained as a special case of Proposition 2.14 in the manuscript [1].

Appendix 1

We present the matrices $\mathcal{J}_1, \ldots, \mathcal{J}_7$ used in Sect. 2 to define the vector fields $X_1, \ldots, X_8$.

$$\mathcal{J}_1 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix},$$

$$\mathcal{J}_2 = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \end{pmatrix},$$

$$\mathcal{J}_3 = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathcal{J}_4 = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathcal{J}_5 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathcal{J}_6 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathcal{J}_7 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Appendix 2

Here we present the components of the Levi–Civita connection for the Riemannian metric on G_7^1 defined in Sect. 2.

$$
\begin{aligned}
&\nabla_{X_1} Z_1 = -\frac{1}{2} X_2, \quad \nabla_{X_2} Z_1 = \frac{1}{2} X_1, \quad \nabla_{X_3} Z_1 = -\frac{1}{2} X_4, \quad \nabla_{X_4} Z_1 = \frac{1}{2} X_3,\\
&\nabla_{X_5} Z_1 = -\frac{1}{2} X_6, \quad \nabla_{X6} Z_1 = \frac{1}{2} X_5, \quad \nabla_{X_7} Z_1 = \frac{1}{2} X_8, \quad \nabla_{X_8} Z_1 = -\frac{1}{2} X_7,\\
&\nabla_{X_1} Z_2 = -\frac{1}{2} X_3, \quad \nabla_{X_2} Z_2 = \frac{1}{2} X_4, \quad \nabla_{X_3} Z_2 = \frac{1}{2} X_1, \quad \nabla_{X_4} Z_2 = -\frac{1}{2} X_2,\\
&\nabla_{X_5} Z_2 = -\frac{1}{2} X_7, \quad \nabla_{X_6} Z_2 = -\frac{1}{2} X_8, \quad \nabla_{X_7} Z_2 = \frac{1}{2} X_5, \quad \nabla_{X_8} Z_2 = \frac{1}{2} X_6,\\
&\nabla_{X_1} Z_3 = -\frac{1}{2} X_4, \quad \nabla_{X_2} Z_3 = -\frac{1}{2} X_3, \quad \nabla_{X_3} Z_3 = \frac{1}{2} X_2, \quad \nabla_{X_4} Z_3 = \frac{1}{2} X_1,\\
&\nabla_{X_5} Z_3 = -\frac{1}{2} X_8, \quad \nabla_{X_6} Z_3 = \frac{1}{2} X_7, \quad \nabla_{X_7} Z_3 = -\frac{1}{2} X_6, \quad \nabla_{X_8} Z_3 = \frac{1}{2} X_5,\\
&\nabla_{X_1} Z_4 = -\frac{1}{2} X_5, \quad \nabla_{X_2} Z_4 = \frac{1}{2} X_6, \quad \nabla_{X_3} Z_4 = \frac{1}{2} X_7, \quad \nabla_{X_4} Z_4 = \frac{1}{2} X_8,\\
&\nabla_{X_5} Z_4 = \frac{1}{2} X_1, \quad \nabla_{X_6} Z_4 = -\frac{1}{2} X_2, \quad \nabla_{X_7} Z_4 = -\frac{1}{2} X_3, \quad \nabla_{X_8} Z_4 = -\frac{1}{2} X_4,\\
&\nabla_{X_1} Z_5 = -\frac{1}{2} X_6, \quad \nabla_{X_2} Z_5 = -\frac{1}{2} X_5, \quad \nabla_{X_3} Z_5 = \frac{1}{2} X_8, \quad \nabla_{X_4} Z_5 = -\frac{1}{2} X_7,\\
&\nabla_{X_5} Z_5 = \frac{1}{2} X_2, \quad \nabla_{X_6} Z_5 = \frac{1}{2} X_1, \quad \nabla_{X_7} Z_5 = \frac{1}{2} X_4, \quad \nabla_{X_8} Z_5 = -\frac{1}{2} X_3,\\
&\nabla_{X_1} Z_6 = -\frac{1}{2} X_7, \quad \nabla_{X_2} Z_6 = -\frac{1}{2} X_8, \quad \nabla_{X_3} Z_6 = -\frac{1}{2} X_5, \quad \nabla_{X_4} Z_6 = \frac{1}{2} X_6,\\
&\nabla_{X_5} Z_6 = \frac{1}{2} X_3, \quad \nabla_{X_6} Z_6 = -\frac{1}{2} X_4, \quad \nabla_{X_7} Z_6 = \frac{1}{2} X_1, \quad \nabla_{X_8} Z_6 = \frac{1}{2} X_2,\\
&\nabla_{X_1} Z_7 = -\frac{1}{2} X_8, \quad \nabla_{X_2} Z_7 = \frac{1}{2} X_7, \quad \nabla_{X_3} Z_7 = -\frac{1}{2} X_6, \quad \nabla_{X_4} Z_7 = -\frac{1}{2} X_5,\\
&\nabla_{X_5} Z_7 = \frac{1}{2} X_4, \quad \nabla_{X_6} Z_7 = \frac{1}{2} X_3, \quad \nabla_{X_7} Z_7 = -\frac{1}{2} X_2, \quad \nabla_{X_8} Z_7 = \frac{1}{2} X_1.
\end{aligned}
$$

References

1. Baudoin, F., Garofalo, N.: Generalized Bochner formulas and Ricci lower bounds for sub-Riemannian manifolds of rank two. arXiv:0904.1623
2. Calin, O., Chang, D.-C., Markina, I.: Geometric analysis on H-type groups related to division algebras. Math. Nachr. **282**(1), 44–68 (2009)
3. Cheeger, J., Ebin, D. G.: Comparison theorems in Riemannian geometry, 168 p. Revised reprint of the 1975 original. AMS Chelsea Publishing, Providence (2008). ISBN: 978-0-8218-4417-5 MR2394158 (2009c:53043))

4. Cowling, M., Dooley, A.H., Korányi, A., Ricci, F.: H-type groups and Iwasawa decompositions. Adv. Math. **87**(1), 1–41 (1991)
5. Godoy Molina, M., Markina, I.: Sub-Riemannian geodesics and heat operator on odd dimensional spheres. Anal. Math. Phys. **2**(2), 123–147 (2012)
6. Hurtado, A., Rosales, C.: Area-stationary surfaces inside the sub-Riemannian three-sphere. Math. Ann. **340**(3), 675–708 (2008)
7. Kaplan, A.: Fundamental solutions for a class of hypoelliptic PDE generated by composition of quadratic forms. Trans. Amer. Math. Soc. **258**(1), 147–153 (1980)
8. Montgomery, R.: A tour of subriemannian geometries, their geodesics and applications. Mathematical Surveys and Monographs, vol. 91. American Mathematical Society, Providence, RI, (2002)
9. Ritoré, M.: A proof by calibration of an isoperimetric inequality in the Heisenberg group H^n. Calc. Var. Partial Differ. Equat. **44**(1–2), 47–60 (2012)
10. Ritoré, M., Rosales, C.: Area-stationary surfaces in the Heisenberg group $\mathbb{H}^1$. Adv. Math. **219**(2), 633–671 (2008)
11. Rumin, M.: Formes différentielles sur les variétés de contact. J. Differential Geom. **39**(2), 281–330 (1994)

Regularization of Inverse Ill-Posed Problems with L^2-BV Penalizers and Applications to Signal Restoration

Gisela L. Mazzieri, Ruben D. Spies and Karina G. Temperini

Abstract Several generalizations of the traditional Tikhonov- Phillips regularization method for ill-posed inverse problems have been proposed during the past two decades. Many of these generalizations are based upon inducing stability throughout the use of different penalizers which allow the capturing of diverse properties of the exact solution (e.g., edges, discontinuities, borders, etc.). However, in some problems in which it is known that the regularity of the exact solution is heterogeneous and/or anisotropic, it is reasonable to think that a much better option could be the simultaneous use of two or more penalizers of different nature. Such is the case, for instance, in some image restoration problems in which preservation of edges, borders, or discontinuities is an important matter. In this work, we present some results on the simultaneous use of penalizers of L^2 and of bounded variation (BV) type. For particular cases, existence and uniqueness results are proved. Open problems are discussed and results to signal restoration problem are presented.

Keywords Inverse problem, Ill-posedness, Regularization, Tikhonov–Phillips, Bounded variation

---

G. L. Mazzieri (✉)
Instituto de Matemática Aplicada del Litoral, IMAL, CCT CONICET Santa Fe, Colectora Ruta Nac. 168, km. 472, Paraje El Pozo, 3000 Santa Fe, Argentina
e-mail: glmazzieri@santafe-conicet.gov.ar

Departamento de Matemática, Facultad de Bioquímica y Ciencias Biológicas, Universidad Nacional del Litoral, Santa Fe, Argentina

R. D. Spies
Instituto de Matemática Aplicada del Litoral, IMAL, CCT CONICET Santa Fe, Colectora Ruta Nac. 168, km. 472, Paraje El Pozo, 3000 Santa Fe, Argentina
e-mail: rspies@santafe-conicet.gov.ar

Departamento de Matemática, Facultad de Ingeniería Química, Universidad Nacional del Litoral, Santiago del Estero 2829, 3000 Santa Fe, Argentina

K. G. Temperini
Instituto de Matemática Aplicada del Litoral, IMAL, CCT CONICET, Santa Fe, Colectora Ruta Nac. 168, km. 472, Paraje El Pozo, 3000 Santa Fe, Argentina
e-mail: ktemperini@santafe-conicet.gov.ar

Departamento de Matemática, Facultad de Humanidades y Ciencias, Universidad Nacional del Litoral, Colectora Ruta Nac. 168, km. 472, Paraje El Pozo, 3000 Santa Fe, Argentina

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_9

1 Introduction and Preliminaries

For our general setting, we consider the problem of finding u in an equation of the form:

$$Tu = v, \tag{1}$$

where $T : \mathcal{X} \to \mathcal{Y}$ is a bounded linear operator between two infinite dimensional Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ (usually they are both function spaces), the range of T is non-closed and v is the data, which is supposed to be known, perhaps with a certain degree of error. It is well known that under these hypotheses, problem (1) is ill-posed in the sense of Hadamard [6] and it must be regularized before any attempt to approximate its solutions is made [5]. The most usual way of regularizing a problem is by means of the use of the *Tikhonov–Phillips regularization method* whose general formulation can be given within the context of an unconstrained optimization problem. In fact, given an appropriate penalizer $W(u)$ with domain $\mathcal{D} \subset \mathcal{X}$, the regularized solution obtained by the Tikhonov–Phillips method and such a penalizer is the minimizer u_α (provided it exits), over $\mathcal{D}$, of the functional:

$$J_{\alpha,W}(u) = \|Tu - v\|^2 + \alpha W(u), \tag{2}$$

where α is a positive constant called regularization parameter. For general penalizers W, sufficient conditions guaranteeing existence, uniqueness, and weak and strong stability of the minimizers under different types of perturbations were found in [9]. In the sequel, unless otherwise specified, we will assume that $\mathcal{X} = L^2(\Omega)$, were $\Omega \subset \mathbb{R}^n$, with $n = 1, 2$, or 3.

Each choice of an appropriate penalizer W originates a different regularization method producing a particular regularized solution possessing particular properties. Thus, for instance, the choice of $W(u) = \|u\|^2$ gives rise to the classical Tikhonov–Phillips method of order zero producing always smooth regularized approximations which approximate, as $\alpha \to 0^+$, the best approximate solution (i.e., the least squares solution of minimum norm) of problem (1) (see [5]). The order-one method corresponds to the choice of $W(u) = \||\nabla u|\|^2$. Similarly, the choice of $W(u) = \|u\|_{\mathrm{BV}}$ (where $\|\cdot\|_{\mathrm{BV}}$ denotes the total variation norm) results in the so-called bounded variation (BV) regularization method [1, 10]. The use of this penalizer is appropriate when preserving discontinuities or edges is an important matter. The method, however, has as a drawback that it tends to produce piecewise constant approximations and therefore, it will most likely be highly inappropriate near regions where the exact solution is smooth [3] producing the so-called staircasing effect.

In certain types of problems, particularly in those in which it is known that the regularity of the exact solution is heterogeneous and/or anisotropic, it is reasonable to think that using and spatially adapting two or more penalizers of different nature could be more convenient. During the past 15 years several regularization methods have been developed in light of this reasoning. Thus, for instance, in 1997 Blomgren et al. [2] proposed the use of the following penalizer, by using variable L^p spaces:

$$W(u) = \int_\Omega |\nabla u|^{p(|\nabla u|)} dx, \tag{3}$$

where $\lim_{u\to 0^+} p(u) = 2$, $\lim_{u\to\infty} p(u) = 1$ and p is a decreasing function. Thus, in regions where the modulus of the gradient of u is small the penalizer is approximately equal to $\||\nabla u|\|^2_{L^2(\Omega)}$ corresponding to a Tikhonov–Phillips method of order one (appropriate for restoration near smooth regions). On the other hand, when the modulus of the gradient of u is large, the penalizer resembles the BV seminorm $\||\nabla u|\|_{L^1(\Omega)}$, whose use, as mentioned earlier, is highly appropriate for border detection purposes. Although this model for W is quite reasonable, proving basic properties of the corresponding generalized Tikhonov–Phillips functional turns out to be quite difficult. A different way of combining these two methods was proposed by Chambolle and Lions [3]. They suggested the use of a thresholded penalizer of the form:

$$W_\beta(u) = \int_{|\nabla u|\le\beta} |\nabla u|^2 \, dx + \int_{|\nabla u|>\beta} |\nabla u| \, dx,$$

where $\beta > 0$ is a prescribed threshold parameter. Thus, in regions where borders are more likely to be present ($|\nabla u| > \beta$), penalization is made with the BV seminorm while a standard order-one Tikhonov–Phillips method is used otherwise. This model was shown to be successful in denoising of images possessing regions with homogeneous intensity separated by borders. However, in the case of images with nonuniform or highly degraded intensities, the model is extremely sensitive to the choice of the threshold β. More recently penalizers of the form:

$$W(u) = \int_\Omega |\nabla u|^{p(x)} dx, \tag{4}$$

for certain functions p with range in $[1, 2]$, were studied in [4] and [8]. It is timely to point out here that all previously mentioned results work only for the case of denoising, i.e., for the case $T = id$.

In this work, we propose the use of a model for general restoration problems, which combines, in an appropriate way, the penalizers corresponding to zero-order Tikhonov–Phillips method and the BV seminorm. Although several mathematical issues still remain open, its use in some signal restoration problems has already proved to be very promising. The purpose of this chapter is to introduce the model, show some mathematical results regarding existence and uniqueness of the corresponding regularized solutions, and present a few results of its application to signal restoration problems.

2 Main Results

In this section, we will state our main results concerning existence and uniqueness of minimizers of certain generalized Tikhonov–Phillips functionals with combined L^2-BV penalizers. We remark that all results will be presented without proofs. More details on those results including complete proofs will appear in a forthcoming chapters. In what follows, Ω will denote a bounded convex region in $\mathbb{R}^n$, $n = 1, 2$, or 3,

whose boundary $\delta\Omega$ is Lipschitz continuous and $\mathcal{M}(\Omega)$ shall denote the set of all real–valued measurable functions defined on Ω and $\theta \in \mathcal{M}(\Omega)$, a function with values in $[0, 1]$.

Definition 1 Given $\theta \in \mathcal{M}(\Omega)$ we define the functional $W_{0,\theta}(u)$ with values on the extended reals by

$$W_{0,\theta}(u) \doteq \sup_{\boldsymbol{v}\in\mathcal{V}_\theta} \int_\Omega -u \operatorname{div}(\theta \boldsymbol{v})\, \mathrm{d}x, \qquad u \in \mathcal{M}(\Omega) \tag{5}$$

where $\mathcal{V}_\theta \doteq \{\boldsymbol{v} : \Omega \to \mathbb{R}^n$ such that $\theta \boldsymbol{v} \in C_0^1(\Omega)$ and $|\boldsymbol{v}(x)| \leq 1 \,\forall\, x \in \Omega\}$.

Remark 1 For any $\theta : \Omega \to [0, 1]$, $\theta \in \mathcal{M}(\Omega)$, it follows easily that

$$W_{0,\theta}(u) \leq J_0(u), \quad \forall\, u \in \mathcal{M}(\Omega), \tag{6}$$

where $J_0(u)$ denotes the BV seminorm given by

$$J_0(u) = \sup_{\boldsymbol{v}\in\mathcal{V}} \int_\Omega -u \operatorname{div}\boldsymbol{v}\, \mathrm{d}x, \tag{7}$$

with $\mathcal{V} \doteq \{\boldsymbol{v} : \Omega \to \mathbb{R}^n$ such that $\boldsymbol{v} \in C_0^1(\Omega)$ and $|\boldsymbol{v}(x)| \leq 1 \,\forall\, x \in \Omega\}$.

Although inequality (6) is important by itself since it relates the functionals $W_{0,\theta}$ and J_0, in order to be able to use the known coercitivity properties of J_0 (see [1]), an inequality of the opposite type is desired. That is, we would like to show that under certain conditions on $\theta(\,\cdot\,)$, there exists a constant $C = C(\theta)$ such that $W_{0,\theta}(u) \geq C\, J_0(u)$ for all $u \in \mathcal{M}(\Omega)$. The following theorem provides sufficient conditions on θ assuring such an inequality.

Theorem 1 *Let $\theta : \Omega \to [0, 1]$ be such that $\frac{1}{\theta} \in L^\infty(\Omega)$ and let J_0, $W_{0,\theta}$ be the functionals defined in (7) and (5), respectively. Then $J_0(u) \leq \|\frac{1}{\theta}\|_{L^\infty(\Omega)}\, W_{0,\theta}(u)$ for all $u \in \mathcal{M}(\Omega)$.*

The following lemma is of fundamental importance in all the upcoming results.

Lemma 1 *The functional $W_{0,\theta}$ defined by (5) is weakly lower semicontinuous with respect to the L^p topology, $\forall\, p \in [1, \infty)$.*

We are now ready to present several results on existence and uniqueness of minimizers of generalized Tikhonov–Phillips functionals with penalizers involving spatially varying combinations of the L^2-norm and of the functional $W_{0,\theta}$, under different hypotheses on the function θ.

Theorem 2 *Let $\Omega \subset \mathbb{R}^n$ be a bounded open convex set with Lipschitz boundary, $\mathcal{X} = L^2(\Omega)$, $\mathcal{Y}$ a normed vector space, $T \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$, $v \in \mathcal{Y}$, α_1, α_2 positive constants, $\theta : \Omega \to [0, 1]$ a measurable function and F_θ the functional defined by*

$$F_\theta(u) \doteq \|Tu - v\|_{\mathcal{Y}}^2 + \alpha_1 \|\sqrt{1-\theta}\, u\|_{L^2(\Omega)}^2 + \alpha_2\, W_{0,\theta}(u), \qquad u \in L^2(\Omega). \tag{8}$$

If there exists $\varepsilon_2 \in \mathbb{R}$, such that $\theta(x) \leq \varepsilon_2 < 1$ for a.e. $x \in \Omega$, then the functional (8) has a unique global minimizer $u^* \in L^2(\Omega)$. If moreover there exists $\varepsilon_1 \in \mathbb{R}$ such that $0 < \varepsilon_1 \leq \theta(x)$ for a.e. $x \in \Omega$, then $u^* \in BV(\Omega)$.

Remark 2 Note that if $\theta(x) = 0\ \forall x \in \Omega$, then in (2), $W(u) = \|u\|^2_{L^2(\Omega)}$ and F_θ as defined in (8) is the classical Tikhonov–Phillips functional of order zero. On the other hand, if $\theta(x) = 1\ \forall x \in \Omega$, then $W(u) = J_0(u)$ is the BV-seminorm and F_θ has a global minimizer provided that the operator T does not anhilates constant functions on Ω, i.e., $T\chi_\Omega \neq 0$ (see [1]).

Theorem 3 *Let $\Omega \subset \mathbb{R}^n$ be a bounded open convex set with Lipschitz boundary, $\mathcal{X} = L^2(\Omega)$, $\mathcal{Y}$ a normed vector space, $T \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$, $v \in \mathcal{Y}$, α_1, α_2 positive constants and $\theta : \Omega \to [0, 1]$ such that $\frac{1}{1-\theta} \in L^1(\Omega)$ and $\frac{1}{\theta} \in L^\infty(\Omega)$. Then the functional (8) has a unique global minimizer $u^* \in BV(\Omega)$.*

Theorem 4 *Let $\Omega \subset \mathbb{R}^n$ be a bounded open convex set with Lipschitz boundary, $\mathcal{X} = L^2(\Omega)$, $\mathcal{Y}$ a normed vector space, $T \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$, $v \in \mathcal{Y}$, α_1, α_2 positive constants, $\theta : \Omega \to [0, 1]$, $\theta \in \mathcal{M}(\Omega)$, and $\Omega_0 \doteq \{x \in \Omega$ such that $\theta(x) = 0\}$. If $\frac{1}{\theta} \in L^\infty(\Omega^c_0)$ and $\frac{1}{1-\theta} \in L^1(\Omega^c_0)$, then the functional (8) has a unique global minimizer $u^* \in L^2(\Omega) \cap BV(\Omega^c_0)$.*

Theorem 5 *Let $n \leq 2$, $\Omega \subset \mathbb{R}^n$ be a bounded open convex set with Lipschitz boundary, $\mathcal{X} = L^2(\Omega)$, $\mathcal{Y}$ a Hilbert space, $T \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$, $v \in \mathcal{Y}$, α_1, α_2 positive constants. Let $\theta : \Omega \to [0, 1]$, $\theta \in \mathcal{M}(\Omega)$ and $\Omega_1 \doteq \{x \in \Omega$ such that $\theta(x) = 1\}$. If $\frac{1}{\theta} \in L^\infty(\Omega^c_1)$, $\frac{1}{1-\theta} \in L^1(\Omega^c_1)$ and $T\chi_\Omega \neq 0$, then the functional (8) has a global minimizer $u^* \in L^2(\Omega) \cap BV(\Omega^c_1)$.*

3 Applications to Signal Restoration

The purpose of this section is to present some applications of the simultaneous use of penalizers of L^2 and of BV type to signal restoration problem.

A basic mathematical model for signal blurring is given by convolution, as a Fredholm integral equation of first kind:

$$v(t) = \int_0^1 k(t, s)u(t)\mathrm{d}s, \tag{9}$$

where $k(t, s) = \frac{1}{\sqrt{2\pi}\sigma_b} \exp\left(-\frac{(t-s)^2}{2\sigma_b^2}\right)$ is a Gaussian kernel, $\sigma_b > 0$, u is the original signal and v is the blurred signal. For the numerical examples that follow, Eq. (9) was discretized in the usual way (using collocation and quadrature), resulting in a discrete model of the form:

$$Af = g, \tag{10}$$

where A is a $(n + 1) \times (n + 1)$ matrix, $f, g \in \mathbb{R}^{n+1}$ ($f_j = u(t_j)$, $g_j = v(t_j)$, $t_j = \frac{j}{n}$, $0 \leq j \leq n$). We took $n = 130$ and $\sigma_b = 0.05$. The data g was contaminated with a 1 % zero-mean Gaussian additive noise (i.e., standard deviation equal to 1 % of the range of g). Figure 1 shows the original signal (unknown in real life problems) and the blurred noisy signal which constitutes the data of the inverse problem.

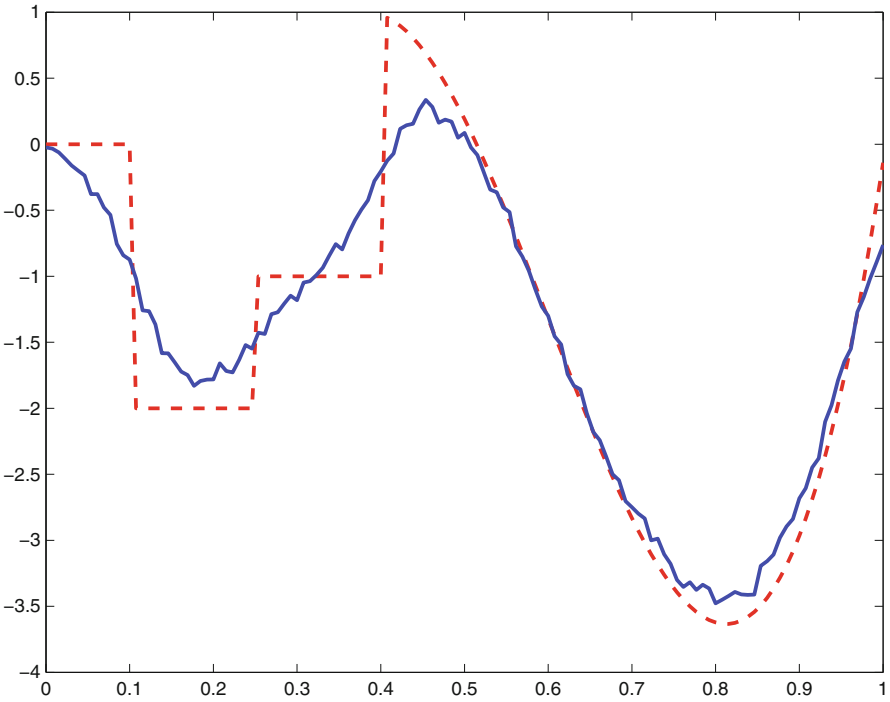


Fig. 1 Original signal (*red dotted line*) and blurred noisy signal (*blue line*)

Figure 2 shows the regularized solutions obtained with the classical Tikhonov–Phillips method of order zero and with penalizer associated to the BV seminorm J_0. As expected, the regularized solution obtained with the J_0 penalizer is significantly better than the one obtained with the classical Tikhonov–Phillips method near jumps and in regions where the exact solution is piecewise constant. The opposite happens where the exact solution is smooth.

Figure 3 shows the regularized solution obtained with the combined L^2-BV method (see (8)). In this case, the function $\theta(t)$ was chosen to be $\theta(t) \doteq 1$ for $t \in (0, 0.4]$ and $\theta(t) \doteq 0$ for $t \in (0.4, 1)$. Although this choice of $\theta(t)$ is clearly based upon "a priori" information about the regularity of exact solution, other choices of θ can be made by using only data-based information. For example, the function $\theta(t)$ can be computed by normalizing in $[0, 1]$ the modulus of the gradient of the regularized solution obtained with a pure zero-order Tikhonov–Phillips method (see Fig. 4). For this function θ, the regularized solution obtained with the combined L^2-BV method is shown in Fig. 5. In all cases, reflexive boundary conditions were used [7] and the regularization parameters were calculated using the Morozov's discrepancy principle with $\tau = 1.1$ [5].

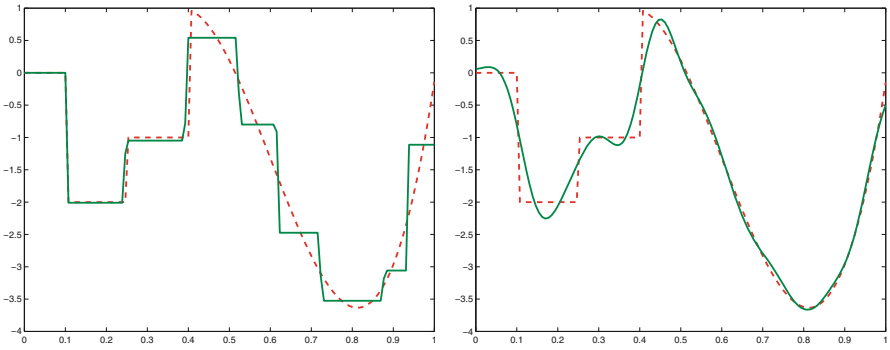


Fig. 2 Original signal (*red dotted line*) and regularized solution (*green line*) obtained with Tikhonov–Phillips (*right*) and bounded variation seminorm (*left*)

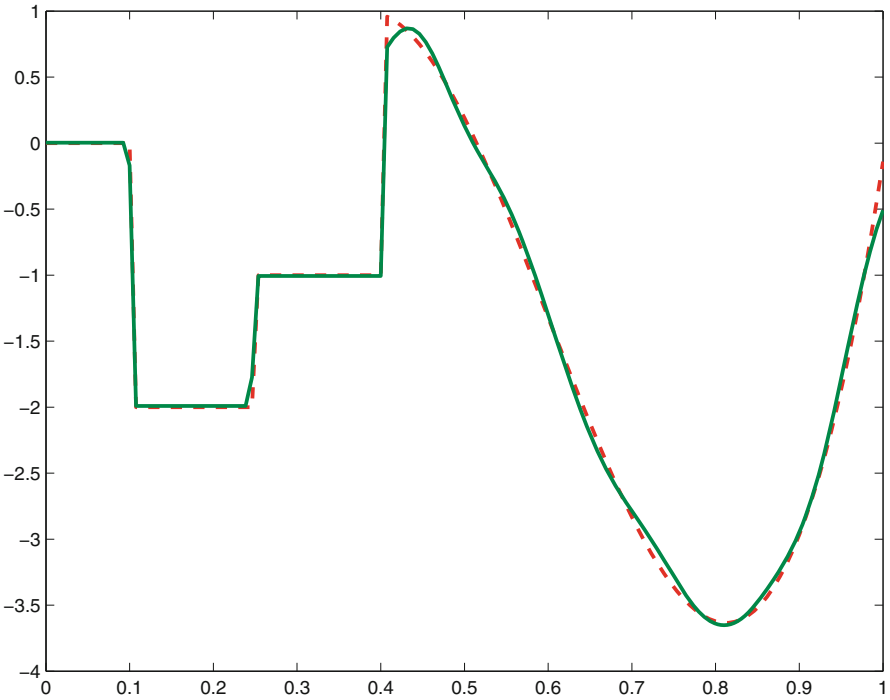


Fig. 3 Original signal (*red dotted line*) and regularized solution (*green line*) obtained with the combined method L^2-bounded variation (BV) and binary function θ

Remark 3 The estimation of the optimal parameters α_1 and α_2 was performed as follows: first optimal regularization parameters δ_1 and δ_2 were estimated independently of one another by using Morozov's discrepancy principle for Tikhonov–Phillips and BV methods, respectively. Then δ_1 and δ_2 were used as weights on each penalizing

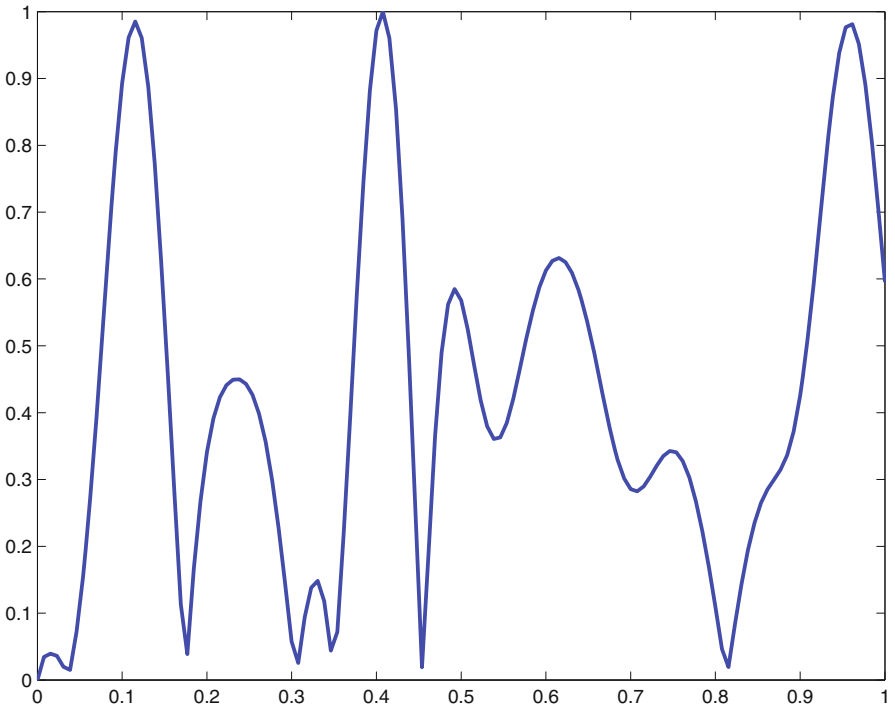


Fig. 4 Function θ computed by normalizing in $[0, 1]$ the modulus of the gradient of the regularized solution with a pure zero-order Tikhonov–Phillips method

term of the mixed L^2-BV method and finally, by using once again the discrepancy principle, a third optimal parameter α multiplying both terms was determined, so that the actual penalizer W considered was:

$W(u) = \alpha \left(\delta_1 \| \sqrt{1-\theta}\, u \|^2_{L^2(\Omega)} + \delta_2\, W_{0,\theta}(u) \right)$. Hence, $\alpha_1 = \alpha\delta_1$ and $\alpha_2 = \alpha\delta_2$.

For the case of binary function θ, the improvement of the combined L^2-BV method with respect to the pure simple cases, Tikhonov–Phillips method of zero order and regularization with penalizer associated to the BV seminorm, is notorious. In that case, however, the function θ was constructed based on a priori information that may not be available in a concrete problem. Nevertheless, the regularized solution obtained with the data-based function θ shown in Fig. 4 is also significantly better than the those obtained with the single-based penalizers. This fact is clearly and objectively reflected by the improved signal-to-noise ratio (ISNR) defined as

$$\mathrm{ISNR} = 10 \log_{10} \left(\frac{\|g - f\|^2}{\|f_\alpha - f\|^2} \right),$$

(where f_α is the restored signal obtained with regularization parameter α). For all the previously shown examples, the ISNR was computed in order to have a parameter

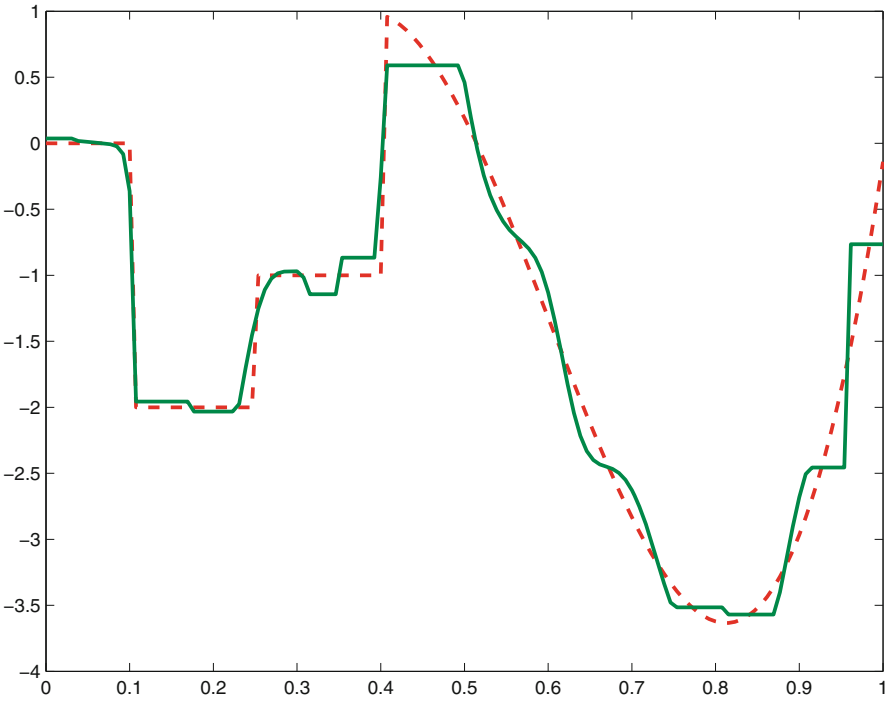


Fig. 5 Original signal (*red dotted line*) and regularized solution (*green line*) obtained with the combined method L^2-bounded variation (BV) and function θ showed in Fig. 4

Table 1 Improved signal-to-noise ratio (ISNR)

Regularization method	ISNR
Tikhonov–Phillips of order zero (T–P)	3.4866
Bounded variation seminorm	0.6459
Combined method L^2-BV with binary θ	14.725
Combined method L^2-BV with θ obtained using regularized T	4.7686

for objectively measuring and comparing the quality of all regularized solutions (see Table 1).

Remark 4 We find it appropriate to make a final remark regarding the performance of the mixed methods for different noise levels. Table 2 shows the ISNR values obtained with the four methods for different noise levels. For low noise levels (less than 1.5 %), both mixed methods consistently performed better than both single ones. For higher noise levels, the mixed method with binary weighting function θ performed better than Tikhonov while the method with θ estimated by convolution did not. Thus, in the presence of high noise levels, this fact points to the need of either having suitable information for estimating θ and/or to the necessity of developing appropriate ways for its adequate estimation.

Table 2 ISNR values for different noise levels

Method	1 % noise	1.5 % noise	4 % noise	5 % noise
Tikhonov–Phillips	3.4866	2.9614	2.5403	2.3045
Bounded variation	0.6459	−0.6568	0.1767	0.5103
Mixed L^2-BV with binary θ	14.725	6.1543	4.1817	3.1271
Mixed L^2-BV with Tikhonov-based θ	4.7686	2.7934	2.2086	1.6797

4 Conclusions

In this chapter, we introduced a generalized Tikhonov–Phillips regularization method in which the penalizer is given by a combination of the L^2 norm and the BV seminorm. For particular cases, existence and uniqueness results were shown. Finally, applications of the model to signal restoration problem were shown.

Although these preliminary results are clearly quite promising, much further research is needed. In particular, in spite of interesting numerical results, no rigorous mathematical proofs are yet known on the existence and uniqueness of minimizers of functional (8) for the case $\theta(t)$ binary (i.e., with $\theta(t)$ taking only the values 0 and 1). Further, the choice of the function $\theta(t)$ in a somewhat optimal way is also a subject which deserves much further attention. Research in all these directions is currently under way.

Acknowledgement This work was supported in part by Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, through PIP 11220130100216-CO, by Universidad Nacional del Litoral, U.N.L., through projects CAI+D 2009-PI-62-315, CAI+D PJov 2011 N° 50020110100055, CAI+D PI N° 50120110100294, and by the Air Force Office of Scientific Research, AFOSR, through Grant FA9550-14-1-0130.

References

1. Acar, R., Vogel, C.R.: Analysis of bounded variation penalty methods for ill-posed problems. Inverse Probl. **10**, 1217–1229 (1994)
2. Blomgren, P., Chan, T.F., Mulet, P., Wong, C.: Total variation image restoration: Numerical methods and extensions. Proceedings of the IEEE International Conference on Image Processing, vol. III, pp. 384–387 (1997)
3. Chambolle, A., Lions, J.L.: Image recovery via total variation minimization and related problems. Numer. Math. **76**, 167–188 (1997)
4. Chen, Y., Levine, S., Rao, M.: Variable exponent, linear growth functionals in image restoration. SIAM J. Appl. Math. **66**, 1383–1406 (2006)
5. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of inverse problems. Mathematics and its applications, vol. 375. Kluwer Academic Publishers Group, Dordrecht (1996)
6. Hadamard, J.: Sur les problèmes aux dérivées partielles et leur signification physique. Princeton Univ. Bull. **13**, 49–52 (1902)
7. Hansen, P. C.: Discrete inverse problems: Insight and algorithms. Fundamentals of algorithms, vol. FA07. Society for Industrial and Applied Mathematics, Philadelphia (2010)

8. Li, F., Li, Z., Pi, L.: Variable exponent functionals in image restoration. Appl. Math. Comput. **216**, 870–882 (2010)
9. Mazzieri, G.L., Spies, R.D., Temperini, K.G.: Existence, uniqueness and stability of minimizers of generalized Tikhonov-Phillips functionals. J. Math. Anal. Appl. **396**, 396–411 (2012)
10. Rudin, L.I., Osher, S., Fatemi E.: Nonlinear total variation based noise removal algorithms. Proceedings of the 11th Annual International Conference of the Center for Nonlinear Studies. Physica D, vol. 60, pp. 259–268 (1992)

Stability Analysis of a Finite Difference Scheme for a Nonlinear Time Fractional Convection Diffusion Equation

Carlos D. Acosta, Pedro A. Amador and Carlos E. Mejía

Abstract The nonlinear time fractional convection diffusion equation (TFCDE) is obtained from a standard nonlinear convection diffusion equation by replacing the first-order time derivative with a fractional derivative (in Caputo sense) of order $\alpha \in (0, 1)$. Developing numerical methods for solving fractional partial differential equations is of increasing interest in many areas of science and engineering. In this chapter, an explicit conservative finite difference scheme for TFCDE is introduced. We find its Courant–Friedrichs–Lewy (CFL) condition and prove encouraging results regarding stability, namely, monotonicity, the total variation diminishing (TVD) property and several bounds. Illustrative numerical examples are included in order to evaluate potential uses of the new method.

Keywords Caputo fractional derivative · Finite difference scheme · Stability · CFL · TVD

1 Introduction

The area of *fractional calculus* is as old as classical calculus, that is, the end of the seventeenth century. In the introduction of [13], there is an annotated chronological bibliography on fractional calculus prepared by Professor Bertram Ross of the University of New Haven. This chapter contains seven pages for the twentieth century (up to 1975), five pages for the nineteenth century, and only two entries for the eighteenth and seventeenth centuries. According to more recent works like [9], [14], and

C. D. Acosta (✉) · P. A. Amador
Departamento de Matemáticas y Estadística, sede Manizales,
Universidad Nacional de Colombia, Manizales, Colombia
e-mail: cdacostam@unal.edu.co

P. A. Amador
e-mail: paamadorr@unal.edu.co

C. E. Mejía
Escuela de Matemáticas, sede Medellín,
Universidad Nacional de Colombia, Medellín, Colombia
e-mail: cemejia@unal.edu.co

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_10

[15], it is quite possible that the last quarter of the twentieth century amounts for as much research on the subject as all the previous works. The reasons are: Better knowledge of mathematics and the generalized use of computers.

We are interested in numerical strategies for Cauchy problems of the form:

$$u_t^\alpha + cu_x = A(u)_{xx}, \qquad 0 < \alpha < 1, \quad (x,t) \in \Pi_T := \mathbb{R} \times (0,T), \quad T > 0 \quad (1)$$

with initial condition given by

$$u(x,0) = u_0(x), \qquad x \in \mathbb{R}.$$

Here c is a positive constant, the integrated diffusion coefficient $A(u)$ is defined by

$$A(u) = \int_0^u a(s)\mathrm{d}s, \quad a(u) \geq 0, \quad a \in L^\infty(\mathbb{R}) \cap L^1(\mathbb{R}) \qquad (2)$$

and u_t^α denotes Caputo's fractional derivative of order α defined by

$$u_t^\alpha(x,t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t \frac{\partial u(x,\xi)}{\partial \xi} \frac{1}{(t-\xi)^\alpha} \mathrm{d}\xi. \qquad (3)$$

The diffusion function $a(s)$ is allowed to vanish on intervals of positive length and thus, in principle, (1) might be a strongly degenerate parabolic equation.

In this chapter, we introduce a new finite difference scheme for Eq. (1), establish some of its main features and conclude with some illustrative numerical examples. Finite difference methods are promising for fractional diffusion equations [3, 7, 8, 12, 16, 17] and for strongly degenerate parabolic equations with or without fractional derivatives [1, 2, 4, 10].

Convergence studies for the numerical solution of problems like (1) are widely unknown and are the subject of current research. Promising perspectives are presented in [4] in which the authors deal with a a fractional nonlinear diffusion term, [5] for a degenerate fractional diffusion term and [6] in which the authors study a scalar conservation law with u_t replaced by the time derivative of a Volterra-type convolution in time of the solution and a kernel k. However, there are still a lot of open questions.

The rest of this chapter consists on three sections dedicated to the numerical method, some stability considerations and a collection of numerical examples and final remarks, respectively.

2 The Numerical Method

We begin our discussion of a finite difference scheme for Eq. (1) by defining a grid of points in the (x,t) strip. Let Δx be a positive real number, N be a positive integer and let us define $\Delta t = \frac{T}{N}$. The grid will be the points $(x_j, t^n) = (j\Delta x, n\Delta t)$ for all $j \in \mathbb{Z}$ and $n = 0, 1, \ldots, N$.

Following [16], the Caputo fractional derivative at time t^{n+1} can be approximated by

$$u_t^{\alpha}\left(x,t^{n+1}\right)=\frac{(\Delta t)^{-\alpha}}{\Gamma(2-\alpha)}\sum_{k=0}^{n}b_k\left[u\left(x,t^{n-k+1}\right)-u\left(x,t^{n-k}\right)\right]+O\left((\Delta t)^{2-\alpha}\right),$$

for $n=0,1,...,N-1$ and weights $b_k=(k+1)^{1-\alpha}-k^{1-\alpha}$ for $k=0,1,...,n$.

The partial derivatives with respect to x are approximated in a straightforward way by

$$\frac{\partial u(x_j,t)}{\partial x}=\frac{u(x_j,t)-(x_{j-1},t)}{\Delta x}+O(\Delta x) \tag{4}$$

and

$$\frac{\partial^2 A(u(x_j,t))}{\partial x^2}=\frac{A(u(x_{j+1},t))-2A(u(x_j,t))+A(u(x_{j-1},t))}{(\Delta x)^2}+O\left((\Delta x)^2\right) \tag{5}$$

Let us denote by v_j^n the numerical approximation of $u(x_j,t^n)$. The numerical method for the solution of (1) is obtained from the previous approximations and is given by the explicit finite difference scheme:

$$\frac{(\Delta t)^{-\alpha}}{\Gamma(2-\alpha)}\sum_{k=0}^{n}b_k\left[v_j^{n-k+1}-v_j^{n-k}\right]+c\frac{v_j^n-v_{j-1}^n}{\Delta x}$$
$$=\frac{A\left(v_{j+1}^n\right)-2A\left(v_j^n\right)+A\left(v_{j-1}^n\right)}{(\Delta x)^2}.$$

Let $\lambda=\Gamma(2-\alpha)(\Delta t)^{\alpha}/\Delta x$, $\mu=\lambda/\Delta x$ and $A_j^n=A\left(v_j^n\right)$. If $n=0$, the numerical scheme can be written as:

$$v_j^1=v_j^0-c\lambda\left(v_j^0-v_{j-1}^0\right)+\mu\left(A_{j+1}^0-2A_j^0+A_{j-1}^0\right). \tag{6}$$

Likewise, if $n\geq 1$, the numerical scheme becomes

$$v_j^{n+1}=v_j^n-c\lambda\left(v_j^n-v_{j-1}^n\right)+\mu\left(A_{j+1}^n-2A_j^n+A_{j-1}^n\right)-\sum_{k=1}^{n}b_k\left[v_j^{n-k+1}-v_j^{n-k}\right]. \tag{7}$$

An alternative way to write scheme (7) is

$$v_j^{n+1}=v_j^n-c\lambda\left(v_j^n-v_{j-1}^n\right)+\mu\left(A_{j+1}^n-2A_j^n+A_{j-1}^n\right) \tag{8}$$
$$-b_1v_j^n+\sum_{k=1}^{n-1}d_kv_j^{n-k}+b_nv_j^0,$$

where $d_k=b_k-b_{k+1}$ for $k=1,2,\ldots,n-1$.

Sometimes it is appropriate to consider the method in sequence form. Let $v^n = (v_j^n)_{j\in\mathbb{Z}}$. Method (6) and (7) are represented by an expression of the form:

$$v^{n+1} = \mathcal{H}\left(v^n, v^{n-1}, \cdots, v^0; j\right) \tag{9}$$

where the right-hand side in (9) corresponds to the right-hand side in (6) or (7), depending on the value of n.

The first feature of scheme (6) and (7) is that it allows a conservative form, which guarantees that the numerical method does not converge to nonsolutions.

Lemma 1 *(6) and (7) is conservative, that is, it admits a conservation form. More precisely,*

$$v_j^{n+1} = v_j^n - \lambda\left(\psi_j^n - \psi_{j-1}^n\right), \tag{10}$$

where

$$\psi_j^0 = cv_j^0 - \frac{1}{\Delta x}\left(A_{j+1}^0 - A_j^0\right), \quad \textit{for } n = 0$$

$$\psi_j^n = cv_j^n - \frac{1}{\Delta x}\left(A_{j+1}^n - A_j^n\right) - \sum_{k=1}^{n} b_k \psi_j^{n-k}, \quad \textit{for } n \geq 1$$

Proof The case $n = 0$ follows from (6). Suppose it is possible to achieve the conservation form (10) for $k = 0, 1, \cdots, n-1$, that is

$$v_j^{k+1} = v_j^k - \lambda\left(\psi_j^k - \psi_{j-1}^k\right).$$

For $k = n$,

$$\begin{aligned} v_j^{n+1} &= v_j^n - c\lambda\left(v_j^n - v_{j-1}^n\right) + \mu\left(A_{j+1}^n - 2A_j^n + A_{j-1}^n\right) \\ &\quad - \sum_{k=1}^{n} b_k\left(v_j^{n-k+1} - v_j^{n-k}\right) \\ &= v_j^n - \lambda\left\{c\left(v_j^n - v_{j-1}^n\right) - \frac{1}{\Delta x}\left[\left(A_{j+1}^n - A_j^n\right) - \left(A_j^n - A_{j-1}^n\right)\right]\right\} \\ &\quad + \lambda\sum_{k=1}^{n} b_k\left(\psi_j^{n-k} - \psi_{j-1}^{n-k}\right). \end{aligned}$$

We end this section by clarifying that convergence issues are not addressed here although they are important. Since nonlinear equations may have several weak solutions, an entropy condition is usually required to identify the *physically correct* solution. These ideas, along with the notion of nonlinear stability, are treated by many authors. For an initial boundary value problem of a strongly degenerate parabolic equation in which the time derivative is not fractional, we recommend [2].

Next section deals with conditional stability and other properties of scheme (6) and (7).

3 Stability Analysis

Explicit schemes require certain restrictions on the discretization parameters in order for the method to be useful. Among them, the inequality known as Courant–Friedrichs–Lewy (CFL) condition is of paramount importance. We begin by introducing the CFL condition for scheme (6) and (7), which is

$$c\lambda + 2\mu \|a\|_\infty \leq 2 - 2^{1-\alpha}. \tag{11}$$

Provided the CFL condition is satisfied, two important properties of the method are derived.

3.1 Monotonicity Property

Let u_j^n and v_j^n be two discrete functions to which method (9) can be applied. The numerical method (9) is called a *monotone* method if

$$u_j^0 \leq v_j^0 \quad \text{for all } j \Longrightarrow u_j^n \leq v_j^n \quad \text{for all } j \text{ and all } n$$

Theorem 1 *If the CFL condition (11) holds, then method (9) is monotone.*

Proof Suposse $u_j^0 \leq v_j^0$ for all $j \in \mathbb{Z}$. For all n, we denote $A_j^n = A\left(v_j^n\right)$ and $\bar{A}_j^n = A(u_j^n)$. For $n = 1$, monotonicity is proved as follows:

$$\begin{aligned} v_j^1 - u_j^1 &= \left(v_j^0 - u_j^0\right) - c\lambda\left(\left(v_j^0 - u_j^0\right) - \left(v_{j-1}^0 - u_{j-1}^0\right)\right) \\ &\quad + \mu\left(\left(A_{j+1}^0 - \bar{A}_{j+1}^0\right) - 2\left(A_j^0 - \bar{A}_j^0\right) + \left(A_{j-1}^0 - \bar{A}_{j-1}^0\right)\right) \\ &= \int_{u_j^0}^{v_j^0} (1 - c\lambda - 2\mu a(u))\,\mathrm{d}u + \mu \int_{u_{j+1}^0}^{v_{j+1}^0} a(u)\mathrm{d}u + \mu \int_{u_{j-1}^0}^{v_{j-1}^0} a(u)\mathrm{d}u \\ &\geq 0 \end{aligned}$$

The CFL condition (11) allows nonnegativity of the first of the three integrals. Now, suppose $u_j^k \leq v_j^k$ for $k = 0, 1, \cdots, n$ and all $j \in \mathbb{Z}$. Thus,

$$\begin{aligned} v_j^{n+1} - u_j^{n+1} &= \int_{u_j^n}^{v_j^n} (1 - c\lambda - 2\mu a(u))\,\mathrm{d}u + \mu \int_{u_{j+1}^n}^{v_{j+1}^n} a(u)\mathrm{d}u + \mu \int_{u_{j-1}^0}^{v_{j-1}^n} a(u)\mathrm{d}u \\ &\quad - b_1 \int_{u_j^n}^{v_j^n} \mathrm{d}u + \sum_{k=1}^{n-1} d_k \left(v_j^{n-k} - u_j^{n-k}\right) + b_n \left(v_j^0 - u_j^0\right) \end{aligned}$$

$$
\begin{aligned}
&= \int_{u_j^n}^{v_j^n} (1 - b_1 - c\lambda - 2\mu a(u))\, \mathrm{d}u + \mu \int_{u_{j+1}^n}^{v_{j+1}^n} a(u)\mathrm{d}u + \mu \int_{u_{j-1}^0}^{v_{j-1}^n} a(u)\mathrm{d}u \\
&\quad + \sum_{k=1}^{n-1} d_k \left(v_j^{n-k} - u_j^{n-k} \right) + b_n \left(v_j^0 - u_j^0 \right) \\
&\geq 0
\end{aligned}
$$

where we have taken into consideration that $1 - b_1 = 2 - 2^{1-\alpha}$ and the CFL condition.

3.2 Stability Bounds

The next theorem establishes two stability bounds in the ∞-norm and the 1-norm, respectively, and it includes a total variation diminishing property of importance in case a convergence analysis is sought.

Theorem 2 *If the CFL condition (11) is satisfied, then the following inequalities hold:*

$$
\begin{aligned}
\left\| v^n \right\|_\infty &\leq \left\| v^0 \right\|_\infty, n = 1, 2, \cdots, N \\
\left\| v^n \right\|_1 &\leq \left\| v^0 \right\|_1, n = 1, 2, \cdots, N \\
\sum_j \left| v_{j+1}^{n+1} - v_j^{n+1} \right| &\leq \sum_j \left| v_{j+1}^n - v_j^n \right|, n = 1, 2, \cdots, N
\end{aligned}
$$

Proof First observe that

$$
\begin{aligned}
v_j^1 &= v_j^0 - c\lambda \left(v_j^0 - v_{j-1}^0 \right) + \mu \left(A_{j+1}^0 - 2A_j^0 + A_{j-1}^0 \right) \\
&= (1 - c\lambda)\, v_j^0 + c\lambda v_{j-1}^0 + \mu \left(\left(A_{j+1}^0 - A_j^0 \right) - \left(A_j^0 - A_{j-1}^0 \right) \right) \\
&= (1 - c\lambda)\, v_j^0 + c\lambda v_{j-1}^0 + \mu \left(a(\zeta_{j+1/2}^0) \left(v_{j+1}^0 - v_j^0 \right) - a(\zeta_{j-1/2}^0) \left(v_j^0 - v_{j-1}^0 \right) \right) \\
&= \left(1 - c\lambda - \mu a(\zeta_{j+1/2}^0) - \mu a(\zeta_{j-1/2}^0) \right) v_j^0 \\
&\quad + c\lambda v_{j-1}^0 + \mu a(\zeta_{j+1/2}^0) v_{j+1}^0 + \mu a(\zeta_{j-1/2}^0) v_{j-1}^0.
\end{aligned}
$$

for some values $\zeta_{j\pm1/2}^0$ between $v_{j\pm1}^0$ and v_j^0, respectively. Then,

$$
\begin{aligned}
\left| v_j^1 \right| &\leq \left(1 - c\lambda - \mu a(\zeta_{j+1/2}^0) - \mu a(\zeta_{j-1/2}^0) \right) \left| v_j^0 \right| + c\lambda \left| v_{j-1}^0 \right| \\
&\quad + \mu a(\zeta_{j+1/2}^0) \left| v_{j+1}^0 \right| + \mu a(\zeta_{j-1/2}^0) \left| v_{j-1}^0 \right| \leq \left\| v^0 \right\|_\infty.
\end{aligned}
$$

Also,

$$
\sum_j \left| v_j^1 \right| \leq \sum_j \left(1 - c\lambda - \mu a(\zeta_{j+1/2}^0) - \mu a(\zeta_{j-1/2}^0) \right) \left| v_j^0 \right| + \sum_j c\lambda \left| v_{j-1}^0 \right|
$$

$$
\begin{aligned}
&+\mu\sum_j a(\zeta_{j+1/2}^0)\left|v_{j+1}^0\right|+\mu\sum_j a(\zeta_{j-1/2}^0)\left|v_{j-1}^0\right| \\
&\leq \sum_j \left|v_j^0\right| - c\lambda\sum_j\left(\left|v_j^0\right|-\left|v_{j-1}^0\right|\right) \\
&\quad -\mu\sum_j\left(a(\zeta_{j+1/2}^0)\left|v_j^0\right|-a(\zeta_{j-1/2}^0)\left|v_{j-1}^0\right|\right) \\
&\quad -\mu\sum_j\left(a(\zeta_{j-1/2}^0)\left|v_j^0\right|-a(\zeta_{j+1/2}^0)\left|v_{j+1}^0\right|\right) \\
&\leq \sum_j \left|v_j^0\right|.
\end{aligned}
$$

Similarly, we get:

$$
\sum_j \left|v_j^1 - v_{j-1}^1\right| \leq \sum_j \left|v_j^0 - v_{j-1}^0\right|.
$$

To conclude the proof, we proceed by induction. Suppose the following inequalities are satisfied:

$$
\begin{aligned}
\left\|v^k\right\|_\infty &\leq \left\|v^0\right\|_\infty,\ k=1,2,\ldots,n-1<N \\
\left\|v^k\right\|_1 &\leq \left\|v^0\right\|_1,\ k=1,2,\ldots,n-1<N \\
\sum_j \left|v_{j+1}^k - v_j^k\right| &\leq \sum_j \left|v_{j+1}^{k-1} - v_j^{k-1}\right|,\ k=1,2,\ldots,n-1<N
\end{aligned}
$$

Thus, for $k=n$, we have:

$$
\begin{aligned}
v_j^{n+1} &= v_j^n - c\lambda\left(v_j^n - v_{j-1}^n\right)+\mu\left(A_{j+1}^n - 2A_j^n + A_{j-1}^n\right) \\
&\quad - b_1 v_j^n + \sum_{k=1}^{n-1} d_k v_j^{n-k} + b_n v_j^0 \\
&= \left(1 - b_1 - c\lambda - \mu a(\zeta_{j+1/2}^n) - \mu a(\zeta_{j-1/2}^n)\right) v_j^n \\
&\quad + c\lambda v_{j-1}^n + \mu a(\zeta_{j+1/2}^n) v_{j+1}^n + \mu a(\zeta_{j-1/2}^n) v_{j-1}^n \\
&\quad + \sum_{k=1}^{n-1} d_k v_j^{n-k} + b_n v_j^0.
\end{aligned}
$$

By the CFL condition (11), we obtain:

$$
\begin{aligned}
\left|v_j^{n+1}\right| &\leq \left(1 - b_1 - c\lambda - \mu a(\zeta_{j+1/2}^n) - \mu a(\zeta_{j-1/2}^n)\right)\left|v_j^n\right| \\
&\quad + c\lambda\left|v_{j-1}^n\right| + \mu a(\zeta_{j+1/2}^n)\left|v_{j+1}^n\right| + \mu a(\zeta_{j-1/2}^n)\left|v_{j-1}^n\right|
\end{aligned}
$$

Table 1 Numerical results for Example 1

	$\alpha = 1/2$		$\alpha = 2/3$		$\alpha = 3/4$	
Δx	L_∞−err	Order	L_∞−err	Order	L_∞−err	Order
1/32	0.099289	–	0.24896	–	0.27227	–
1/64	0.0062903	3.9804	0.029929	3.0563	0.046724	2.5428
1/128	0.0003967	3.987	0.0038678	2.952	0.0076661	2.6076

$$+\sum_{k=1}^{n-1} d_k \left|v_j^{n-k}\right| + b_n \left|v_j^0\right|$$
$$\leq \left(1 - b_1 + \sum_{k=1}^{n-1} d_k + b_n\right) \left\|v^0\right\|_\infty = \left\|v^0\right\|_\infty .$$

since $\sum_{k=1}^{n-1} d_k = b_n - b_1$. Similarly, we obtain the other inequalities.

4 Numerical Examples and Final Remarks

4.1 Numerical Experiments

Example 1 Fractional linear diffusion.
This experiment is a linear time fractional diffusion equation with constant diffusion $a(u) = \bar{a} = 0.001$ for all u. The right-hand side term $f(x,t)$ is chosen in such a way that the equation has a unique polynomial solution (Fig. 1). The problem is the following:

$$u_t^\alpha = 0.001 u_{xx} + f(x,t),\ 0 < \alpha < 1,\ x \in [0,1],\ 0 < t \leq 1,$$

The exact solution is given by

$$u(x,t) = 10x^2(1-x)(t+1)^2$$

For this problem, the CFL condition (11) becomes

$$2\Gamma(2-\alpha)\bar{a}\frac{(\Delta t)^\alpha}{(\Delta x)^2} \leq 2 - 2^{1-\alpha}$$

and indicates that Δt behaves like $O\left((\Delta x)^{\frac{2}{\alpha}}\right)$

Table 1 shows results for three different values of α and suggests that the order of accuracy is about $\frac{2}{\alpha}$ for Δx as the main discretization parameter. This is consistent with the theory, because monotone numerical methods are at most first-order accurate (see [11, Theorem 15.6]).

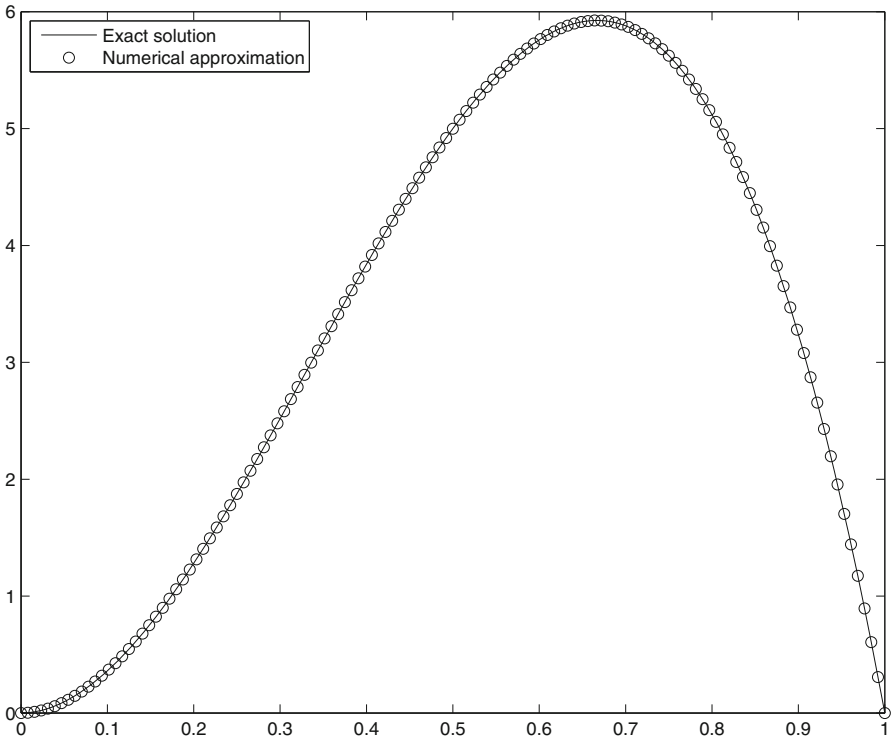


Fig. 1 Comparison of the exact solution and numerical solution for Example 1 with $\alpha = 2/3$ and $\Delta x = 1/128$

Example 2 This is a nonlinear time fractional convection diffusion equation (TFCDE) and, as before, the right-hand side function $f(x,t)$ is chosen so that the equation has a unique polynomial solution (Fig. 2):

$$u_t^{\alpha} + cu_x = A(u)_{xx} + f(x,t), \quad 0 < \alpha < 1,\ x \in [0,2],\ 0 < t \leq 1$$

$$c = 1,\ A(u) = 4\varepsilon u^2 \left(\frac{1}{2} - \frac{u}{3}\right),\ \varepsilon = 0.001$$

The exact solution is given by

$$u(x,t) = t^2 x(2-x)$$

Table 2 summarizes our results for three different values of α and several discretization parameters. It suggests that the order of accuracy is about 1 for Δx as the main discretization parameter.

Example 3 Once again, this is a nonlinear TFCDE and, as before, the right-hand side function $f(x,t)$ is chosen so that the equation has a unique closed form solution (Fig. 3):

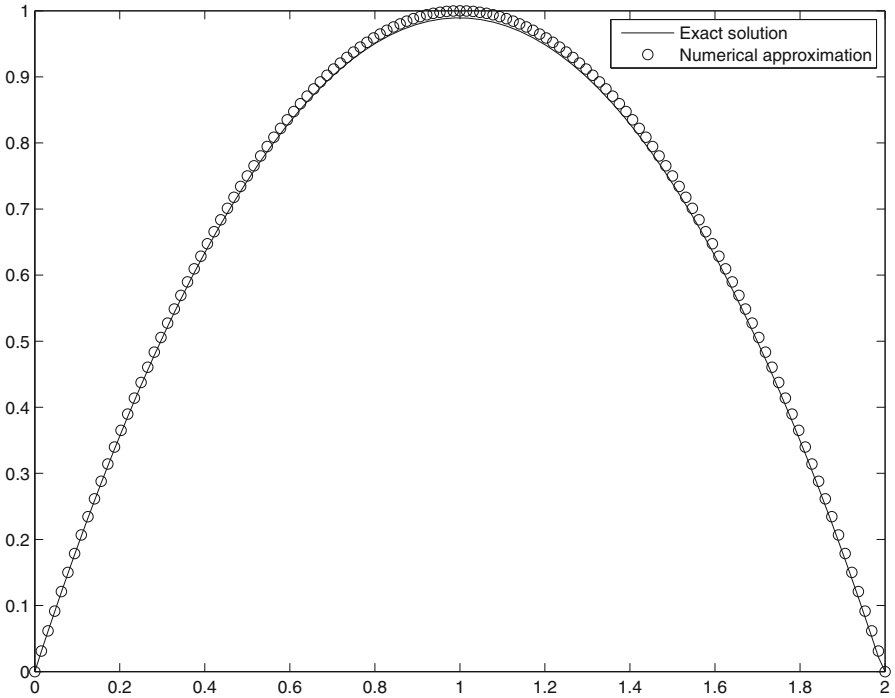


Fig. 2 Comparison of the exact solution and numerical solution for Example 2 with $\alpha = 2/3$ and $\Delta x = 1/128$

Table 2 Numerical results for Example 2

	$\alpha = 1/2$		$\alpha = 2/3$		$\alpha = 3/4$	
Δx	L_∞−err	Order	L_∞−err	Order	L_∞−err	Order
1/32	0.025104	–	0.029653	–	0.034915	–
1/64	0.012526	1.003	0.013879	1.0953	0.016232	1.105
1/128	0.0062491	1.0032	0.0065186	1.0903	0.0074625	1.1211

$$u_t^\alpha + cu_x = A(u)_{xx} + f(x,t), \quad 0 < \alpha < 1,\ x \in [0,1],\ 0 < t \leq 1,$$

$$c = 1,\ A(u) = \varepsilon \frac{u^{n+1}}{n+1},\ \varepsilon = 0.001,\ n = 2$$

The exact solution is

$$u(x,t) = t^2 \sin(2\pi x)$$

The numerical results are presented in Table 3 which includes experiments for three different values of α and several discretization parameters. As in the previous example, the table suggests that the order of accuracy is about 1 for Δx as the main discretization parameter.

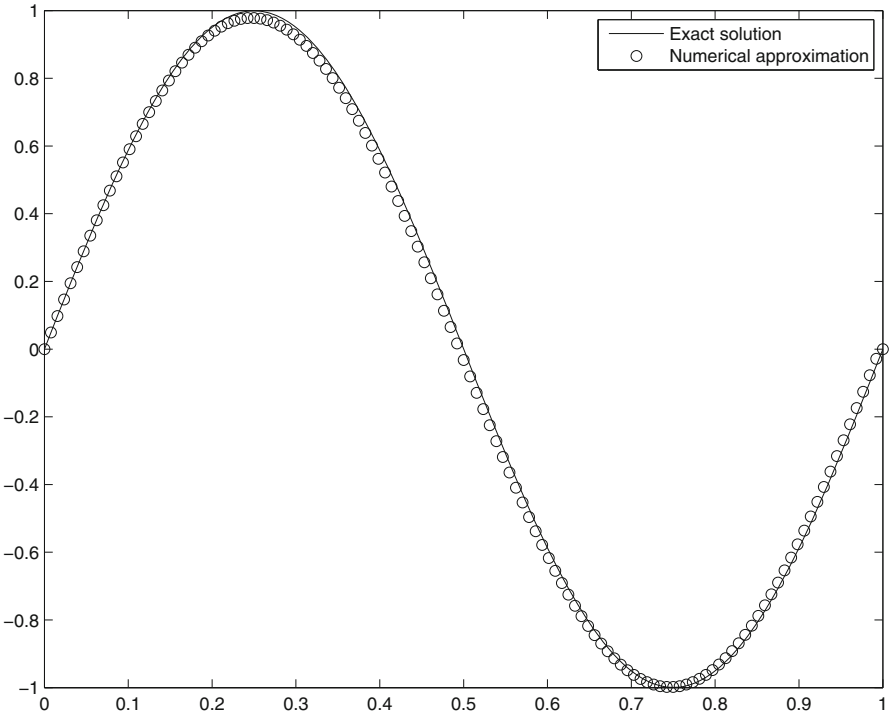


Fig. 3 Comparison of the exact solution and numerical solution for Example 3 with $\alpha = 2/3$ and $\Delta x = 1/128$

Table 3 Numerical results for Example 3

	$\alpha = 1/2$		$\alpha = 2/3$		$\alpha = 3/4$	
Δx	$L_\infty-$err	Order	$L_\infty-$err	Order	$L_\infty-$err	Order
1/32	0.22037	–	0.21317	–	0.2107	–
1/64	0.11147	0.98327	0.10815	0.97897	0.10725	0.97421
1/128	0.056076	0.9912	0.054304	0.9939	0.053776	0.99594

4.2 Concluding Remarks

A new method for the numerical solution of (1) has been presented. It is an explicit conservative finite difference method which under a reasonable CFL condition satisfies standard stability estimates. If some day an entropy solution of (1) is defined, method (8) is an excellent candidate to be convergent. This is so due to the properties proved in Sect. 3 and the encouraging numerical examples of Sect. 4.

Acknowledgments CDA and CEM acknowledge support by Universidad Nacional de Colombia through the project *Mathematics and Computation,* Hermes code 20305. CDA and PAA acknowledge support by COLCIENCIAS through the project *Programa jóvenes investigadores e innovadores 2012 (contrato 566),* Hermes code 16243. CEM acknowledge support by Universidad Nacional de Colombia through the project *Fortalecimiento del grupo de computación científica,* Hermes code 16084.

References

1. Acosta, C.D., Bürger, R., Mejía, C.E.: Monotone difference schemes stabilized by discrete mollification for strongly degenerate parabolic equations. Numer. Methods Partial Differ. Equat. **28**, 38–62 (2012)
2. Bürger, R., Coronel, A., Sepúlveda, M.: A semi-implicit monotone difference scheme for an initial-boundary value problem of a strongly degenerate parabolic equation modeling sedimentation-consolidation processes. Math. Comput. **75**, 91–112 (2005)
3. Chen, W., Ye, L., Sun, H.: Fractional diffusion equations by the Kansa method. Comput. Math. Appl. **59**, 1614–1620 (2010)
4. Cifani, S., Jakobsen, E.R.: Entropy solution theory for fractional degenerate convection-diffusion equations. Ann. Inst. H. Poincaré Anal. Non Linéaire. **28**, 413–441 (2011)
5. Cifani, S., Jakobsen, E.R.: On numerical methods and error estimates for degenerate fractional convection-diffusion equations. Numer. Math. **127**, 447–483 (2014)
6. Cockburn, B., Gripenberg, G., Londen, S.-O.: On convergence to entropy solutions of a single conservation law. J. Differ. Equat. **128**(1), 206–251 (1996)
7. Dalir, M., Bashour, M. Applications of fractional calculus. Appl. Math. Sci. **4**, 1021–1032 (2010)
8. Diethelm, K.: An algorithm for the numerical solution of differential equations of fractional order. Electron. Trans. Numer. Anal. **5**, 1–6 (1997)
9. Diethelm, K.: The analysis of fractional differential equations. Springer, Berlin (2010)
10. Holden, H., Karlsen, K., Risebro, N.: On uniqueness and existence of entropy solutions of weakly coupled systems of nonlinear degenerate parabolic equations. Electron. J. Differ. Equat. **2003**(46), 1–31 (2003)
11. LeVeque, R.: Numerical methods for conservation laws, 2nd edn. Birkhäuser-Verlag, Basel (1992)
12. Mohammadi, A., Manteghian, M., Mohammadi, A.: Numerical solution of one-dimensional advection-diffusion equation using simultaneously temporal and spatial weighted parameters. Aust. J. Basic Appl. Sci. **5**, 1536–1543 (2011)
13. Oldham, K.B., Spanier, J.: The fractional calculus. Dover Publications, Mineola (2006)
14. Pierantozzi, T.: Estudio de generalizaciones fraccionarias de las ecuaciones estándar de difusión y de ondas. Universidad Complutense de Madrid, Servicio de Publicaciones, Madrid (2007)
15. Podlubny, I.: Fractional differential equations. Academic, San Diego (1999)
16. Shen, S., Liu, F., Anh, V., Turner, I.: Detailed analysis of a conservative difference approximation for the time fractional diffusion equation. J. Appl. Math. Comput. **22**, 1–19 (2006)
17. Shen, S., Liu, F., Chen, J., Turner, I., Anh, V.: Numerical techniques for the variable order time fractional diffusion equation. Appl. Math. Comput. **218**, 10861–10870 (2012)

Dealing with Uncertainties in Computing: From Probabilistic and Interval Uncertainty to Combination of Different Types of Uncertainty

Vladik Kreinovich

Abstract To predict values of future quantities, we apply algorithms to the current and past measurement results. Because of the measurement errors and model inaccuracy, the resulting estimates are, in general, different from the desired values of the corresponding quantities. There exist methods for estimating this difference, but these methods have been mainly developed for the two extreme cases: the case when we know the exact probability distributions of all the measurement errors and the interval case, when we only know the bounds on the measurement errors. In practice, we often have some partial information about the probability distributions which goes beyond these bounds. In this chapter, we show how the existing methods of estimating uncertainty can be extended to this generic case.

Keywords Error estimation · Measurement errors · Model inaccuracy · Interval computations

1 Need to Deal with Uncertainty in Computing

Need for Data Processing To make a proper decision, we need to be able to predict the results of making a certain decision (or of not making any decision at all). In many real-life situations, we know how the desired future value y of each corresponding quantity depends on the current values of relevant quantities $q_1, \ldots, q_n$; in other words, we have an algorithm that, given the values $q_1, \ldots, q_n$, produces the estimate $y = A(q_1, \ldots, q_n)$. This algorithm can be as simple as a straightforward computation by using an explicit formula, or it can be as complex as a solution of the corresponding system of partial differential equations (as in weather prediction).

Sometimes, the quantities $q_1, \ldots, q_n$ can be measured directly; in such cases, to predict the future value y, we measure the current values of these quantities and use the algorithm f to predict the future value y.

V. Kreinovich (✉)
University of Texas at El Paso, El Paso, TX 79968, USA
e-mail: vladik@utep.edu

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_11

In many practical situations, however, some of the quantities q_i are difficult (or even impossible) to measure directly. For example, to make predictions in geosciences, we must know the densities and stresses at different depths, including areas much deeper that current boreholes can reach. In such situations, instead of directly measuring the corresponding quantity q_i, we can measure it *indirectly:* namely, we measure the auxiliary quantities $a_1, \ldots, a_m$ which are related to q_i by a known dependence, and then use a known algorithm to estimate q_i based on the results of these measurements. For example, to estimate the density at different depths, we measure gravity at different Earth locations, we measure travel times of seismic waves, etc. As a result, we arrive at the following problem:

- First, we (directly) measure some quantities; we will denote these quantities by $x_1, \ldots, x_n$. Some of these quantities may be the easy-to-measure quantities q_i, some may be auxiliary quantities whose measurement is needed to estimate difficult-to-measure quantities q_i.
- Then, we use the results $\widetilde{x}_1, \ldots, \widetilde{x}_n$ of the measured quantities $x_1, \ldots, x_n$ to compute the estimate $\widetilde{y}$ for the desired future value y. We will denote the corresponding algorithm by f, so that $\widetilde{f} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$. This algorithm usually consists of two parts:
  - First, we use the values $\widetilde{x}_j$ to estimate the quantities q_i.
  - Then, we use the estimated values of q_i to predict the value y.

Computation of $\widetilde{y}$ from $\widetilde{x}_i$ constitutes *data processing*.

Need to Deal with Uncertainty in Data Processing Measurements are never absolutely accurate. As a result, the measurement results $\widetilde{x}_i$ are, in general, different from the actual (unknown) values x_i of the corresponding quantity. In other words, in general, we have a nonzero *measurement error* $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$. Due to this difference, even when the model is exact, i.e., when the actual values y and x_i satisfy the condition $y = f(x_1, \ldots, x_n)$, the estimated value $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ is, in general, different from the actual value y.

In some cases, the model itself is only approximate, in the sense that y is only approximately equal to $f(x_1, \ldots, x_n)$. In this case, there is an additional *model inaccuracy* $\Delta x_0 \stackrel{\text{def}}{=} f(x_1, \ldots, x_n) - y$, and hence, the estimate $\widetilde{y}$ is even more different from y.

To make a proper decision based on the estimate $\widetilde{y}$, it is important to know how accurate is this estimate. For example, if the estimate for the amount of water is an underground aquifer of 200 million tons, and it is 200 ± 10, then it is a good idea to start digging and exploiting this water; on the other hand, if it is 200 ± 300, then it may be that there is no water available at all—in which case, further measurements may be needed before we invest money in exploiting this possible source of water.

In general, it is important to get some information about the estimation error $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$.

2 Processing Uncertainty: General Formulation of the Problem

Toward the General Formulation of the Problem We are interested in the difference $\Delta y = \widetilde{y} - y$.

- We know that $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.
- By definition of the model inaccuracy, we have $y = f(x_1, \ldots, x_n) - \Delta x_0$. By definition of the measurement error, we have $x_i = \widetilde{x}_i - \Delta x_i$, so

$$y = f(\widetilde{x}_1 - \Delta x_1, \ldots, \widetilde{x}_n - \Delta x_n) - \Delta x_0.$$

Substituting these expressions for $\widetilde{y}$ and y into the above formula for Δy, we conclude that

$$\Delta y = f(\widetilde{x}_1, \ldots, \widetilde{x}_n) - f(\widetilde{x}_1 - \Delta x_1, \ldots, \widetilde{x}_n - \Delta x_n) + \Delta x_0. \tag{1}$$

Measurement Errors Are Usually Relatively Small The measurement errors are usually relatively small; we may have measurement accuracy 10, 5, 1 %. In all these cases, the squares of the measurement errors can be safely ignored: e.g., for $\Delta x_i \approx 10\,\%$, we have $(\Delta x_i)^2 \approx 1\,\% \ll 10\,\%$. Due to this, we can expand the formula (1) in Taylor series in Δx_i and ignore terms which are quadratic (or of higher order) in Δx_i. We thus get $f(\widetilde{x}_1 - \Delta x_1, \ldots, \widetilde{x}_n - \Delta x_n) = f(\widetilde{x}_1, \ldots, \widetilde{x}_n) - \sum_{i=1}^{n} c_i \cdot \Delta x_i$, where $c_i \stackrel{\text{def}}{=} \dfrac{\partial f}{\partial x_i}$ and therefore, we get a linear dependence:

$$\Delta y = \sum_{i=1}^{n} c_i \cdot \Delta x_i + \Delta x_0. \tag{2}$$

Measurement Errors Corresponding to Different Measurements Are Usually Independent. Measurement errors Δx_i corresponding to different measurements are usually independent from each other—and from the model inaccuracy Δx_0. Therefore, it makes sense to assume that all $n + 1$ random variables $\Delta x_1, \ldots, \Delta x_n$, and Δx_0 are independent.

What We Do in This Chapter? In this chapter, we describe how to estimate Δy in a general situation, when we may have a combination of probabilistic and interval uncertainty. To provide this description, we need to first recall how uncertainty is usually estimated—so that it will be clear what are the assumptions underlying the usual techniques, and what needs to be modified when these assumptions are not satisfied.

3 Traditional Engineering Approach to Processing Uncertainty: Brief Reminder

Usual Assumptions: That All Distributions Are Normal with Zero Mean. In engineering practice, it is usually assumed that all the measurement errors are normally distributed with zero mean.

The normality assumption comes from the fact that for each measurement, the measurement error comes from many different sources. Usually, manufacturers of the measuring instrument try their best to eliminate all major sources of measurement errors. As a result, the remaining measurement error does not contain any large components, it is a joint effort of numerous small error components coming from different sources. According to the central limit theorem (see, e.g., [17]), the distribution of the sum of a large amount of small independent random components is close to Gaussian—and the more components we have, the closer the resulting distribution to Gaussian. Thus, it make sense to assume that the measurement errors are normally distributed—and indeed, empirical analysis shows that more than half of the measuring instruments have normal distribution [13, 14].

The zero mean assumption comes from the fact that the measuring instruments are usually calibrated before their use; see, e.g., [15]. One of the purposes of the calibration is to find the instrument's *bias*—i.e., the mean value of the measurement error—and to compensate for this bias. After the compensation, the mean is zero.

To describe a normal distribution, it is sufficient to describe the mean and the standard deviation. Since the mean of the variable Δx_i is zero, all we need to do to describe the measurement error is to provide the standard deviation σ_i. Similarly, we can eliminate the main sources of the model inaccuracy, and we can delete the model's bias as well. As a result, we can conclude that the model's inaccuracy Δx_0 is also normally distributed, with zero mean. We will denote its standard deviation by σ_0.

Estimating uncertainty under the usual assumptions: Derivation of the resulting formulas. According to the formula (2), the estimation error Δy is a linear combination of measurement errors Δx_i and of the model inaccuracy Δx_0. These quantities are independent, and (under the above assumptions) normally distributed. It is known that a linear combination of independent Gaussian random variables is also normally distributed, so Δy is also normally distributed.

To describe a normal distribution, it is sufficient to describe the mean and the standard deviation. Since the means of all the variables Δx_i and Δx_0 are zeros, the mean value of Δy is also equal to 0. Thus, under the usual engineering assumptions, to describe the probability distribution for Δy, it is sufficient to describe its standard deviation σ. The variance of the sum of independent random variables is equal to the sum of the variances, so from (2), we conclude that

$$\sigma^2 = \sum_{i=1}^{n} c_i^2 \cdot \sigma_i^2 + \sigma_0^2. \tag{3}$$

How to Actually Estimate σ: Toward the First Algorithm How can we actually estimate σ? To use this formula, we need to know the values c_i. These values are partial derivatives of the function $f(x_1, \ldots, x_n)$ describing the data-processing algorithm. When this algorithm consists of a straightforward application of an explicit formula, we can simply differentiate this formula and get an explicit expression for the corresponding derivatives. However, in general, the function $f(x_1, \ldots, x_n)$ is given as a complex algorithm, so it is not possible to perform an explicit differentiation.

A reasonable alternative is to use *numerical differentiation*. Numerical differentiation is based on the definition of the derivative as a limit:

$$\frac{\partial f}{\partial x_i} = \lim_{h_i \to 0} \frac{f(x_1, \ldots, x_{i-1}, x_i + h_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)}{h_i}.$$

By the definition of the limit, this means that for small h, we have

$$\frac{\partial f}{\partial x_i} \approx \frac{f(x_1, \ldots, x_{i-1}, x_i + h_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)}{h_i}.$$

For small h_i, we expand the expression $f(x_1, \ldots, x_{i-1}, x_i + h_i, x_{i+1}, \ldots, x_n)$ in Taylor series and keep only terms which are linear in h, getting

$$f(x_1, \ldots, x_{i-1}, x_i + h_i, x_{i+1}, \ldots, x_n) = f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) + h_i \cdot c_i.$$

From this formula, we can estimate c_i as the ratio:

$$c_i = \frac{f(x_1, \ldots, x_{i-1}, x_i + h_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)}{h_i}. \quad (4)$$

Substituting these expressions into the formula (3), we get

$$\sigma^2 = \sum_{i=1}^{n} \left(\frac{f(\ldots, x_{i-1}, x_i + h_i, x_{i+1}, \ldots) - f(\ldots, x_{i-1}, x_i, x_{i+1}, \ldots)}{h_i} \right)^2 \cdot \sigma_i^2 + \sigma_0^2.$$

Which values $h_1, \ldots, h_n$ should we use? Once we know the values of the function f, this formula uses subtraction, addition, multiplication, and division to estimate σ^2. In the computer, division is the most time-consuming operation, so ideally, we should select h_i so as to avoid divisions. Division can indeed be avoided if we take $h_i = \sigma_i$. In this case, the above formula takes the simplified form:

$$\sigma^2 = \sum_{i=1}^{n} (f(\ldots, x_{i-1}, x_i + \sigma_i, x_{i+1}, \ldots) - f(\ldots, x_{i-1}, x_i, x_{i+1}, \ldots))^2 + \sigma_0^2. \quad (5)$$

Thus, we arrive at the following algorithm.

First Algorithm: Sensitivity Analysis We are given the values $\widetilde{x}_1, \ldots, \widetilde{x}_n$, the algorithm f, and the standard deviations $\sigma_1, \ldots, \sigma_n$, and σ_0:

- First, we perform the original data processing, i.e., compute the value $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.
- Then, for $i = 1, \ldots, n$, we compute $y_i \stackrel{\text{def}}{=} f(\widetilde{x}_1, \ldots, \widetilde{x}_{i-1}, \widetilde{x}_i + \sigma_i, \widetilde{x}_{i+1}, \ldots, \widetilde{x}_n)$.
- Finally, we compute $\sigma^2 = \sum_{i=1}^{n} (y_i - \widetilde{y})^2 + \sigma_0^2$.

Comment. Our recommendation to use $h_i = \sigma_i$ differs from the usual numerical methods recommendation to use $h_i \approx \sqrt{\varepsilon}$, where ε is the machine epsilon. This usual recommendation makes perfect sense in the situations in which:

- The algorithm f provides very accurate computation of the corresponding function $f(x_1, \ldots, x_n)$, and
- The objective is to find the most accurate estimate of the derivatives.

In our application, none of these two conditions are satisfied.

First, since the input data come from measurement and are, thus, only approximately known, the data processing algorithms provide only approximate computation of the corresponding value—it makes no sense to compute, e.g., $\ln(x)$ with eight-digit accuracy if we only know x with accuracy 1 % (which, by the way, means very accurate measurements). Such an approximate algorithm f may not even take into account the much smaller difference $h_i \approx \sqrt{\varepsilon}$ between the values x_i and $x_i + \sqrt{\varepsilon}$, but this algorithm will definitely react to the difference of order σ_i between the values x_i and $x_i + \sigma_i$, since a difference of this order of magnitude corresponds to practically distinguishable difference between data values.

Second, by applying linearization—i.e., by replacing the exact formula (1) with an approximate formula (2)—we have already ignored quadratic terms in the expression σ, terms which even for very accurate measurements, with accuracy 1 %, leads to relative accuracy 10^{-4} of computing σ. Since the formula (3) is only valid with this accuracy, it does not make sense to spend additional computation time on estimating c_i too accurately.

In this case, as we have mentioned, the need to save computation time leads to $h_i = \sigma_i$.

Limitations of the First Algorithm As we have mentioned, the data processing algorithm f can be very time consuming. Thus, the more times we call this algorithm, the longer our estimation of σ. The above algorithm requires $n + 1$ calls to the algorithm f (n more calls than a simple data processing). In many practical problems—e.g., in geosciences—we process thousands of data points, so n is in thousands. If it takes several hours on a high-performance computer to estimate each value of f, then, to compute σ, the above algorithm requires thousands time more time—i.e., several months. This is not realistic, we need a faster method.

Towards a Second Algorithm The possibility to process uncertainty faster comes from the fact that a similar expression for σ arises if we *simulate* normally distributed random errors. Namely, if we add, to the original values $\widetilde{x}_i$, simulated random errors δx_i which are normally distributed with 0 mean and standard deviation σ_i, and use a random variable δx_0 which is normally distributed with mean 0 and standard deviation σ_0, then the difference:

$$f(\widetilde{x}_1 + \delta x_1, \ldots, \widetilde{x}_n + \delta x_n) - f(\widetilde{x}_1, \ldots, \widetilde{x}_n) + \delta x_0 = \sum_{i=1}^{n} c_i \cdot \delta x_i + \delta x_0$$

is also normally distributed with 0 mean and the desired standard deviation σ. We can thus use the standard formulas for estimating standard deviation from a sample to estimate σ. We therefore arrive at the following algorithm:

Second Algorithms Monte Carlo Simulations We are given the values $\widetilde{x}_1, \ldots, \widetilde{x}_n$, the algorithm f, and the standard deviations $\sigma_1, \ldots, \sigma_n$, and σ_0:

- First, we perform the original data processing, i.e., compute the value $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.
- Then, we select the number of iterations N. For each k from 1 to N, we generate $n+1$ random numbers $r_{k1}, \ldots, r_{kn}, r_{k0}$ each of which is normally distributed with mean 0 and standard deviation 1.
- For each k, we compute $y_k = f(\widetilde{x}_1 + \sigma_1 \cdot r_{k1}, \ldots, \widetilde{x}_n + \sigma_n \cdot r_{kn}) + \sigma_0 \cdot r_{k0}$.
- Finally, we estimate $\sigma^2 = \dfrac{1}{N} \cdot \sum_{k=1}^{N} (y_k - \widetilde{y})^2$.

Advantages and Limitations of the Second Algorithm The above method requires $N+1$ calls to the algorithm f. The number of iterations N depends on the accuracy with which we want to estimate σ. In general, the relative standard deviation of determining σ from a sample of size N is equal to $\sqrt{\dfrac{2}{N}}$; so, e.g., to find σ with accuracy 20 % and reliability 95 % (which corresponds to two standard deviations), we need to make sure that $2 \cdot \sqrt{\dfrac{2}{N}} \le 0.2$, i.e., $N \ge 200$. For $n \gg 1000$, this is much faster than the sensitivity analysis—this is the main advantage of this method.

The limitation is that, in contrast to the sensitivity analysis method, we do not get the exact value σ, only an approximate value.

Possibility of Parallelization In both methods for estimating σ, the most time-consuming step is calling the algorithm f. If we have at our disposal several processors which can work in parallel, then we can make all these calls in parallel and thus, drastically decrease the computation time.

Comment In situations when we know the actual step-by-step code of the data processing algorithm f, there is another way to save computation time. Namely, we can apply, to the known code, the procedure of reverse differentiation (also known as *backpropagation* or *adjoint methods*) which allows us to compute the values x_i of all n partial derivatives c_i in time which, theoretically, is no more than three times longer than the time needed to compute the value f itself; see, e.g., [5, 18, 19]. Once we have computed all the values c_i of the gradient, we can use the formula (3) to compute the desired standard deviation σ.

This method is indeed effectively used, e.g., in neural networks [18, 19]. However, in many practical situations, the actual computational overhead of using reverse differentiation is much higher to the extent than Monte Carlo methods are faster.

Besides, in some practical situations, data processing uses proprietary programs, programs for which the code is not provided to the user. The only way to use these programs is to treat the data processing algorithm as a "black box": the only thing we can compute are the output values $f(x)$ corresponding to different inputs x. For such programs, it is not possible to use reverse differentiation, and the only possibility to reduce the computation time in comparison with sensitivity analysis is to use Monte Carlo techniques.

4 Case of Interval Uncertainty

Need for Interval Uncertainty The traditional approach is based on the assumption that for each measuring instrument, we know the exact distribution of the corresponding measurement error Δx_i. In practice, this probability distribution can be established if we compare the results $\widetilde{x}_i^{(k)}$ produced by our measuring instrument with the results $\widetilde{x}_{i,\mathrm{st}}^{(k)}$ produced by a much more accurate ("standard") measuring instrument. As the standard measuring instrument is much more accurate, we can ignore its measurement errors and assume that its measurement results are equal to the exact values of the corresponding quantity: $\widetilde{x}_{i,\mathrm{st}}^{(k)} \approx x_i^{(k)}$. In this approximation, the differences $\widetilde{x}_i^{(k)} - \widetilde{x}_{i,\mathrm{st}}^{(k)}$ are equal to the corresponding measurement errors $\Delta x_i^{(k)} = \widetilde{x}_i^{(k)} - x_i^{(k)}$. By accumulating a sample of such values, we get a probability distribution for Δx_i.

However, there are two situations when we cannot do it. First is the case of state-of-the-art measurements. For example, it would be nice if near the Hubble telescope, there would be another one, five times more accurate, which we could use to calibrate the Hubble telescope—but the Hubble telescope is the best we have. Similarly, it would be nice if we had geophysical methods which were five times more accurate than the current ones—but our methods are the best we have. In such situations, at best, we can have upper bounds Δ_i on the corresponding measurement errors. We know that $|\Delta x_i| \le \Delta_i$, i.e., that Δx_i is located on the interval $[-\Delta_i, \Delta_i]$, but we do not have any information about which values from this interval are more probable and which values are less probable. This situation is known as *interval uncertainty*; see, e.g., [6, 11, 15].

Interval uncertainty also occurs in *manufacturing,* where, in principle, we can calibrate every sensors, but since sensors are relatively cheap and their calibration is very expensive, they are not calibrated—instead, we rely on the upper bounds Δ_i provided by the manufacturer.

Similarly, we only know a bound Δ_0 on the model inaccuracy Δx_0: $|\Delta x_0| \le \Delta_0$.

Estimating Uncertainty Under Interval Uncertainty: Derivation of the Resulting Formulas The sum (2) is the largest when each term $c_i \cdot \Delta x_i$ attains its largest possible value on the corresponding interval $[-\Delta_i, \Delta_i]$.

- When $c_i \ge 0$, the function $c_i \cdot \Delta x_i$ is increasing, so its largest value is attained for the largest possible value $\Delta x_i = \Delta_i$. This largest value is equal to $c_i \cdot \Delta_i$.
- When $c_i \le 0$, the function $c_i \cdot \Delta x_i$ is decreasing, so its largest value is attained for the smallest possible value $-\Delta x_i = \Delta_i$. This largest value is equal to $-c_i \cdot \Delta_i$.

In both cases, the largest possible value is $|c_i| \cdot \Delta_i$. Similarly, in both cases, the smallest possible value is $-|c_i| \cdot \Delta_i$. Thus, the range of possible values of Δy is equal to $[-\Delta, \Delta]$, where

$$\Delta = \sum_{i=1}^{n} |c_i| \cdot \Delta_i + \Delta_0. \tag{6}$$

Why Not Use Maximum Entropy Approach? In statistics, situations when we do not know the exact probability distribution are frequent. In such case, if we have several possible distributions consistent with our knowledge, a reasonable idea is to select the distribution with the largest value of the *entropy* $S \stackrel{\text{def}}{=} -\int \rho(x) \cdot \ln(\rho(x))\,\mathrm{d}x$, where $\rho(x)$ is the probability density; see, e.g., [7]. If we only know that the random variable is located on an interval, then this maximum entropy approach leads to a uniform distribution on this interval. (For several variables, if we know nothing about their correlation, the maximum entropy approach implies that they are independent.)

At first glance, this makes perfect sense—and this is how many practitioners deal with interval uncertainty. However, we can show that this approach can drastically underestimate the uncertainty Δy. We can illustrate it on the example of the simplest possible dependence, when $f(x_1, \ldots, x_n) = x_1 + \cdots + x_n$ and therefore, $\Delta y = \Delta x_1 + \cdots + \Delta x_n$. For simplicity, we can assume that all the upper bounds are the same: $\Delta_1 = \cdots = \Delta_n$. In this case, the formula (6) implies that $\Delta = n \cdot \Delta_1$. This is possible, e.g., if each measurement error is exactly equal to Δ_1.

On the other hand, according to the maximum entropy approach, each value Δx_i is uniformly distributed on the interval $[-\Delta_1, \Delta_1]$. This distribution has mean 0 and variance $\frac{1}{3} \cdot \Delta_1^2$. For large n, the sum Δy of these independent random variables is approximately normally distributed (the same central limit theorem that we cited earlier). The mean of Δy is equal to the sum of 0s, i.e., to 0, and its variance is equal to the sum of the variances, i.e., $\sigma^2 = \frac{n}{3} \cdot \Delta_1^2$. For a normal distribution, the values are located in the six-sigma interval with practically absolute certainty; thus, we can take $6\sigma \sim \sqrt{n}$ as an upper bound for Δy. For large n, this is much smaller than the above upper bound $n \cdot \Delta_1$. Thus, the maximum entropy approach is not applicable, and we have to use the formula (6).

How to Actually Estimate Δ: Toward the First Algorithm How can we actually estimate Δ? If we substitute the above numerical differentiation formula for c_i into the formula (6), we conclude that

$$\Delta = \sum_{i=1}^{n} \left| \frac{f(\ldots, x_{i-1}, x_i + h_i, x_{i+1}, \ldots) - f(\ldots, x_{i-1}, x_i, x_{i+1}, \ldots)}{h_i} \right| \cdot \Delta_i + \Delta_0.$$

Which values $h_1, \ldots, h_n$ should we use? Similarly to the traditional case, we select the values h_i for which we can avoid division and thus, speed up computations. Division can indeed be avoided if we take $h_i = \Delta_i$. In this case, the above formula takes the simplified form:

$$\sigma^2 = \sum_{i=1}^{n} \left| f(\ldots, x_{i-1}, x_i + \Delta_i, x_{i+1}, \ldots) - f(\ldots, x_{i-1}, x_i, x_{i+1}, \ldots) \right| + \Delta_0. \quad (7)$$

Thus, we arrive at the following algorithm.

First Algorithm: Sensitivity Analysis We are given the values $\widetilde{x}_1, \ldots, \widetilde{x}_n$, the algorithm f, and the bounds deviations $\Delta_1, \ldots, \Delta_n$, and δ:

- First, we perform the original data processing, i.e., compute the value $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.
- Then, for $i = 1, \ldots, n$, we compute $y_i \stackrel{\text{def}}{=} f(\widetilde{x}_1, \ldots, \widetilde{x}_{i-1}, \widetilde{x}_i + \Delta_i, \widetilde{x}_{i+1}, \ldots, \widetilde{x}_n)$.
- Finally, we compute $\Delta = \sum_{i=1}^{n} |y_i - \widetilde{y}| + \Delta_0$.

Limitations of the First Algorithm Similarly to the traditional case, this algorithm requires $n + 1$ calls to the algorithm f and is, thus, often too slow.

Toward a Second Algorithm The possibility to process uncertainty faster comes from the fact that for random variables distributed according to the Cauchy distribution, with probability density $\rho(x) = \frac{1}{\pi \cdot \Delta} \cdot \frac{1}{(x/\Delta)^2 + 1}$, a linear combination (2) of variables Δx_i which are Cauchy distributed with parameters Δ_i is Cauchy distributed with parameter Δ determined by the formula (7). We therefore arrive at the following algorithm [8, 9]:

Second Algorithm: Monte Carlo Simulations We are given the values $\widetilde{x}_1, \ldots, \widetilde{x}_n$, the algorithm f, and the bounds $\Delta_1, \ldots, \Delta_n$, and s_m:

- First, we perform the original data processing, i.e., compute the value $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.
- Then, we select the number of iterations N. For each k from 1 to N, we generate $n + 1$ random numbers $r_{k1}, \ldots, r_{kn}$ each of which is uniformly distributed on the interval $[0, 1]$.
- Then, we compute Cauchy distributed values $c_{ki} = \tan(\pi \cdot (r_{ki} - 0.5))$.
- We compute the maximum $K_k = \max_i |c_{ki}|$ so that we will be able to normalize the simulated approximation errors and apply f to the values that are within the box of possible values.
- For each k, we compute:

$$\Delta y_k = K_k \cdot \left(f\left(\widetilde{x}_1 + \Delta_1 \cdot \frac{c_{k1}}{K_k}, \ldots, \widetilde{x}_n + \Delta_n \cdot \frac{c_{kn}}{K_k} \right) - \widetilde{y} \right).$$

- We compute Δ' by applying the bisection method to solve the equation:

$$\frac{1}{1 + \left(\frac{\Delta y^{(1)}}{\Delta'} \right)^2} + \cdots + \frac{1}{1 + \left(\frac{\Delta y^{(N)}}{\Delta'} \right)^2} = \frac{N}{2}.$$

- Finally, we return $\Delta = \Delta' + \Delta_0$.

Advantages and Limitations of the Second Algorithm The above method requires $N + 1$ calls to the algorithm f. Similarly to the usual Monte Carlo method, the number of iterations N depends on the accuracy with which we want to estimate σ. For $n \gg 200$, this is much faster than the sensitivity analysis—this is the main advantage of this method. The limitation is that, in contrast to the sensitivity analysis method, we do not get the exact value Δ, but only an approximate value.

Possibility of Parallelization Similarly to the statistical case, in both methods for estimating σ, the most time consuming step is calling the algorithm f. So, if we have at our disposal several processors which can work in parallel, then we can make all these calls in parallel and thus, drastically decrease the computation time.

Comment A numerical example of using this Cauchy-based Monte Carlo method is given in [9].

5 Need to Go Beyond Traditional and Interval Cases

What We Have Considered So Far Up to now, we considered two extreme cases:

- The traditional case, when all measurement errors are normally distributed with zero mean.
- The interval case, when we only know the upper bounds on the measurement errors.

Need to Go Beyond These Cases In practice, we often have cases in between:

- In some cases, we know the distributions, and these distributions are non-Gaussian. This is actually the case for almost half (40 %) of the measuring instruments; see, e.g., [13, 14].
- In some other cases, we do not know the exact probability distributions—but we have some partial information about these distributions which go beyond the upper bounds.

What We Do in This Chapter In this chapter, we describe how to estimate uncertainty in the general case.

6 Case of Known Non-Gaussian Distributions

Formulation of the Problem Let us first consider the case when we know the probability distributions of all the measurement errors Δx_i, and the probability distribution of the model error Δx_0. For example, these probability distributions are represented in terms of the probability density functions $\rho_i(\Delta x_i)$ and $\rho_0(\Delta x_0)$.

We know that the corresponding variables are independent. Our goal is to find the probability distribution of the quantity Δy—as described by the formula (2).

Two Types of Algorithms Similarly to the above two cases, we will consider two types of algorithms for solving this problem: algorithms which produce the exact answer, and faster Monte Carlo-type algorithms which produce approximate answers.

Algorithm for Exact Computation: General Idea

- First, we use numerical differentiation (4) to estimate the coefficients c_i.
- For each i, we can then compute the probability density functions corresponding to $t_i \stackrel{\text{def}}{=} c_i \cdot \Delta x_i$ as $d_i(t_i) = \dfrac{1}{c_i} \cdot \rho_i\left(\dfrac{t_i}{c_i}\right)$.
- Then, we can apply several times the known convolution formula $\rho_c(x) = \int \rho_a(t) \cdot \rho_b(x - t)\,\mathrm{d}t$ for the probability density of the sum $c = a + b$ of independent random variables to find the probability density corresponding to the sum $\Delta y = \sum_{i=1}^{n} t_i + \Delta x_0$:
  - First, we combine the probability distributions of t_1 and t_2 to compute the probability density of the sum $t_1 + t_2$.
  - Then, we combine the probability distributions of $t_1 + t_2$ and t_3 to compute the probability density of the sum $t_1 + t_2 + t_3$.
  - ...
  - Finally, we combine the probability distributions of $\sum_{i=1}^{n} t_i$ and Δx_0 to compute the probability density of $\Delta y = \sum_{i=1}^{n} t_i + \Delta x_0$.

How to Compute Convolutions Faster One possibility to compute the probability density function of the sum is to perform a straightforward computation of each convolution integral $\rho_c(x) = \int \rho_a(t) \cdot \rho_b(x - t)\,\mathrm{d}t$. If we represent each of the probability density functions by its values at M different points $\rho_a(v_k)$ and $\rho_b(v_k)$ for $v_k = k \cdot \Delta v$, then each computation takes the form $\rho_c(v_k) = \sum_\ell \rho_a(v_\ell) \cdot \rho_b(v_{k-\ell}) \cdot \Delta v$. This computation requires M^2 computational steps: M steps for each value k.

It is known, however, that we can speed up the computation of convolution if we use *Fourier transforms*, i.e., if instead of the original probability density functions $\rho_a(x)$ and $\rho_b(x)$, we use the corresponding *characteristic functions:*

$$\chi_a(\omega) \stackrel{\text{def}}{=} E[\exp(\mathrm{i} \cdot \omega \cdot a)] = \int \exp(\mathrm{i} \cdot x \cdot \omega) \cdot \rho_a(x)\,\mathrm{d}x$$

and

$$\chi_b(\omega) \stackrel{\text{def}}{=} E[\exp(\mathrm{i} \cdot \omega \cdot b)] = \int \exp(\mathrm{i} \cdot x \cdot \omega) \cdot \rho_b(x)\,\mathrm{d}x.$$

Namely, it is known that the characteristic function of the sum is equal to the product of the characteristic functions. Thus, we can compute the convolution as follows; see, e.g., [1]:

- First, we the use the fast Fourier transform algorithm to compute the Fourier transforms $\chi_a(\omega)$ and $\chi_b(\omega)$ of the corresponding probability density functions. This computation takes $O(M \cdot \log(M))$ computational steps.
- Then, we multiply the corresponding values of the Fourier transform element by element to compute $\chi_c(\omega) = \chi_a(\omega) \cdot \chi_b(\omega)$. To compute M values of this new characteristic function, we need M computational steps.
- Finally, we apply the inverse fast Fourier transform algorithm to the function $\chi_c(\omega)$ and thus, find the desired probability density function $\rho_c(x)$. This computation also takes $O(M \cdot \log(M))$ computational steps.

Thus, overall, we need $O(M \cdot \log(M)) + O(M) + O(M \cdot \log(M)) = O(M \cdot \log(M))$ computational steps to compute the convolution, which, for large M, is much smaller than M^2 steps needed for the straightforward computation of the convolution.

Faster Computation of the Convolution Can Speed Up the Computation of the Probability Density Function $\rho(\Delta y)$ For the sum Δy of $n+1$ random variables $t_1, \ldots, t_n$, and Δx_0, the characteristic function $\chi(\omega)$ is equal to the product of the characteristic functions $\chi_i(\omega)$ and $\chi_0(\omega)$ of these random variables. Thus:

- First, we the use the fast Fourier transform algorithm to compute the Fourier transforms $\chi_i(\omega)$ and $x_m(\omega)$ of the corresponding probability density functions $d_i(t_i)$ and $\rho_0(\Delta x_0)$. This computation takes $(n+1) \cdot O(M \cdot \log(M))$ computational steps.
- Then, we multiply the corresponding values of the Fourier transform element-by-element to compute $\chi(\omega) = \chi_1(\omega) \cdot \ldots \chi_n(\omega) \cdot \chi_0(\omega)$. To compute M values of this new characteristic function, we need $n \cdot M$ computational steps.
- Finally, we apply the inverse fast Fourier transform algorithm to the function $\chi(\omega)$ and thus, find the desired probability density function corresponding to Δy. This computation takes $O(M \cdot \log(M))$ computational steps.

Overall, we thus need $O(n \cdot M \cdot \log(M))$ computational steps.

Possible use of Central Limit Theorem: Discussion The larger the number n of inputs $x_1, \ldots, x_n$, the more computation time we need. On the other hand, when n is large, this means that most of the contributions $t_i = c_i \cdot \Delta x_i$ to the overall error Δy are relatively small. In this case, as we have mentioned earlier, we can invoke the central limit theorem and conclude that the probability distribution for the sum of these small contributions is close to Gaussian.

A Gaussian distribution is uniquely determined by its mean and standard deviation (or, equivalently, variance), and the mean and variance of the sum of several independent random variables is equal to the sum of the corresponding means and variances. Thus, for the small components, there is no need to use their full probability density functions: it is sufficient to estimate their means and variances, then add them, and then add the resulting Gaussian sum to the few non-small components.

Thus, we arrive at the following algorithm.

Use of the Central Limit Theorem: Resulting algorithm This algorithm requires that we know the list of non-small components. Without losing generality, let us assume that the components $t_1, \ldots, t_k$, $k \ll n$ (and Δx_0) are non-small, and that all the other components $t_{k+1}, \ldots, t_n$ are small.

For each small component t_i, we use the probability distribution $d_i(t_i)$ to compute the mean $\mu_i = \int x \cdot d_i(x)\,\mathrm{d}x$ and the variance $\sigma_i^2 = \int (x - \mu_i)^2 \cdot d_i(x)\,\mathrm{d}x$. Then, we compute the overall mean and variance of the sum of all the small components as $\mu = \sum_{i=k+1}^{n} \mu_i$ and $\sigma^2 = \sum_{i=k+1}^{n} \sigma_i^2$, and we form a probability distribution function:

$$\rho_{\mathrm{sm}} = \frac{1}{\sqrt{\pi} \cdot \sigma} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

This is a probability distribution for the sum $\sum_{i=k+1}^{n} t_i$ of all small components.

Then, we combine the probability distributions for $t_1, \ldots, t_k$, $\sum_{i=k+1}^{n} t_i$, and Δx_0.

Monte Carlo-Type Algorithm To decrease the number of calls to the algorithm f and thus, to speed up the computations, we can simulate the measurement errors. To simulate a measurement error t_i distributed according to the probability density $d_i(t_i)$, we can perform the following preliminary computations:

- Form the cumulative distribution function (cdf) $F_i(x) = \int^x \mathrm{d}_i(t)\,\mathrm{d}t$,
- Form its inverse function $F_i^{-1}(p)$ – by computing, for every value $p \in [0, 1]$, the value $x = F^{-1}(p)$ for which $F_i(x) = p$.

After that, on each iteration k, we generate a random number r_{ki} which is uniformly distributed on the interval [0, 1], and compute $c_{ki} = F^{-1}(r_{ki})$. Similarly, we simulate a number c_{k0} which is distributed according to the probability density function $\rho_0(\Delta x_0)$.

We then compute simulated values:

$$\Delta y^{(k)} = (f(\widetilde{x}_1 - c_{k1}, \ldots, \widetilde{x}_n - c_{kn}) + c_{k0}) - \widetilde{y}.$$

Based on the sample of these values, we can now determine the probability distribution for Δy.

Use of the Central Limit Theorem Due to the central limit theorem, for small components, instead of simulating their exact distributions, we can simulate normally distributed random variables with the same values of mean and standard deviation.

Parallelization Can Lead to a Further Speed Up In all these methods, the most time-consuming step is calling the algorithm f. If we have at our disposal several processors which can work in parallel, then we can make all these calls in parallel and thus, drastically decrease the computation time.

It is also possible to parallelize further processing of these values. For example, in the algorithm using Fourier transforms, we can compute each of $n + 1$ Fourier transforms in parallel—and if we have more than $n + 1$ processors, then we can also perform each fast Fourier transform in parallel. In the case of unlimited number of processors, this can be done in time $O(\log(M))$.

Similarly, each of the products $\chi(\omega)$ can be computed in parallel, and, if needed, each computation of a product can also be parallelized:

- First, we multiply all the neighboring pairs $\chi_{2i-1}(\omega) \cdot \chi_{2i}(\omega)$.
- Then, we multiply product of neighboring pairs into products of neighboring four tuples,
- etc.

In this manner, in the ideal case of unlimited number of processors, we compute all the products in time $O(\log(M))$—and thus, finish all the computations in time $O(\log(M))$.

7 Case of Partial Information about Probabilities: How to Represent this Partial Information?

Need to Select a Representation In many real-life situations, we have only partial information about the probability distribution of measurement errors. How can we represent this partial information?

In principle, we can represent a probability distribution in many different forms:

- By its probability density function
- By its cdf
- By its moments, etc.

Which representation should we use?

To Select a Representation, We Need to Take into Account the Ultimate Goal of Decision Making As we have mentioned, one of the main reasons why we need to take into account uncertainty in data processing is that this uncertainty affects our decisions. From the viewpoint of decision making, what is the best way to represent partial information about the probabilities?

It is known that a consistent decision making can be described as optimizing an expected value of a special function $u(x)$ known as *utility*; see, e.g., [4, 10, 12, 16]. The utility function $u(x)$ describes the user preferences. Thus, it makes sense to select characteristics of the probability distribution which can help us compute this expected utility $\int \rho(x) \cdot u(x)\,\mathrm{d}x$.

In particular, for measurement errors $\Delta x_i = \widetilde{x}_i - x_i$, the loss of utility is caused by the fact that while the only information that we can use about x_i is the measurement result $\widetilde{x}_i$, the actual value x_i is, in general, different from $\widetilde{x}_i$. For example, if we want to dress appropriately for the weather, we must know the exact temperature; if we know it approximately, then there is a strong chance that we will dress either too warm or too cold. In general, the expected utility has the form $\int \rho_i(\Delta x_i) \cdot u(\Delta x_i)\,\mathrm{d}\Delta x_i$.

Ideally, the perfect situation is when $\Delta x_i = 0$ and the actual value x_i is exactly equal to the measurement result $\widetilde{x}_i$. In this case, we prepare for exactly the proper conditions, so the utility attains its maximum value.

It is, however, possible that we know that the measuring instrument has a bias, and we know the approximate value of this bias b. In this case, when the measurement result is $\widetilde{x}_i$, we prepare for the de-biased value $x_i = \widetilde{x}_i - b$. So, even if $\Delta x_i = 0$, the actual condition $x_i = \widetilde{x}_i$ is somewhat different from the value $x_i = \widetilde{x}_i - b$ for which we are prepared.

Case of Smooth Utility Functions: Analysis of the Problem Let us first consider the case when the utility function smoothly changes with Δx_i. We consider the case when measurement errors are relatively small. This means that the values Δx_i are close to 0, so we can expand the utility function $u(\Delta x_i)$ in Taylor series and keep only the first few terms in this expansion.

In Sect. 2, we made a similar statement about the function f, and for this function, we decided to keep only linear terms, terms determined by its first derivatives c_i taken

at the point $\widetilde{x}_i$ (i.e., at the point $x_i = \widetilde{x}_i - \Delta x_i$ corresponding to $\Delta x_i = 0$). For the utility function, this is not always possible: As we have mentioned, for the unbiased measuring instrument, the utility function attains its maximum when $\Delta x_i = 0$ and thus, its first derivative is equal to 0. So, for the utility function, we also need to take into account second-order terms: $u(\Delta x_i) = u_0 + u_1 \cdot \Delta x_i + u_2 \cdot (\Delta x_i)^2 + \ldots$, for some values u_0 and u_2.

Since the values Δx_i are assumed to be small, we can thus ignore cubic and higher-order terms in this expansion, and conclude that $u(\Delta x_i) = u_0 + u_1 \cdot \Delta x_i + u_2 \cdot (\Delta x_i)^2$. For this utility function, the expected utility has the form:

$$\int \rho_i(\Delta x_i) \cdot u(\Delta x_i)\, \mathrm{d}\Delta x_i = u_0 + u_1 \cdot \int \Delta x_i \cdot \rho_i(\Delta x_i)\, \mathrm{d}\Delta x_i + u_2 \cdot \int (\Delta x_i)^2 \cdot \rho_i(\Delta x_i)\, \mathrm{d}\Delta x_i,$$

i.e., the form $u_0 + u_1 \cdot \mu_i + u_2 \cdot M_i$, where μ_i is the expected value of the measurement error (*bias*) and M_i is the second moment of the measurement error. So, in the case of a smooth utility function, to describe the probability distribution, it is reasonable to use its first two moments.

Our goal is not just to represent these measurement errors Δx_i, we also want to use this information to characterize the linear combination (2) of these measurement errors. From this viewpoint, it is more convenient, instead of the second moments M_i, to use variances $\sigma_i^2 = M_i - \mu_i^2$, since the variance is the easiest to process: The variance of the sum of two independent random variables is equal to the sum of the corresponding variances. Therefore, a reasonable representation of a probability distribution should consist of the mean μ_i and the standard deviation σ_i. Similarly, a reasonable way to describe the probability distribution of the model error Δx_0 is to describe its mean μ_0 and standard deviation σ_0.

In terms of metrology (measurement theory and practice), μ_i is a *systematic error component*, and σ_i is known as a standard deviation of the *random error components*; see, e.g., [15].

Partial information means that we do not know the exact values of μ_i and σ_i. Instead, we only know the *bounds* on these values, i.e., we know the intervals $[\underline{\mu}_i, \overline{\mu}_i]$ and $[\underline{\sigma}_i, \overline{\sigma}_i]$ that contain the actual (unknown) values of mean and standard deviation.

Which characteristics of Δy should we compute based on these values? A similar analysis shows that we want to know the values of the corresponding mean μ and standard deviation σ.

Different possible values μ_i and σ_i from the corresponding intervals lead, in general, to different values of μ and σ; so, what we really want to compute are the ranges of possible values of μ and σ. Thus, we arrive at the following problem.

Case of a Smooth Utility Function: Precise Formulation of the Resulting Computational Problem We know:

- The intervals $[\underline{\mu}_i, \overline{\mu}_i]$ and $[\underline{\sigma}_i, \overline{\sigma}_i]$ containing the means and standard deviations of $n + 1$ independent random variables Δx_i
- The algorithm f

We want to find the ranges $[\underline{\mu}, \overline{\mu}]$ and $[\underline{\sigma}, \overline{\sigma}]$ of possible values of the mean μ and standard deviation σ of the quantity Δy described by the formulas (1) and (2).

How to Compute the Range of Possible Values of μ: Analysis of the Problem The mean of a linear combination is equal to the linear combination of the means, so we have

$$\mu = \sum_{i=1}^{n} c_i \cdot \mu_i + \mu_0.$$

We want to use the above interval-computation formulas from Sect. 4 to find the range of values of this linear function. For that purpose, we need to represent the corresponding intervals in the form $[\widetilde{\mu}_i - \Delta_i, \widetilde{\mu}_i + \Delta_i]$. By equating $\widetilde{\mu}_i - \Delta_i$ with $\underline{\mu}_i$ and $\widetilde{\mu}_i + \Delta_i$ with $\overline{\mu}_i$, we get a system of two equations with two unknowns $\widetilde{\mu}_i$ and Δ_i, from which we can conclude that:

- The value $\widetilde{\mu}_i$ is equal to the midpoint.
- The value Δ_i is equal to the half width of the corresponding interval:

$$\widetilde{\mu}_i = \frac{\underline{\mu}_i + \overline{\mu}_i}{2} \text{ and } \Delta_i = \frac{\overline{\mu}_i - \underline{\mu}_i}{2}.$$

For the differences $\Delta\mu_i \stackrel{\text{def}}{=} \widetilde{\mu}_i - \mu_i$, we have a limitation $|\Delta\mu_i| \le \Delta_i$. Thus, the general formulas for the range of a function f (from Sect. 4) lead to a conclusion that the range of possible values of μ is equal to $[\widetilde{\mu} - \Delta, \widetilde{\mu} + \Delta]$, where $\widetilde{\mu} = \sum_{i=1}^{n} c_i \cdot \widetilde{\mu}_i + \widetilde{\mu}_0$ and $\Delta = \sum_{i=1}^{n} |c_i| \cdot \Delta_i + \Delta_0$.

Due to the formulas (1) and (2), the value $\widetilde{\mu}$ can be computed simply as $\widetilde{y} - f(\widetilde{x}_1 - \widetilde{\mu}_1, \ldots, \widetilde{x}_n - \widetilde{\mu}_n) + \widetilde{\mu}_0$. The value Δ can be computed by using one of the two interval computations algorithms. Thus, we arrive at the following algorithms.

How to Compute the Range of Possible Values of μ: Algorithm

- First, we perform the original data processing, and compute the value $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.
- Then, for all i, we compute $\widetilde{\mu}_i = \frac{\underline{\mu}_i + \overline{\mu}_i}{2}$ and $\Delta_i = \frac{\overline{\mu}_i - \underline{\mu}_i}{2}$.
- After that, we compute the value $\widetilde{\mu} = \widetilde{y} - f(\widetilde{x}_1 - \widetilde{\mu}_1, \ldots, \widetilde{x}_n - \widetilde{\mu}_n) + \widetilde{\mu}_0$, and we use one of the two interval computation algorithms from Sect. 4 to compute $\Delta = \sum_{i=1}^{n} |c_i| \cdot \Delta_i + \Delta_0$.
- Finally, we compute the desired range $[\widetilde{\mu} - \Delta, \widetilde{\mu} + \Delta]$.

How to Compute the Range of Possible Values of σ: Analysis of the Problem The variance of the sum is equal to the sum of the variances, so we have $\sigma^2 = \sum_{i=1}^{n} c_i^2 \cdot \sigma_i^2 + \sigma_0^2$. This expression is increasing in σ_i, so:

- Its largest possible value $\overline{\sigma}$ is attained when each of the values σ_i attains its largest possible value $\overline{\sigma}_i$, so we have $(\overline{\sigma})^2 = \sum_{i=1}^{n} c_i^2 \cdot (\overline{\sigma}_i)^2 + (\overline{\sigma}_0)^2$.
- Its smallest possible value $\underline{\sigma}$ is attained when each of the values σ_i attains its smallest possible value $\underline{\sigma}_i$, so we have $(\underline{\sigma})^2 = \sum_{i=1}^{n} c_i^2 \cdot (\underline{\sigma}_i)^2 + (\underline{\sigma}_0)^2$.

Each of these formulas is of type (3), so we can use the two algorithms from Sect. 3 to perform these computations. In other words, we arrive at the following algorithm.

How to Compute the Range of Possible Values of σ: Algorithm

- First, we use one of the algorithms from Sect. 3 to compute the value $\overline{\sigma}$ from the formula $(\overline{\sigma})^2 = \sum_{i=1}^{n} c_i^2 \cdot (\overline{\sigma}_i)^2 + (\overline{\sigma}_0)^2$.
- Then, we use the same algorithm to compute the value $\underline{\sigma}$ from the formula $(\underline{\sigma})^2 = \sum_{i=1}^{n} c_i^2 \cdot (\underline{\sigma}_i)^2 + (\underline{\sigma}_0)^2$.

Case of Discontinuous Utility Function In some cases, the utility function is not smooth, but discontinuous. For example, at a manufacturing plant, we want to make sure that the possible pollution does not exceed a certain legal level. In such situations, there are stiff penalties for exceeding the level.

The expected value of this utility function is thus proportional to the probability of exceeding (or not exceeding) a certain level. For a random variable η, the corresponding probabilities $F(x) \stackrel{\text{def}}{=} \text{Prob}(\eta \le x)$ form a cdf. For such utility functions, an appropriate representation of the probability distribution is thus the cdf $F(x)$.

Partial information means that we may not know all the values $F(x)$ of the cdf; instead, we only know *bounds* $[\underline{F}(x), \overline{F}(x)]$. Such bounds are known as a *probability box,* or a *p-box,* for short; see, e.g., [2, 3]. So, we arrive at the following problem.

Case of a Discontinuous Utility Function: Precise Formulation of the Resulting Computational Problem We know:

- The p-boxes $[\underline{F}_i(x), \overline{F}_i(x)]$ describing the probability distribution of $n + 1$ independent random variables Δx_i, and
- The algorithm f.

We want to find the ranges $[\underline{F}(x), \overline{F}(x)]$ of possible values of the cdf $F(x)$ for the quantity Δy described by the formulas (1) and (2).

How to Compute the Corresponding p-Box: Analysis of the Problem The desired quantity Δ is the sum of several terms $t_i = c_i \cdot \Delta x_i$ and $t_0 = \Delta x_0$. Thus, it makes sense to first find the p-boxes $[\underline{G}_i(t), \overline{G}_i(t)]$ which describe the range of possible values of the cdf $G_i(x)$ characterizing each term t_i.

For $c_i > 0$, the inequality $c_i \cdot \Delta x_i \le t$ is equivalent to $\Delta x_i \le \dfrac{t}{c_i}$, so,

$$G_i(t) = \text{Prob}(t_i \le t) = \text{Prob}\left(\Delta x_i \le \frac{t}{c_i}\right) = F_i\left(\frac{t}{c_i}\right).$$

In this case:

- The smallest possible value of $G_i(t)$ corresponding to the smallest possible values $\underline{F}_i$ of F_i.
- The largest possible value of $G_i(t)$ corresponding to the largest possible values $\overline{F}_i$ of F_i.

So, we have $\underline{G}_i(t) = \underline{F}_i(\frac{t}{c_i})$ and $\overline{G}_i(t) = \overline{F}_i(\frac{t}{c_i})$.

For $c_i < 0$, the inequality $c_i \cdot \Delta x_i \le t$ is equivalent to $\Delta x_i \ge \dfrac{t}{c_i}$, so

$$G_i(t) = \text{Prob}(t_i \le t) = 1 - \text{Prob}\left(\Delta x_i \ge \frac{t}{c_i}\right) = 1 - F_i\left(\frac{t}{c_i}\right).$$

In this case:

- The smallest possible value of $G_i(t)$ corresponding to the largest possible values $\overline{F}_i$ of F_i.
- The largest possible value of $G_i(t)$ corresponding to the smallest possible values $\underline{F}_i$ of F_i.

So, we have $\underline{G}_i(t) = 1 - \overline{F}_i(\frac{t}{c_i})$ and $\overline{G}_i(t) = 1 - \underline{F}_i(\frac{t}{c_i})$.

In general, the lower bound $\underline{F}(x)$ corresponds to the smallest possible probability of smaller values—and thus, to the largest possible probability of larger values. Similarly, the upper bound $\overline{F}(x)$ corresponds to the largest possible probability of smaller values. Thus:

- To find the lower bound $\underline{F}(x)$ corresponding to Δy, we must use probability distributions $G_i(\Delta x_i)$ for which the values t_i are the largest, i.e., the values $\underline{G}_i(t)$.
- Similarly, to find the upper bound $\overline{F}(x)$, we must use probability distributions $G_i(\Delta x_i)$ for which the values t_i are the smallest, i.e., the values $\overline{G}_i(t)$.

So, we arrive at the following algorithm.

Algorithm for Exact Computation of p-Box $[\underline{F}(x), \overline{F}(x)]$: General Idea

- First, we use numerical differentiation (4) to estimate the coefficients c_i.
- For each i, we can then compute the p-boxes $[\underline{G}_i(t), \overline{G}_i(t)]$ corresponding to $t_i \stackrel{\text{def}}{=} c_i \cdot \Delta x_i$ as follows:
  - When $c_i > 0$, we take $\underline{G}_i(t) = \underline{F}_i(\frac{t}{c_i})$ and $\overline{G}_i(t) = \overline{F}_i(\frac{t}{c_i})$.
  - When $c_i < 0$, we take $\underline{G}_i(t) = 1 - \overline{F}_i(\frac{t}{c_i})$ and $\overline{G}_i(t) = 1 - \underline{F}_i(\frac{t}{c_i})$.
- Then, to find the p-box $[\underline{F}(x), \overline{F}(x)]$ corresponding to the sum $\Delta y = \sum_{i=1}^{n} t_i + \Delta x_0$, we do the following:
  - To compute $\underline{F}(x)$, we apply the convolution formula:

$$\rho_c(x) = \int \rho_a(t) \cdot \rho_b(x - t)\, \mathrm{d}t,$$

    for the probability density of the sum $c = a + b$ to independent random variables with cdf's $\underline{G}_i(t)$
  - To compute $\overline{F}(x)$, we apply the same convolution formula to independent random variables with cdf's $\overline{G}_i(t)$.

To compute convolutions, we use the above algorithm based on fast Fourier transform.

Possible Use of the Central Limit Theorem Similar to the case when we know the exact non-Gaussian distributions, we can speed up computations if we know the list of non-small components. In this case, we know the sum $t_{k+1} + \cdots + t_n$ of small

components is normally distributed. Normal distribution can be described by the mean μ and standard deviation σ; ranges $[\underline{\mu}, \overline{\mu}]$ and $[\underline{\sigma}, \overline{\sigma}]$ for μ and σ can be found by using the same methods as in the case of smooth utility function.

In general, cdf for a normal distribution has the form $F(t) = F_0(\frac{t-\mu}{\sigma})$, where $F_0(t)$ is the cdf of the "standard" normal distribution, with mean 0 and standard deviation 1. The function $F_0(t)$ is increasing. Thus, if we know the bounds on μ and on σ:

- The smallest value of $F(t)$ is attained when μ and σ are the largest.
- The largest value of $F(t)$ is attained when μ and σ are the smallest.

In other words, $\underline{F}_{\rm sm}(x) = F_0(\frac{t-\overline{\mu}}{\overline{\sigma}})$ and $\overline{F}_{\rm sm}(x) = F_0(\frac{t-\underline{\mu}}{\underline{\sigma}})$.

The p-box for Δy can then be obtained by combining p-boxes corresponding to $t_1, \ldots, t_k, t_0$, and the above Gaussian p-box $[\underline{F}_{\rm sm}(x), \overline{F}_{\rm sm}(x)]$ for $\sum_{i=k+1}^{n} t_i$.

Acknowledgements This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721.

The author would like to thank all the participants of the International Conference on Applied Mathematics and Informatics ICAMI'2013 (San Andres Island, Colombia, November 24–28, 2013), especially Anibal Sosa, for valuable suggestions, and to the anonymous referees for their useful comments.

References

1. Cormen, T.H., et al.: Introduction to Algorithms. MIT Press, Cambridge (2009)
2. Ferson, S.: Risk Assessment with Uncertainty Numbers: RiskCalc. CRC, Boca Raton (2002)
3. Ferson, S., et al.: Experimental uncertainty estimation and statistics for data having interval uncertainty. Sandia National Laboratories, Report SAND2007-0939, May 2007. http://www.ramas.com/intstats.pdf (2007)
4. Fishburn, P.C.: Nonlinear Preference and Utility Theory. John Hopkins, Baltimore (1988)
5. Griewank, A.: Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation. SIAM, Philadelphia (2000)
6. Jaulin, L., et al.: Applied Interval Analysis. Springer, London (2001)
7. Jaynes, E.T., Bretthorst, G.L.: Probability Theory: The Logic of Science. Cambridge University Press, Cambridge (2003)
8. Kreinovich, V.: Interval computations and interval-related statistical techniques: tools for estimating uncertainty of the results of data processing and indirect measurements, In: Pavese, F., Forbes, A.B. (eds.) Data Modeling for Metrology and Testing in Measurement Science, pp. 117–145. Birkhauser-Springer, Boston (2009)
9. Kreinovich, V., Ferson, S.: A new Cauchy-based black-box technique for uncertainty in risk analysis. Reliab. Eng. Syst. Safety **85**(1–3), 267–279 (2004)
10. Luce, R.D., Raiffa, R.: Games and Decisions: Introduction and Critical Survey, Dover, New York (1989)
11. Moore, R.E., Kearfott, R.B., Cloud, M.J.: Introduction to Interval Analysis. SIAM, Philadelphia (2009)
12. Nguyen, H.T., et al.: Computing Statistics Under Interval and Fuzzy Uncertainty. Springer, Berlin (2012)
13. Novitskii, P.V., Zograph, I.A.: Estimating the Measurement Errors. Energoatomizdat, Leningrad (1991) (in Russian)
14. Orlov, A.I.: How often are the observations normal? Ind. Lab. **57**(7), 770–772 (1991)

15. Rabinovich, S.G.: Measurement Errors and Uncertainty. Theory and Practice. Springer, Berlin (2005)
16. Raiffa, H.: Decision Analysis, McGraw-Hill, Columbus (1997)
17. Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures. Chapman & Hall/CRC, Boca Raton (2011)
18. Werbos, P.J.: Beyond regression: New tools for prediction and analysis in the behavioral sciences. PhD thesis, Harvard University (1974)
19. Werbos, P.J.: The Roots of Backpropagation. From Ordered Derivatives to Neural Networks and Political Forecasting. Wiley, New York (1994)

A Unified Approach to Piecewise Linear Hopf and Hopf-Pitchfork Bifurcations

Enrique Ponce, Javier Ros and Elísabet Vela

Abstract We propose for symmetric three-dimensional piecewise linear systems with three zones a unified approach to analyze both Hopf and Hopf-pitchfork bifurcations. For the equilibrium at the origin, the crossing of a complex eigenvalue pair through the imaginary axis of complex plane, with the possible simultaneous crossing of a real eigenvalue, is considered. Some results related to the bifurcation of limit cycles are provided, and an illustrative example is included.

Keywords Piecewise linear systems · Hopf-pitchfork bifurcation · Limit cycles

1 Introduction

The class of piecewise linear differential (PWL) systems is very important within the realm of nonlinear dynamical systems. In fact, this kind of models is frequent in applications from electronic engineering and nonlinear control systems, where piecewise linear models cannot be considered as idealized ones, see [5] and references therein; they are used in mathematical biology as well, see [14, 15], where they constitute approximate models. On the other hand, since piecewise linear characteristics can be considered as the uniform limit of smooth nonlinearities, the global dynamics of smooth models has been sometimes approximated by piecewise linear models and viceversa, see [9, 16], obtaining a good qualitative agreement between the two modeling approaches. In practice, nonlinear characteristics use to have a saturated part, which is difficult to be approximated by polynomial functions but

E. Ponce (✉) · J. Ros · E. Vela
Depto. Matemática Aplicada II, Camino Descubrimientos,
E.T.S. Ingeniería, 41092 Sevilla, Spain
e-mail: eponcem@us.es

J. Ros
e-mail: javieros@us.es

E. Vela
e-mail: elivela@us.es

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_12

suitable to be modeled by linear pieces, leading to what we could call a "global linearization."

The pioneering investigation of piecewise linear systems in a rigorous way was due to Andronov et al. [1]. Their book *Theory of Oscillations* remains nowadays an obligated reference, still being a source of ideas. More recently, the analysis of piecewise linear systems received growing attention due to the interest on PWL chaotic systems, see for instance [10] and references therein.

While bifurcation theory is rather well established for smooth vector fields, the nonsmooth case and the PWL case in particular are nowadays an area of active research, see [2, 5, 8, 13] among others. It is in this context, where we want to advance in the theory; more precisely, we consider three-dimensional symmetric continuous piecewise linear systems with three zones paying special attention to the bifurcation of limit cycles. Limit cycles are isolated periodic orbits that, after equilibrium points, correspond with the most important solutions of dynamical systems. Their determination is a difficult task, so that new results in this direction are of great relevance in real applications, see [7]. In the case of piecewise smooth systems, there are very few known results about, see again [5].

We study the analogous situation to Hopf bifurcation in smooth systems, allowing that such a bifurcation be simultaneous with a pitchfork bifurcation, and proposing a unified approach for both settings.

To be more specific, we consider the following family of PWL systems written in the Luré form:

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}) = A_R\mathbf{x} + \mathbf{b}\,\mathrm{sat}(x), \tag{1}$$

where $\mathbf{x} = (x, y, z)^T \in \mathbb{R}^3$, the saturation function is given by

$$\mathrm{sat}(u) = \begin{cases} 1 & \text{if } u > 1, \\ u & \text{if } |u| \leq 1, \\ -1 & \text{if } u < -1, \end{cases}$$

the dot represents derivative with respect to the time τ. Under generic assumptions, see [4], there is no loss of generality in assuming that

$$A_R = \begin{pmatrix} t & -1 & 0 \\ m & 0 & -1 \\ d & 0 & 0 \end{pmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} T - t \\ M - m \\ D - d \end{pmatrix}, \tag{2}$$

where the coefficients t, m, d and T, M, D are the linear invariants (trace, sum of principal minors and determinant) of the matrices A_R and A_C, respectively. Note that in the region with $|x| \leqslant 1$, it becomes the homogeneous system:

$$\dot{\mathbf{x}}(\tau) = A_C\mathbf{x}(\tau) = \begin{pmatrix} T & -1 & 0 \\ M & 0 & -1 \\ D & 0 & 0 \end{pmatrix} \mathbf{x}(\tau). \tag{3}$$

We observe that $A_C = A_R + \mathbf{b}\mathbf{e}_1^T$, where $\mathbf{e}_1 = (1, 0, 0)^T$, and that the considered family of systems is in the generalized Liénard form. Thus, under generic conditions for every system of the form (1) after some change of variables, we can get the matrices in the form given in (2) and (3).

2 Statements of Main Results

In this work, we consider a more general structure of eigenvalues than the one appeared in [6] and [12], which includes both the piecewise linear analogue of Hopf bifurcation and the one of Hopf-pitchfork bifurcation, also called Hopf-zero bifurcation. Let us take ε as the main bifurcation parameter and assume the following expressions for the eigenvalues of the linear part at the origin A_C,

$$\lambda(\varepsilon),\ \sigma(\varepsilon) \pm \mathrm{i}\omega(\varepsilon),$$

where,

$$\begin{aligned}
\lambda(\varepsilon) &= \lambda_0 + \lambda_1\varepsilon + O(\varepsilon^2),\\
\sigma(\varepsilon) &= \sigma_1\varepsilon + O(\varepsilon^2),\\
\omega(\varepsilon) &= \omega_0 + \omega_1\varepsilon + O(\varepsilon^2),
\end{aligned}$$

with $\omega_0 > 0$. Here we will assume that both σ_1 and λ_1 do not vanish; these vanishing cases, much more involved, are left for future work. Clearly, when ε passes from negative values to positive ones, a pair of complex eigenvalues crosses the imaginary axis, which is the usual requirement for a Hopf bifurcation.

With this choice of the eigenvalues, the trace, principal minor of order two and determinant must have the following form:

$$\begin{aligned}
T(\varepsilon) &= \lambda(\varepsilon) + 2\sigma(\varepsilon),\\
M(\varepsilon) &= \sigma^2(\varepsilon) + \omega^2(\varepsilon) + 2\lambda(\varepsilon)\sigma(\varepsilon),\\
D(\varepsilon) &= \lambda(\varepsilon)(\sigma^2(\varepsilon) + \omega^2(\varepsilon)),
\end{aligned} \tag{4}$$

where

$$\begin{aligned}
T_0 &= T(0) = \lambda_0,\\
M_0 &= M(0) = \omega_0^2,\\
D_0 &= D(0) = \lambda_0\omega_0^2,
\end{aligned}$$

so that $D_0 - M_0T_0 = 0$.

When $\lambda_0 \neq 0$, by moving the parameter ε through zero, we reproduce the piecewise linear Hopf or focus-center-limit cycle bifurcation analyzed in [6], by

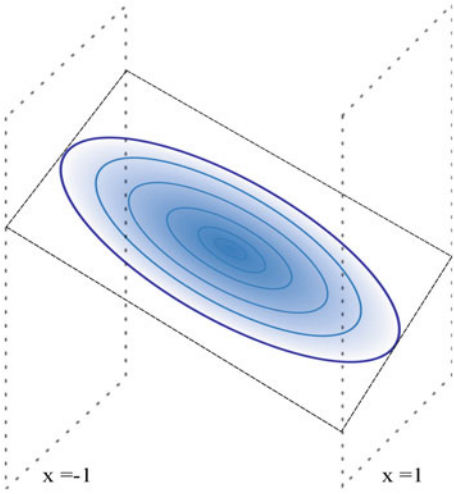


Fig. 1 Structure of periodic orbits in the central zone for $\lambda_0 \neq 0$ and $\varepsilon = 0$; the periodic orbits determine a center configuration located at the focal plane $\lambda_0^2 x - \lambda_0 y + z = 0$

considering M and D constant; thus, we require here less restricted assumptions. Furthermore, when $\lambda_0 = 0$, as it is assumed $\lambda_1 \neq 0$, by moving ε we have the simultaneous crossing of a zero eigenvalue and a complex pair, a situation analogous to the Hopf-zero bifurcation in smooth systems. Thus, our analysis unifies the study done in [12], with the one in [6] allowing also to consider new degenerate cases, not yet analyzed.

In particular, the current formulation allows to pave the way for analysing the situations $\lambda_0 = \lambda_1 = 0$, or the case $\omega_0 = 0$, where we should get a more degenerate case for the focus-center-limit cycle bifurcation or a triple-zero case, respectively; the analysis of such degenerate cases is lacking and far from being solved.

To start with, we emphasize in the next result an invariant property of systems (1)–(2), whose proof is direct and will be omitted.

Proposition 1 *Systems (1)–(2) are invariant under the following transformation:*

$$(x, y, z, \tau, t, m, d, \varepsilon) \longrightarrow (x, -y, z, -\tau, -t, m, -d, -\varepsilon).$$

This symmetry property is useful for simplifying the analysis of the family under consideration.

First, we consider the bifurcation for $\varepsilon = 0$ under the hypothesis $\lambda_0 \neq 0$, $\omega_0 > 0$, and $\sigma_1 \neq 0$. Under these conditions, it is very easy to show that in the focal plane $\lambda_0^2 x - \lambda_0 y + z = 0$ there exists a center configuration when $\varepsilon = 0$, see Fig. 1. From the periodic orbit of this center that is tangent to the planes $x = \pm 1$, we can assure the bifurcation of one limit cycle as follows.

Theorem 1 *Let us consider systems (1)–(2) under condition (4) where it is assumed $\omega_0 > 0$, $\lambda_0 \neq 0$, $\sigma_1 \neq 0$. Thus, we have $MT - D = 0$ for $\varepsilon = 0$ with $M_0 > 0$. Assuming*

$$\delta = d - t\omega_0^2 + \lambda_0(\omega_0^2 - m) \neq 0,$$

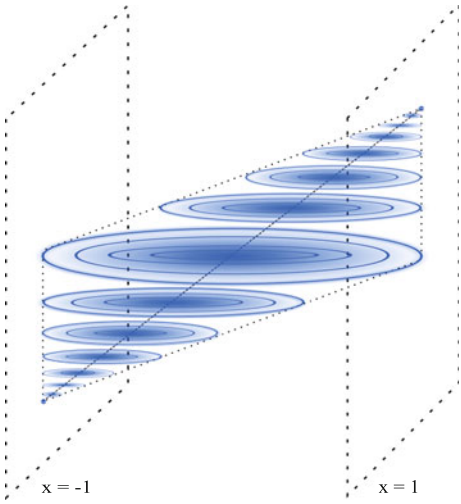


Fig. 2 Structure of periodic orbits in the central zone for $\lambda_0 = 0$ and $\varepsilon = 0$. The two solid cones are completely foliated by periodic orbits surrounding the segment of equilibrium points $\{(x, 0, x\omega_0^2)^{\mathrm{T}} : |x| \leqslant 1\}$

for $\varepsilon = 0$, the system undergoes a focus-center-limit cycle bifurcation, that is, from the linear center configuration in the central zone, which exists for $\varepsilon = 0$, one limit cycle appears for $\delta\sigma_1\varepsilon > 0$ and $|\varepsilon|$ sufficiently small.

The period P and the amplitude A (measured as the maximum of $|x|$) of the periodic orbit are analytic functions at 0, in the variable $\varepsilon^{1/3}$, namely

$$P = \frac{2\pi}{\omega_0} + \frac{2\pi}{\omega_0^3\delta}\left[\lambda_0\sigma_1(t\omega_0^2 - d) + \omega_0^2\sigma_1(\omega_0^2 - m) - \omega_0\omega_1\delta\right]\varepsilon + O(\varepsilon^{4/3}),$$

$$A = 1 + \frac{1}{2}\left(\frac{3\pi}{2}\frac{\sigma_1(\lambda_0^2 + \omega_0^2)}{\delta}\right)^{2/3}\varepsilon^{2/3} + O(\varepsilon^{4/3}).$$

In particular, if $\lambda_0 < 0$ and $\delta > 0$, then the limit cycle bifurcates for $\sigma_1\varepsilon > 0$ and is orbitally asymptotically stable.

For sake of brevity, the proof of Theorem 1, being rather similar to the one given in [6], will be omitted.

The case $\lambda_0 = 0$ with $\lambda_1 \neq 0$ would lead to a richer structure of periodic orbits when $\varepsilon = 0$, see Fig. 2, and then the following assertions about possible equilibrium points of the family are relevant.

Proposition 2 *For systems (1) and (2) under condition (4) with $\omega_0 > 0$, $\lambda_0 = 0$, and $\lambda_1 \neq 0$, the following statements hold:*

(a) If $d\lambda_1\varepsilon > 0$, then the unique equilibrium point is the origin.
(b) If $d\lambda_1\varepsilon < 0$, then the equilibria are the origin and the two points

$$\mathbf{x}_\varepsilon^+ = \frac{1}{d}\left(d - D(\varepsilon),\, dT(\varepsilon) - tD(\varepsilon),\, dM(\varepsilon) - mD(\varepsilon)\right)^{\mathrm{T}}, \mathbf{x}_\varepsilon^- = -\mathbf{x}_\varepsilon^+.$$

(c) If $\varepsilon = 0$, then all the points of the segment

$$\{(x, y, z)^{\mathrm{T}} \in \mathbb{R}^3 : (x, y, z)^{\mathrm{T}} = \mu(1, 0, \omega_0^2)^{\mathrm{T}}, |\mu| \leqslant 1\}$$

are equilibria for the system. If furthermore $d \neq 0$, the above segment captures all the equilibrium points.

For a proof of Proposition 2, see the similar result in [12]. From the above statement (c), we see that at $\varepsilon = 0$ systems (1)–(2) have a *degenerate pitchfork bifurcation*. Note that for $d\lambda_1\varepsilon > 0$, the points $\mathbf{x}_\varepsilon^{\pm}$ are vanishing points for the vector field corresponding to $|x| > 1$ but they are out of their corresponding zones. They do not constitute real equilibria, although they still organize the dynamics of external regions. This type of equilibrium is usually called a *virtual equilibrium point*.

Our first result when $\lambda_0 = 0$ concerns the possible bifurcation of symmetrical periodic orbits using the three zones. We note that if $\lambda_0 = 0$, we now have $\delta = d - t\omega_0^2$, which characterizes the criticality of the bifurcation, in a similar way to what happens in the cases considered in [3] and [6].

Theorem 2 *Let us consider systems (1)–(2) under condition (4) where it is assumed $\lambda_0 = 0$, $\lambda_1 \neq 0$, $\omega_0 > 0$, and $\delta = d - t\omega_0^2 \neq 0$. For $\varepsilon = 0$, the systems (1)–(2) undergo a trizonal limit cycle bifurcation, that is, from the configuration of periodic orbits that exists in the central zone for $\varepsilon = 0$, one limit cycle appears for $\delta\sigma_1\varepsilon > 0$ and $|\varepsilon|$ sufficiently small. It is symmetric with respect to the origin and bifurcates from the ellipse $\{(x, y, z)^{\mathrm{T}} \in \mathbb{R}^3 : \omega_0^2 x^2 + y^2 = \omega_0^2, z = 0\}$. This limit cycle has period:*

$$P = \frac{2\pi}{\omega_0} + 2\pi \frac{\omega_0\sigma_1(\omega_0^2 - m) - \omega_1\delta}{\omega_0^2\delta}\varepsilon + O(\varepsilon^{4/3}),$$

and its amplitude in x measured as $\max\{x\} - \min\{x\}$ is

$$A = 1 + \frac{1}{2}\left(\frac{3\pi}{2}\frac{\omega_0^2\sigma_1}{\delta}\right)^{2/3}\varepsilon^{2/3} + O(\varepsilon^{4/3}).$$

Furthermore, the bifurcating limit cycle is stable if and only if $t < 0$, $d < 0$, and $\delta > 0$.

By using Proposition 1, we could add a new assertion saying that the bifurcating limit cycle is completely unstable (the two characteristic exponents have positive real part) if and only if $t > 0$, $d > 0$, and $\delta < 0$.

Our last result, which also assumes $\lambda_0 = 0$, gives account of the bifurcation of a symmetrical pair of limit cycles that only use two linearity zones. This result requires extra assumptions, but when they are fulfilled allow us to assure the simultaneous bifurcation of three limit cycles.

Theorem 3 *Let us consider system 1–2 under conditions (4) where it is assumed $\delta = d - t\omega_0^2 \neq 0$, $\lambda_0 = 0$, $\lambda_1 \neq 0$, and $\omega_0 > 0$ fixed. Thus, if we have $\sigma_1 \neq 0$,*

$d\sigma_1 - \lambda_1\delta \neq 0$, and

$$0 < \widehat{z} = \frac{d\sigma_1\omega_0^2}{d\sigma_1 - \lambda_1\delta} < \omega_0^2$$

and fixed, a bifurcation takes place for the critical value $\varepsilon = 0$. Thereby, a symmetrical pair of limit cycles appears when $\delta\sigma_1\varepsilon > 0$ and $|\varepsilon|$ is sufficiently small. They are stable if and only if $t < 0$ and $\lambda_1\sigma_1 < 0$, or $t = 0$ and $d\sigma_1(\lambda_1 + 2\sigma_1) < 0$. Their common period is

$$P = \frac{2\pi}{\omega_0} + \frac{2\pi\left[\omega_0\sigma_1(\omega_0^2 - m) - \omega_1\delta\right]}{\omega_0^2\delta}\varepsilon + O(\varepsilon^{5/3}),$$

and their common amplitude in x measured as $\max\{x\} - \min\{x\}$ is

$$A = \frac{2\lambda_1\delta}{\lambda_1\delta - d\sigma_1} - \frac{2(3\pi)^{2/3}}{5}\left(\frac{\sigma_1}{\omega_0\delta}\right)^{2/3}\frac{d\lambda_1\sigma_1\omega_0^2[2t(\omega_0^2 - m) - 3\delta]}{(\lambda_1\delta - d\sigma_1)^2}\varepsilon^{2/3} + O(\varepsilon).$$

For a proof of both Theorems 2 and 3, one can follow the procedure given in [12]. The results included here are similar to the ones in such a quoted paper, but we emphasize that here the number of auxiliary fixed parameters describing the eigenvalue configuration has been increased from two (ρ and ω) to five (λ_0, λ_1, σ_1, ω_0, and ω_1), allowing a unified approach that encompasses both referred bifurcations, including cases not analyzed in [6] nor in [12] and paving the way for future analysis of more degenerate situations.

3 An Illustrative Example: An Electronic Oscillator

In this section, as an illustrative example of the usefulness of above results, we consider an extended Bonhoeffer–van der Pol (BVP) electronic oscillator, which consists of two capacitors, an inductor, a linear resistor, and a nonlinear conductance, as shown in Fig. 3.

To obtain more information about this circuit, see [11], where a smooth nonlinearity is assumed for the conductance and a rich variety of dynamical behaviors is found. The circuit equations can be written as:

$$C\frac{dv_1}{dt} = -i - g(v_1), \quad C\frac{dv_2}{dt} = i - \frac{v_2}{r}, \quad L\frac{di}{dt} = v_1 - v_2,$$

where v_1 and v_2 are the voltages across the capacitors, the symbol i stands for the current through the inductance L, and the $v-i$ characteristics of the nonlinear resistor is written as $g(v) = -av - b\,\mathrm{sat}\,(cv)$, where $a, b, c > 0$. Note that, here we adopt a PWL version of the nonlinearity considered in [11].

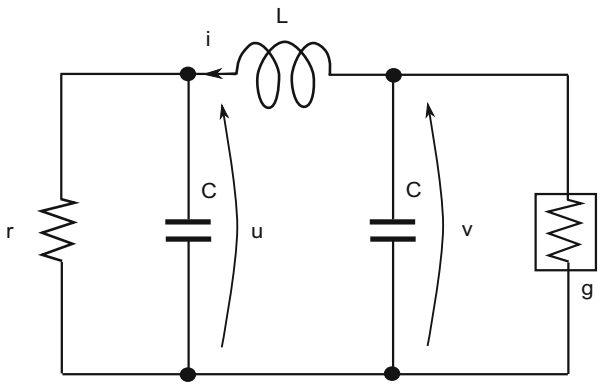


Fig. 3 The extended Bonhoeffer–van der Pol (BVP) oscillator proposed in [11]

After some standard manipulations, the normalized equations of the extended BVP oscillator become

$$\begin{cases} \dot{x} = & -z + \alpha x + \text{sat}\,(\beta x), \\ \dot{y} = & z - \gamma y, \\ \dot{z} = & x - y, \end{cases}$$

where the dot represents derivative with respect to the new time τ, and

$$\tau = \frac{1}{\sqrt{LC}}t, \quad \alpha = a\sqrt{\frac{L}{C}}, \quad \beta = bc\sqrt{\frac{L}{C}}, \quad \gamma = \frac{1}{r}\sqrt{\frac{L}{C}},$$

$$x = \frac{v_1}{b}\sqrt{\frac{C}{L}}, \quad y = \frac{v_2}{b}\sqrt{\frac{C}{L}}, \quad z = \frac{i}{b}.$$

Making now the change of variables $X = \beta x$, we obtain the system in its Luré form,

$$\dot{\mathbf{x}} = \begin{pmatrix} \alpha & 0 & -\beta \\ 0 & -\gamma & 1 \\ 1/\beta & -1 & 0 \end{pmatrix} \mathbf{x} + \begin{pmatrix} \beta \\ 0 \\ 0 \end{pmatrix} \text{sat}\,(\mathbf{e}_1^T \mathbf{x}), \tag{5}$$

and we will rename X as x in the sequel, for convenience. Then, it can be written in the form 1–2, and so we will try to apply Theorems 1, 2, and 3 under the corresponding assumptions. Effectively, with a linear change of variables given by the matrix:

$$P = \frac{1}{\beta}\begin{pmatrix} \beta & 0 & 0 \\ \gamma^2 - 1 & \gamma & 1 \\ \gamma & 1 & 0 \end{pmatrix},$$

we can write system (5) in its Liénard form as

$$\dot{\mathbf{x}} = \begin{pmatrix} \alpha-\gamma & -1 & 0 \\ 2-\alpha\gamma & 0 & -1 \\ \alpha-\gamma & 0 & 0 \end{pmatrix} \mathbf{x} + \begin{pmatrix} \beta \\ -\beta\gamma \\ \beta \end{pmatrix} \text{sat}(x), \tag{6}$$

where now the trace, the sum of second-order principal minors, and the determinant in the different zones are evident, namely

$$\begin{aligned} T &= \alpha+\beta-\gamma, & t &= \alpha-\gamma, \\ M &= 2-\gamma(\alpha+\beta), & m &= 2-\alpha\gamma, \\ D &= \alpha+\beta-\gamma, & d &= \alpha-\gamma. \end{aligned} \tag{7}$$

From (7), we observe that T and D are identically equal, what implies that an extra condition for eigenvalues must be fulfilled. Thus, taking into account the structure of T and D given in (4), we must impose for all values of ε,

$$T(\varepsilon) - D(\varepsilon) = \lambda_0(1-\omega_0^2) + (\lambda_1 - \lambda_1\omega_0^2 + 2\sigma_1 - 2\lambda_0\omega_1)\varepsilon + O(\varepsilon^2) = 0.$$

We will take γ as the only bifurcation parameter, keeping α and β fixed. In looking for the bifurcations analyzed in Sect. 2 to take place at $\varepsilon = 0$, we need first $\lambda_0(1-\omega_0^2) = 0$. If we assume $\lambda_0 \neq 0$, then we must conclude the two conditions

$$\omega_0 = 1 \text{ and } \sigma_1 = \lambda_0\omega_1.$$

Consequently, $M(0) = \omega_0^2 = 1$, and we get for the bifurcation parameter $\gamma(\varepsilon)$ the condition $\gamma(0) = \gamma_0$, with

$$0 < \gamma_0 = \frac{1}{\alpha+\beta} \neq 1, \text{ so that } \lambda_0 = \frac{1-\gamma_0^2}{\gamma_0}.$$

To apply Theorem 1, we compute for $\varepsilon = 0$,

$$m = 2 - \frac{\alpha}{\alpha+\beta}, \text{ and so } \delta = \lambda_0(1-m) \neq 0.$$

From (7), writing $\gamma = \gamma_0 + \varepsilon$, we also obtain

$$\lambda_1 = -\gamma_0^4 \frac{1+\gamma_0^2}{1+\gamma_0^6}, \quad \sigma_1 = \frac{\gamma_0^4 - 1}{2(1+\gamma_0^6)}, \text{ and } \omega_1 = -\frac{\gamma_0(1+\gamma_0^2)}{2(1+\gamma_0^6)}.$$

Thus, the following result is a direct consequence of Theorem 1.

Proposition 3 *Let us consider system (5) or equivalently system (6) with $\alpha > 0$, $\beta > 0$, and $\gamma_0 = 1/(\alpha+\beta)$ fixed. For $\gamma = \gamma_0$, the system undergoes a focus-center-limit cycle bifurcation, that is, from the linear center configuration in the central zone, which exists for $\gamma = \gamma_0$, one limit cycle appears for $\gamma - \gamma_0 > 0$ and sufficiently*

small. In particular, if $\alpha + \beta < 1$, *then* $\gamma_0 > 1$ *and the bifurcating limit cycle is asymptotically stable.*

On the other hand, if we assume $0 < \omega_0 \neq 1$, then to get the consistency between (4) and (7), we need $\lambda_0 = 0$, and therefore,

$$\sigma_1 = \frac{\lambda_1}{2}(\omega_0^2 - 1),$$

getting for the bifurcation parameter $\gamma(\varepsilon)$ the condition $\gamma(0) = \gamma_0$, with

$$0 < \gamma_0 = \alpha + \beta < \sqrt{2},$$

with the additional requirement that $\alpha + \beta \neq 1$; otherwise, $\omega_0 = 1$ and $\sigma_1 = 0$, precluding the use of both Theorems 2 and 3. Note that $\omega_0^2 = 2 - \gamma_0^2$ and so when $\gamma_0 < 1$ we have $\omega_0 > 1$ and vice versa.

Using 7, we obtain for $\varepsilon = 0$ that $t = d = -\beta$ and $\delta = \beta(\omega_0^2 - 1) \neq 0$. Writing $\gamma = \gamma_0 + \varepsilon$, we also obtain

$$\lambda_1 = -\frac{1}{\omega_0^2}, \quad \sigma_1 = \frac{1 - \omega_0^2}{2\omega_0^2}, \text{ and } \omega_1 = -\frac{\gamma_0}{2\omega_0}.$$

We note that in Theorem 3,

$$\widehat{z} = \frac{d\sigma_1\omega_0^2}{d\sigma_1 - \lambda_1\delta} = \frac{\omega_0^2}{3} \in \left(0, \omega_0^2\right).$$

Thus, from Theorems 2 and 3, we get the following result.

Proposition 4 *Considering system (5) or equivalently system (6) with* $\alpha > 0$, $\beta > 0$ *and* $1 \neq \gamma_0 = \alpha + \beta < \sqrt{2}$ *and fixed, the following statements hold:*

(a) For $\gamma > \gamma_0$, *the origin is the only equilibrium of the system. Furthermore, if* $\gamma\gamma_0 < 1$, *then the origin is asymptotically stable.*

(b) For $\gamma = \gamma_0$, *the system undergoes a PWL analogue of the Hopf-zero bifurcation; from the periodic set existing at such critical situation, for* $\gamma - \gamma_0 < 0$ *and sufficiently small in absolute value, the bifurcation leads to the simultaneous appearance of three limit cycles (one trizonal and two bizonal ones) along with two additional equilibrium points.*
Furthermore, if $\gamma_0 < 1$ $(1 < \gamma_0 < \sqrt{2})$, *then the bifurcating trizonal limit cycle is stable (unstable) while the bifurcating bizonal limit cycles are unstable (stable). The bifurcating equilibrium points are stable whenever* $\gamma_0 < 1$ *and, in the case* $1 < \gamma_0 < \sqrt{2}$, *when* $\gamma_0 < 1/\alpha$.

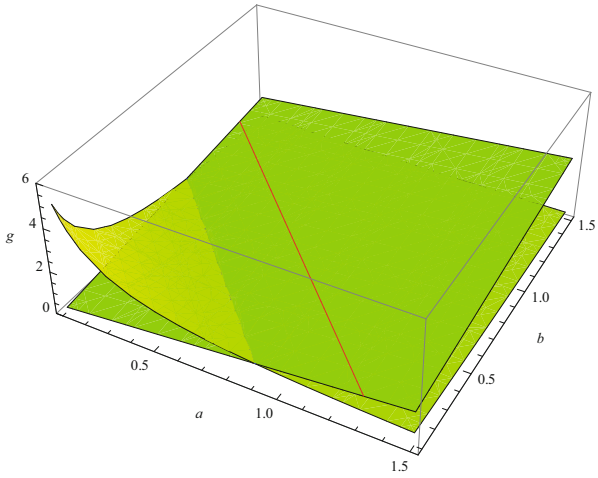


Fig. 4 Partial bifurcation set of system (5), showing the two main bifurcation surfaces corresponding to the piecewise linear Hopf and Hopf-pitchfork bifurcations, namely the surface $\gamma = 1/(\alpha + \beta)$ and the plane $\gamma = \alpha + \beta$. It is also shown the red straight-line $\gamma = \alpha + \beta = \sqrt{2}$, where $T = M = D = 0$

In Fig. 4, we show the two main bifurcation surfaces corresponding to the piecewise linear Hopf and Hopf-pitchfork bifurcations, namely the surface $\gamma = 1/(\alpha+\beta)$ and the plane $\gamma = \alpha + \beta$. It is also shown as the straight-line $\gamma = \alpha + \beta = \sqrt{2}$, where $T = M = D = 0$ and so, a triple-zero bifurcation is involved. The analysis of such a bifurcation is left for future work.

Acknowledgment Authors are partially supported by the *Ministerio de Ciencia y Tecnología, Plan Nacional I+D+I*, in the frame of projects MTM2010-20907 and MTM2012-31821, and by the *Consejería de Economía, Innovación, Ciencia y Empleo de la Junta de Andalucía* under grant FQM-1658.

References

1. Andronov, A.A., Vitt, A.A., Khaikin, S.E.: Theory of Oscillators. Pergamon, Oxford (1966)
2. Biemond, J.J.B., van der Wouw, N., Nijmeier, H.: Nonsmooth bifurcations of equilibria in planar continuous systems. Nonlinear Anal. **4**, 451–475 (2010)
3. Carmona, V., Freire, E., Ponce, E., Ros, J., Torres, F.: Limit cycle bifurcation in 3D continuous piecewise linear systems with two zones. Application to Chua's circuit. Int. J. Bifurcation Chaos **15**, 2469–2484 (2005)
4. Carmona, V., Freire, E., Ponce, E., Torres, F.: On simplifying and classifying piecewise-linear systems. IEEE Trans. Circuits Syst. **49**, 609–620 (2002)
5. Di Bernardo, M., Budd, C., Champneys, A.R., Kowalczyk, P.: Piecewise-smooth dynamical systems: Theory and applications. Applied mathematical sciences, vol. 163. Springer (2007)
6. Freire, E., Ponce, E., Ros, J.: The focus-center-limit cycle bifurcation in symmetric 3D piecewise linear systems. SIAM J. Appl. Math. **65**, 1933–1951 (2005)
7. Kuznetsov, Y.A.: Elements of applied bifurcation theory. Applied mathematical sciences, vol. 112, 3rd edn. Springer, New York (2004)
8. Leine, R.I., Nijmeijer, H.: Dynamics and bifurcations of non-smooth mechanical systems. Lecture Notes in Applied and Computational Mechanics. Springer, Berlin (2004)

9. Lipton, J.M., Dabke, K.P.: Softening the nonlinearity in Chua's circuit. Int. J. Bifurcation Chaos **6**(1), 179–183 (1996)
10. Madan, R.N.: Chua's circuit: A paradigm for chaos. World scientific series on nonlinear science. World Scientific, Singapore (1993)
11. Nishiuchi, Y., Ueta, T., Kawakami, H.: Stable torus and its bifurcation phenomena in a simple three-dimensional autonomous circuit. Chaos Soliton Fract. **27**, 941–951 (2006)
12. Ponce, E., Ros, J., Vela, E.: Unfolding the fold-Hopf bifurcation in piecewise linear continuous differential systems with symmetry. Physica D **250**, 34–46 (2013)
13. Simpson, D.J.W.: Bifurcations in piecewise-smooth continuous systems. World Scientific series on nonlinear science: A. World Scientific Publishing Company, Inc., Singapore (2010)
14. Thul, R., Coombes, S.: Understanding cardiac alternans: A piecewise linear modeling framework. Chaos **20**, 045102-1–045102-13 (2010)
15. Tonnelier, A., Gerstner, W.: Piecewise linear differential equations and integrate-and-fire neurons: Insights from two-dimensional membrane models. Phys. Rev. E **67**, 21908 (2003)
16. Wolf, H., Kodvanj, J., Bjelovučić-Kopilović, S.: Effect of smoothing piecewise-linear oscillators on their stability predictions. J. Sound Vib. **270**, 917–932 (2004)

Optimal Decision Making for Breast Cancer Treatment in the Presence of Cancer Regression and Type II Error in Mammography Results

Sergio A. Vargas, Shengfan Zhang and Raha Akhavan-Tabatabaei

Abstract Breast cancer is the leading cause of cancer death among women worldwide. While breast cancer-screening policies have been widely studied in order to achieve early detection, not much research has been done to optimize treatment decisions once a screening policy is established. In this chapter, we propose a dynamic decision model to determine optimal breast cancer treatment decisions that consider both the impact of overtreatment and the potential delay in cancer detection; these two failures are caused by spontaneous cancer regression and type II error in mammography results, respectively. We measure the impact of medical treatment by means of quality-adjusted life years (QALYs) and our goal is to maximize this metric for a given patient.

Keywords Breast cancer · Screening policies · Dynamic treatment decisions · Cancer regression · Mammography · Markov decision processes · QALY

1 Introduction

Breast cancer is often defined as an uncontrolled growth of breast cells caused by a genetic abnormality. In 2011, the American Cancer Society (ACS) estimated more than 450,000 deaths caused by breast cancer and more than 1,000,000 new cases worldwide [16]. The same year, according to the ACS, the lifetime risk of developing invasive female breast cancer was about 12 %.

S. Zhang (✉)
University of Arkansas, Fayetteville, AR 72701, USA
e-mail: shengfan@uark.edu

S. A. Vargas
Universidad de los Andes, Cra 1 A No. 18-10, Bogotá, Colombia
e-mail: sa.vargas61@uniandes.edu.co

R. Akhavan-Tabatabaei
Universidad de los Andes, Cra 1 A No. 18-10, Bogotá, Colombia
e-mail: r.akhavan@uniandes.edu.co

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_13

Mammography is currently considered to be the most effective technology for breast cancer screening [23, 33]. A mammogram is an X-ray image to examine female breast. The benefits of mammography include early detection of breast cancer as it can identify problems before any symptoms (e.g., lumps) appear. There have been randomized clinical trials indicating that mammography may reduce breast cancer mortality by at least 24 % [13, 20]. However, there are two types of risk that need to be considered when performing mammography. Similar to other binary tests, mammography has two statistical measures of performance, sensitivity, and specificity. Sensitivity is the probability of detecting breast cancer when it is truly present while specificity is the probability of correctly identifying a patient as normal when no cancer exists [15]. The possible failures generated by specificity and sensitivity have raised the need to take into account this fact when developing optimal mammography screening policies for various populations [2, 24, 25].

In most screening and treatment decision models, breast cancer is typically modeled as a progressive disease, under the assumption that cancer does not disappear in the absence of treatment. For example, the Markov chain model proposed by Chen et al. [10] is often presented to describe the natural history of breast cancer, only allowing an early state of cancer to transit to a more advanced cancer state, or to an absorbing death state. However, there has been medical evidence suggesting that at an early stage, breast cancer may actually spontaneously regress without treatment [22]. While there has been a lot of debate in the medical community regarding cancer regression, there has been limited research about the consequences of considering this medical fact when determining treatment policies.

Schaefer et al. [32] discussed the fact that medical treatment decisions are often sequential and uncertain. Therefore, Markov decision processes (MDPs) are an adequate operations research tool to tackle this problem. They pointed out the advantages of MDPs when modeling and solving problems where stochasticity is involved in dynamic decisions such as the ones taken when treating a patient. Among those advantages, the authors mentioned the flexibility that allows a simple representation of future states and possible transitions that may occur until a patient dies. The authors also note that rather than evaluating a decision tree based on a one-time decision (as is often the case in traditional decision trees and Markov models), MDPs allow the "do-nothing" option in each time period and consider the "do-something" option at any later decision epoch. Finally, one of the most important advantages of an MDP approach for medical treatment is that the goal of this technique is to provide an optimal policy, which is a decision strategy to optimize a particular criterion such as maximizing a total discounted reward and it guarantees that no better policy exists.

Zhang and Ivy [41] proposed a finite-horizon MDP model to establish an optimal treatment policy in the presence of breast cancer regression. Their model assumes perfect information in screening results and fixes ACS recommendations reported in [35] as the screening policy upon which treatment decisions are made. The objective of their model is to minimize the loss of quality-adjusted life years (QALYs) due to overtreatment. Finally, their results showed a significant participation of no-treatment decisions in patients diagnosed with noninvasive breast cancer.

We propose an extension of the model presented by Zhang and Ivy [41] that is based on the relaxation of the assumption regarding perfect information in mammography results; more specifically, we incorporate the impact of type II error in the outcomes of the test. We define type II error as a false negative result caused by the sensitivity of mammography and incorporate this risk and its subsequent failure in our model. The rest of the chapter is organized as follows. In Sect. 2, we review medical and operations research literature related to our problem. In Sect. 3, we describe the model for optimal treatment policies. In Sect. 4, we present our computational experiments and results. Finally, Sect. 5 concludes the chapter and outlines the future work.

2 Literature Review

This section presents a summary of literature references considered relevant for purposes of this study. First, we review research regarding cancer regression and imperfection in mammography results. Then, analytical studies concerning screening and treatment policies are presented.

2.1 Cancer Regression and Imperfection of Mammography

There is a heated discussion in the medical community regarding overdiagnosis of cancer. Overdiagnosis may happen when the cancer never progresses, or in fact, regresses. The literature review on the medical exploration of breast cancer spontaneous regression has been summarized by Zhang and Ivy [40]. Multiple sources [7, 21, 22, 27, 39] have indicated that although the phenomenon is rare, there is ample evidence to confirm that spontaneous regression of breast cancer does exist, and it may lead to overdiagnosis if ignored. Since the current protocol recommends women to seek treatment after diagnosis [36], it is difficult to observe the natural history of breast cancer progression and regression. And thus, it is not easy to calculate the probability of breast cancer regression directly. In this chapter, we explore the impact of cancer regression on treatment decisions by varying the probability of regression. We also present the results when regression is not incorporated. This can facilitate treatment decisions based on the belief of regression.

The human intervention during mammographic interpretation (detection and classification) makes mammography results subject to possible failure when interpreting mammographic images. Several studies have shown differences among radiologists when interpreting mammograms [4, 18, 34]. Kerlikowske et al. [19] observed interpretation differences among two radiologists with wide experience in reading mammograms, finding the overall sensitivity ranging from 72.8 to 78.2 % in a study that considered 71,713 screening examination. Elmore and Carney [12] claim that there exists a clinical and significant variation among radiologists when interpreting

mammograms. They suggested that such variation may be attributed to personal, clinical, financial, and legal characteristics of the radiologists, and/or the characteristics of the mammography facility. Beam et al. [3] tested 110 radiologists who interpreted screening mammograms from the same 148 women. They found that sensitivity in the sample of radiologists ranged from 59 to 100 %, and specificity ranged from 35 to 98 %. In addition, they discussed how lack of skill maintenance or improvement mechanisms may affect the interpretation of mammographic images.

2.2 Screening and Treatment Policies

The optima frequency to perform screening tests for breast cancer early detection is a well-studied problem in the operations research literature. Ayer et al. [2] proposed a partially observable MDP model to determine mammography screening decisions based on personalized risk factors. Their model uses QALYs as the measure to be maximized in the decision process and considers the effect of breastself-exam. They concluded that age should not be the only risk factor to be taken into account for screening recommendations and that personalized screening strategies may be more beneficial in decreasing death rates.

Maillart et al. [24] used sample-path enumeration to assess a broad range of different screening recommendations. They considered the imperfect nature of screening in their models and their numerical experiments included mammography screening policies with different starting ages, first screening interval, policy-switching age, second screening interval, and end age; the measure used to compare the performance of different policies is the lifetime breast cancer mortality risk. They developed efficient frontiers for optimal policies considering a trade-off between lifetime mortality risk and the expected number of mammograms.

Michaelson et al. [25] developed a simulation model to determine optimal screening intervals using biologically based data regarding tumor growth and spread. They compared different screening frequencies and estimated the reduction in incidence of distant metastases. As a conclusion of their study, they suggested that death rate from breast cancer could be positively and significantly impacted by an increase in the frequency of mammograms.

Chhatwal et al. [11] formulated a finite-horizon discrete-time MDP to determine when a woman should be sent for biopsy based on her mammographic features and demographic factors. They concluded that the decision to biopsy should take the age of the patient into account with older women having a higher biopsy threshold than younger women. Additionally, they concluded that false-positive interpretation of mammography may lead to unnecessary invasive procedures causing complications in older patients with comorbidities.

In the medical community, screening policies have been widely studied by means of cost-effectiveness analysis [5, 26, 29, 38] in which various screening tests are compared in order to establish their effectiveness and accuracy in reducing mortality rates. On the other hand, although treatment decisions and techniques have had an

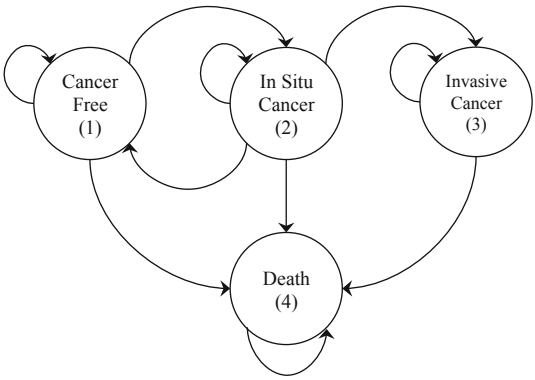


Fig. 1 Four-state discrete-time Markov chain for breast cancer natural history

important place in medical research [8, 17, 31], there are few studies in the operations research literature regarding this topic. The inclusion of cancer regression as a fairly new concept has led to studies such as [41] where treatment decisions are considered dynamic and patients may benefit more from watchful waiting.

3 Methods

This section is devoted to describe our model for optimal breast cancer treatment. Details of the model formulation and sources of input data are presented.

3.1 Model Formulation

In order to find the optimal treatment policy for breast cancer that considers the medical facts discussed in Sects. 2.1 and 2.2, we formulate a discrete-time, finite-horizon MDP model. The objective is to maximize the total expected QALYs of a patient. QALY is the arithmetic product of life expectancy and a measure of the quality of the remaining life- years. It is used to assess the extent of the benefits gained from a variety of interventions in terms of health-related quality of life and survival for the patient [28]. We model the natural history of breast cancer using a discrete-time Markov chain with the following four states: cancer free, in situ cancer, invasive cancer, and death. As seen in Fig. 1, we consider cancer progression and steadiness for every cancer state. On the other hand, spontaneous breast cancer regression is considered only from in situ cancer.

We propose a model from the patient's perspective in the sense that treatment or watchful waiting will directly affect a woman's health. On the other hand, the decision maker is assumed to be the doctor. In our model, at every decision epoch, a woman undergoes a mammogram that is examined by a radiologist who determines whether any abnormality is present or not. If the mammography result is negative

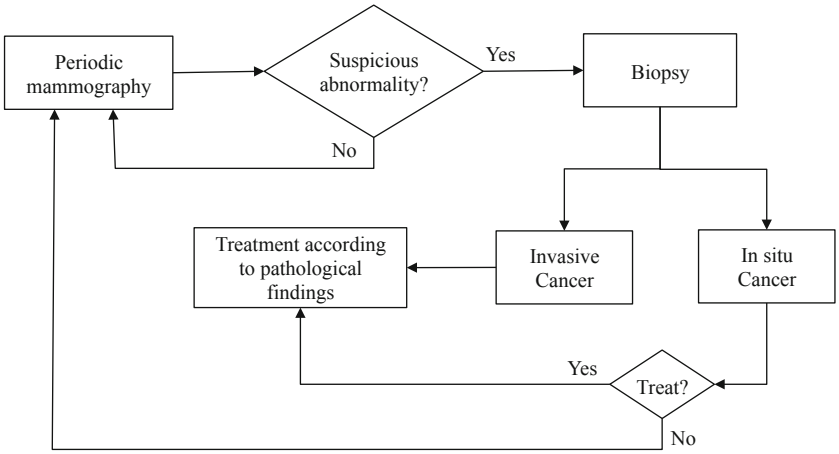


Fig. 2 Decision process for breast cancer early diagnosis and treatment

and hence no abnormality is observed, no further tests are performed until the next decision epoch. On the other hand, if the mammography result turns out to be positive, a follow-up biopsy test is performed in order to confirm the existence of cancer. As reported in the literature, breast biopsy sensitivity is very close to 100 % [14], and therefore this test is assumed to be perfect.

We assume that whenever the observed state of a patient is cancer free, the decision that will be made is to wait. In addition, if a patient is diagnosed with invasive cancer, the decision maker will always decide to treat that patient. These assumptions have been studied and established as optimal in medical guidelines regarding breast cancer treatment [9]. However, when the observed state is in situ (noninvasive) cancer, unlike the ACS recommendation, our model not only considers treatment but also evaluates the possibility of waiting. This assumption differs from other existing models and is based on the inclusion of cancer regression as an established medical fact. Figure 2 presents our model decision process for breast cancer detection.

It is worth mentioning that our model does not include type I error in mammography results (false positive results). Therefore, a positive outcome in mammography always implies the patient has either noninvasive or invasive cancer since perfect biopsy tests are used to confirm the presence of the disease. On the contrary, type II error is included in the model and hence a negative result in mammography does not necessarily imply that the patient is healthy.

Given all the particularities above, we define an MDP model with the following components:

- Set of decision epochs $\Upsilon = \{40, 41, 42, ..., 100\}$. According to the ACS recommendation, a woman should receive annual mammography screening from the age of 40 [36]. We adopt this ACS recommendation as the screening schedule of our model and define the upper boundary of life as 100 years in accordance with the maximum life span reported in the US Life Table for 2012 [1].

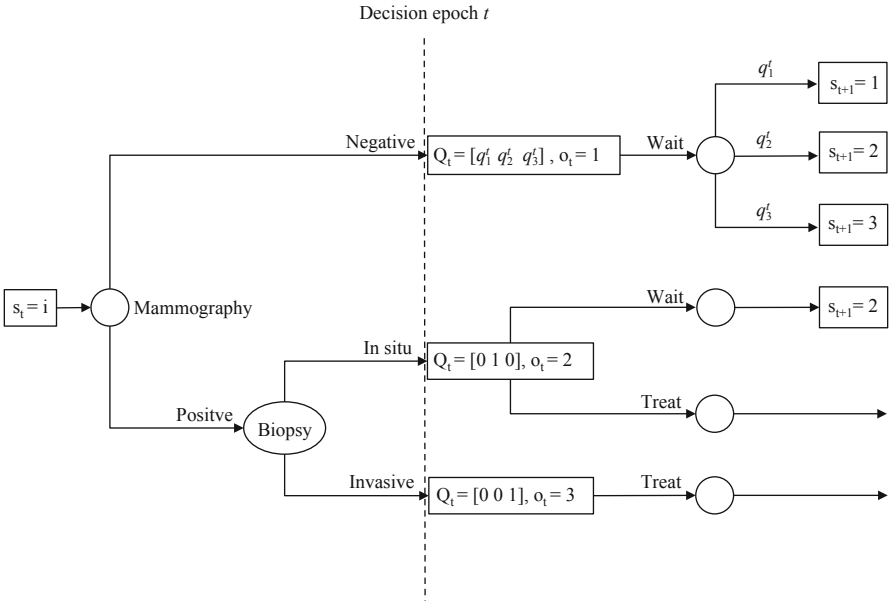


Fig. 3 Decision-making process for breast cancer treatment

- State space $S = \{1, 2, 3, 4\}$, where the cancer state of a patient at decision epoch t is defined as $s_t \in S\ \forall\, t \in \Upsilon$. In particular, 1 represents a cancer-free patient, 2 represents a patient with in situ (noninvasive) cancer, 3 represents a patient with invasive cancer, and 4 represents a death state.
- Postdiagnosis cancer distribution is denoted by $Q_t(S)\ \forall\ t\ \in\ \Upsilon$. We define the *postdiagnosis cancer distribution* as a discrete probability distribution once the diagnosis procedure is finished. The element q_s^t represents the probability that the state of a patient is s at decision epoch t after the patient has undergone a mammogram or a mammogram and a biopsy test and therefore, $Q_t(S) = \{q_s^t : s \in S \setminus \{4\}\}$.
- Observed cancer state space $\Omega = \{1, 2, 3\}$. After the diagnosis procedure is finished the observed cancer state $o_t\ \forall\ t\ \in\ \Upsilon$ can be healthy (1), with in situ cancer (2) or invasive cancer (3). Since we do not consider type I error, the observed and the real cancer states are the same when malignant cells are present in a patient. On the contrary, when the observed cancer state is healthy, we use the postdiagnosis cancer distribution to describe the real cancer state of a patient.
- Actions space $A = \{W, T\}$, where W and T represent *wait* and *treat,* respectively. Our model assumes that the only feasible action is to wait when the cancer-free state is observed; when invasive cancer is observed, the only feasible action is to treat; and finally, the decision maker may suggest to wait or to treat for a patient whose observed state is in situ cancer. Figure 3 presents the decision-making process of our model.

- Transition probability matrices $P(a_t)\ \forall\ t\ \in\ \Upsilon$. We select a set of transition probability matrices presented by Maillart et al. [24] that aim at quantitatively describing the natural history of breast cancer. These matrices are divided into 5-year age groups and the element $p_{ij}(a_t)$ represents the probability that a woman at age group $a_t\ \forall\ t\ \in\ \Upsilon$ in state i transitions to state j within 1 year of no treatment. We assume the matrices proposed by Maillart et al. [24] describe the natural history of breast cancer with one-year transitions despite the fact they were designed as 6-month transition probability matrices. This assumption is made due to the lack of quantitative and detailed information regarding natural history of breast cancer. An example of a transition probability matrix from Maillart et al. [24] is presented below. Note that the following matrix has two different death states which are combined to form the death state in our model.

$$P(a_t) = \begin{pmatrix} p_{11}(a_t) & p_{12}(a_t) & 0 & 0 & p_{15}(a_t) \\ 0 & p_{22}(a_t) & p_{23}(a_t) & 0 & p_{25}(a_t) \\ 0 & 0 & p_{33}(a_t) & p_{34}(a_t) & p_{35}(a_t) \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \tag{1}$$

where
1 → Cancer free
2 → In situ cancer
3 → Invasive cancer
4 → Death from breast cancer
5 → Death from other causes
It is worth mentioning that these matrices do not consider cancer regression. We use an analytical methodology presented by Zhang and Ivy [40] in order to include this medical fact. Zhang and Ivy [40] proposed the following modification to the original matrix for including cancer regression:

$$P^{\star}(a_t) = \begin{pmatrix} p_{11}(a_t) & p_{12}(a_t) & 0 & 0 & p_{15}(a_t) \\ p_{21}(a_t) & p^{*}_{22}(a_t) & p^{*}_{23}(a_t) & 0 & p_{25}(a_t) \\ 0 & 0 & p_{33}(a_t) & p_{34}(a_t) & p_{35}(a_t) \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{2a}$$

$$p_{21}(a_t) = u \cdot p_{22}(a_t) + v \cdot p_{23}(a_t) \tag{2b}$$

$$p^{*}_{22}(a_t) = (1 - u) \cdot p_{22}(a_t) \tag{2c}$$

$$p^{*}_{23}(a_t) = (1 - v) \cdot p_{23}(a_t) \tag{2d}$$

$$0 \leq u, v \leq 1 \tag{2e}$$

where u and v are fractions of the self-loop and the progression transition probabilities respectively; these proportions are used to extract information from the existing probabilities and build the regression transition.

- Immediate rewards $r_t(s,a)\ \forall\ s_t \in S,\ a \in A,\ t \in \Upsilon$. At every decision epoch, we measure the impact of treatment in terms of QALYs as a function of the age, cancer state, and action. In this notation, $r_t(s,a)$ represents the total expected QALYs accumulated at decision epoch t, when the cancer state of a patient is s and action a is taken. We use the estimations done by Stout et al. [37] regarding QALYs at in situ and invasive cancer states when the decision is *wait*. These estimations were derived from EuroQol EQ-5D quality-of-life utility scores along with a series of modifications to estimate the QALYs accrued for a woman with in situ or invasive cancer. The EQ-5D is a standardized measure for general health developed by the EuroQol Group [6].
  On the other hand, when the decision is *treat,* our model uses a life expectancy estimation after the necessary treatment is performed. Zhang and Ivy [41] proposed a methodology to estimate age-specific 5-year QALYs for breast cancer treatment and calculate the expected total QALYs taking into account different survival probabilities depending on cancer state.
- Discount factor λ. We select a discount factor of 0.97 that has been previously used in dynamic decisions models regarding medical treatment [11].

3.2 Type II Error in Mammography Results

Our model considers type II error of mammography results which means no abnormality may be identified on the mammogram image when in fact such an abnormality exists. This type of error generates uncertainty about the real cancer state of a patient once the mammography result is negative. We model this uncertainty by means of a discrete probability distribution that describes the cancer state after diagnosis. As seen in Fig. 3, when the diagnosis includes a biopsy intervention the uncertainty disappears, thanks to the high accuracy of this test. However, when a negative result is given and no further tests are performed, there exists a positive probability that a woman has in situ or invasive cancer.

As reported in the literature, mammography sensitivity depends on both the age of the patient and the cancer state [15]. We define sens_s^t as the sensitivity at decision epoch t when the cancer state of a woman is s. Therefore, $Q_t(S)$ is defined in terms of the sensitivity as follows:

$$Q_t(S) = \begin{bmatrix} q_1^t & q_2^t & q_3^t \end{bmatrix} \tag{3a}$$

$$q_3^t = 1 - \text{sens}_3^t \tag{3b}$$

$$q_2^t = 1 - \text{sens}_2^t \tag{3c}$$

$$q_1^t = 1 - q_2^t - q_3^t \tag{3d}$$

The mammography sensitivity function is defined in [24]. Here, q_1^t is implicitly a theoretical estimation of mammography specificity but as previously mentioned we do not consider the error caused by this statistical measure of performance. Our model also considers the potential misdiagnosis and consequent delay in cancer detection caused by sensitivity. We propose a series of modifications to the transition probability matrices to consider type II error in mammography. These modifications are based on the idea that transitions occur between observed cancer states o_t and not between real cancer states s_t as proposed by [24, 40]. Below, we present the modified transition probability matrix of $P(a_t)$, with the inclusion of breast cancer regression and type II error in mammography results.

$$P'(a_t) = \begin{pmatrix} p_{11}(a_t) & p_{12}(a_t) & 0 & p_{14}(a_t) \\ p'_{21}(a_t) & p'_{22}(a_t) & p'_{23}(a_t) & p_{24}(a_t) \\ p'_{31}(a_t) & 0 & p'_{33}(a_t) & p_{34}(a_t) \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{4a}$$

$$p'_{21}(a_t) = \underbrace{p_{21}(a_t)}_{\text{Natural transition}} + \underbrace{p^*_{22}(a_t) \cdot \overbrace{(1 - \text{sens}_2^t)}^{\text{Type II error}} + p^*_{23}(a_t) \cdot \overbrace{(1 - \text{sens}_3^t)}^{\text{Type II error}}}_{\text{Transitions due to imperfect mammography}} \tag{4b}$$

$$p'_{22}(a_t) = p^*_{22}(a_t) \cdot \text{sens}_2^t \tag{4c}$$

$$p'_{23}(a_t) = p^*_{23}(a_t) \cdot \text{sens}_3^t \tag{4d}$$

$$p'_{31}(a_t) = p_{33}(a_t) \cdot (1 - \text{sens}_3^t) \tag{4e}$$

$$p'_{33}(a_t) = p_{33}(a_t) \cdot \text{sens}_3^t \tag{4f}$$

3.3 Optimality Equations

We denote by $V_t(o)$ the maximum total expected QALYs the patient can attain when the current observed cancer state is o at decision epoch t. Then,

$$V_t(o) = \begin{cases} \sum_{u \in S} q_u^t \left(r_t(u, W) + \lambda \sum_{s' \in \Omega} p_{us'}(a_t) V_{t+1}(s') \right) & o_t = 1 \\ \max \begin{cases} r_t(o, W) + \lambda \sum_{s' \in \Omega} p_{os'}(a_t) V_{t+1}(s') \\ r_t(o, T) \end{cases} & o_t = 2 \\ r_t(o, T) & o_t = 3 \end{cases} \tag{5}$$

For the case when the observation is cancer free, we use the postdiagnosis cancer distribution to calculate the immediate and discounted future QALYs that a patient

Table 1 Sources of model inputs

Parameter	Notation	Data source
Cancer state transition probabilities	$p_{ij}(a_t)$	Maillart et al. [24]
Sensitivity and specificity of mammography	sens_s^t	Maillart et al. [24]
Immediate rewards	$r_t(s, W)$	Stout et al. [37]
Immediate rewards	$r_t(s, T)$	Zhang and Ivy [41]

may obtain. On the other hand, if in situ cancer is observed, the model will decide either to wait or treat depending on the difference between the immediate, plus the discounted future QALYs, and the estimation of expected QALYs for the remaining life. Finally, when invasive cancer is observed, $V_t(o)$ always equals the estimation of expected QALYs for the remaining life. The terminal values at year 100 for the decision-making process are defined below:

$$V_{100}(o) = \begin{cases} \sum_{u \in S} q_u^{100} \left(r_{100}(u, W)\right) & o_t = 1 \\ \max \begin{cases} r_{100}(o, W) \\ r_{100}(o, T) \end{cases} & o_t = 2 \\ r_{100}(o, T) & o_t = 3 \end{cases} \tag{6}$$

3.4 Implementation and Sources of Model Inputs

Solving the optimality equations introduced in Sect. 3.3 can generate optimal decisions at each decision epoch given an observed state. We implement these iterative equations using MATLAB and solve the model to optimality using the *backward induction* algorithm [30]. To do so, we use a series of input data obtained from different sources which are listed in Table 1.

Maillart et al. [24] provided relevant input data regarding transition probability matrices and functions of sensitivity and specificity described in Sects. 3.1 and 3.2, respectively. The immediate rewards are directly calculated by Stout et al. [37] when the decision is *wait*. Given the considerations described in Sect. 3.1, the estimations by Zhang and Ivy [41] are adopted as our immediate rewards when treatment is recommended.

4 Results and Discussion

We defined four different scenarios in order to assess the optimal treatment policy if a given patient is diagnosed with in situ cancer at a given age. In the first scenario, we evaluate the performance of our model when none of the considerations discussed in Sect. 3 regarding cancer regression and type II error in mammography results are taken into account. The second scenario includes the modifications proposed by Zhang and Ivy [41] related to cancer regression but does not consider type II error in mammography results; in this scenario, u and v are assumed to be 0.2 and 0.2, which results in an average regression rate of 20 %. This assumption is made due to lack of evidence supporting a specific function or relation between u and v. To minimize bias, we assume the probability of regression in the Markov model comes from self-loop and progression probabilities with equal probability. The third scenario assesses the optimal treatment policy when type II error in mammography results is taken into account but cancer regression is not. Finally, our proposal is described by the fourth scenario, which presents the optimal policy including all considerations discussed in Sect. 3; this scenario considers the same u and v values as in the second scenario to incorporate cancer regression. Table 2 presents the results obtained for each scenario. From these results, we can conclude with respect to each scenario that:

Scenario 1. When cancer regression and type II error in mammography results are ignored, a patient between the ages 40–60 (inclusive) who is diagnosed with in situ breast cancer should always be treated. However, for patients older than 60, the recommendation is to wait until the next screening period except at ages 65, 70, 75, 80, 85, and 100. This trend may be explained by the nature of the data, used to describe the natural history of breast cancer. As discussed in Sect. 3.1, our model uses age-specific transition probability matrices as input. Specifically, the information that the algorithm uses iteratively is updated every 5 years and causes this behavior. Note that between ages 85–100, there are only two information updates, since the matrix that describes the natural history of breast cancer is the same for the past 15 years.

Scenario 2. The optimal policy proposed in this scenario clearly reflects the impact of cancer regression in treatment decisions. As discussed in Sect. 2, if cancer regression is ignored, treatment policies may lead to overtreatment and therefore, a decrease in quality of life. Our model handles this undesirable situation by increasing the waiting decisions along the decision horizon. The optimal results suggest that more waiting decisions should be made, especially for older patients.

As the results shown, treatment is the optimal decision for a patient aged between 40–50 (inclusive) who is diagnosed with in situ cancer. On the other hand, patients older than 50 with the same diagnosis should wait until the next screening period except at age 55 and 60. This behavior is explained by the structure of the immediate rewards $r_t(s, a)$ used to describe the impact of treatment in quality life. According to the estimations proposed in [37] and [41], once a patient is diagnosed with in situ cancer the impact of waiting decreases as the age at diagnosis increases. Clearly, there is a trade-off between cancer regression and decrease in quality of life when waiting.

Table 2 Optimal treatment policy for in situ breast cancer

Age	No regression or type II error	Regression	Type II error	Regression and type II	Age	No regression or type II error	Regression	Type II error	Regression and type II
40	T	T	T	T	71	W	W	W	W
41	T	T	T	T	72	W	W	W	W
42	T	T	T	T	73	W	W	W	T
43	T	T	T	T	74	W	W	T	T
44	T	T	T	T	75	T	W	T	T
45	T	T	T	T	76	W	W	W	W
46	T	T	T	T	77	W	W	W	W
47	T	T	T	T	78	W	W	W	W
48	T	T	T	T	79	W	W	W	T
49	T	T	T	T	80	T	W	T	T
50	T	T	T	T	81	W	W	W	W
51	T	W	T	T	82	W	W	W	W
52	T	W	T	T	83	W	W	W	W
53	T	W	T	T	84	W	W	W	W
54	T	W	T	T	85	T	W	T	T
55	T	T	T	T	86	W	W	W	W
56	T	W	T	T	87	W	W	W	W

Table 2 (Continued)

Age	No regression or type II error	Regression	Type II error	Regression and type II	Age	No regression or type II error	Regression	Type II error	Regression and type II
57	T	W	T	T	88	W	W	W	W
58	T	W	T	T	89	W	W	W	W
59	T	W	T	T	90	W	W	W	W
60	T	T	T	T	91	W	W	W	W
61	W	W	T	T	92	W	W	W	W
62	W	W	T	T	93	W	W	W	W
63	W	W	T	T	94	W	W	W	W
64	W	W	T	T	95	W	W	W	W
65	T	W	T	T	96	W	W	W	W
66	W	W	W	T	97	W	W	W	W
67	W	W	W	T	98	W	W	W	W
68	W	W	T	T	99	W	W	W	W
69	W	W	T	T	100	T	T	T	T
70	T	W	T	T					

Scenario 3. In Sect. 3.2, we mentioned a potential delay in cancer detection as a direct consequence of sensitivity of mammography results. When our model only considers type II error in mammography results, the optimal policy shows how this delay is avoided through an increase in treatment decisions. As shown in Table 2, the optimal decisions include treatment till a later age (65) as opposed to 60 in *Scenario 1*, and 50 in *Scenario 2*. Here, a patient diagnosed with in situ breast cancer should be treated if the diagnosis is done aged between 40–70 except at age 66 and 67. When the age exceeds 70, a patient should wait until the next screening period unless the age is 74, 75, 80, or 85. As in *Scenario 1,* this trend may be explained by the nature of the data used to describe the natural history of breast cancer.

The clear increase in treatment decisions reveals the response of our model to type II error in mammography results. In this case and given the uncertainty regarding the real cancer state, our model handles the situation by suggesting more treatment decisions which prevents patients from being diagnosed at a later cancer stage.

Scenario 4. This scenario considers the impact of both cancer regression and type II error in mammography results. In *Scenarios 2* and *3,* we showed how the inclusion of cancer regression and type II error in mammography results would lead to an increase in waiting and treatment decisions, respectively. Therefore, the simultaneous inclusion of these medical facts proposes a trade-off between unnecessary treatment and delay in cancer detection. Our optimal treatment policy can be compared to the current policy in which treatment is always recommended but only between age 40–70. In addition, our results suggest for patients 80 years and older, *wait* is always the optimal decision if in situ cancer is diagnosed. As previously mentioned, this is a significant difference between our model and the existing work that does not consider no-treatment decisions for patients diagnosed of breast cancer.

In addition to the scenarios analysis, our model allowed us to assess how determinant cancer regression might be in treatment decisions but taking into account that our study is not aimed at robustly determining the rate at which this phenomenon occurs. As mentioned in Sect. 2.1, breast cancer regression is a medical fact currently considered rare but not improbable in the medical community. Unlike the way, type II error in mammography results is included in our model; there are not statistical measures currently available in the literature that provide a reliable estimation of the rate at which cancer regresses, though analytical methodologies have been proposed [40, 41] in order to include this medical fact into a dynamic decision process. That said, we propose a sensitivity analysis on the cancer regression rate to assess its impact on the optimal treatment policy for breast cancer.

Table 3 presents a sensitivity analysis on the regression rate for *Scenario 4*. The u and v values are equally fixed in such a way that the average regression rate over all age groups corresponds to the values presented at the column fields. As seen in Table 3, the policies appear to be stable when varying the probability. This can be explained by the inclusion of type II error which does not allow waiting decisions to sharply increase due to the trade-off described in *Scenario 4*. The regression probability is currently believed to be 22 % according to [39]. Considering a regression rate of 20 % and once in situ breast cancer is diagnosed, the vast majority of decisions between

Table 3 Sensitivity analysis on breast cancer regression rate

Age	0 %	5 %	10 %	15 %	20 %	25 %	Age	0 %	5 %	10 %	15 %	20 %	25 %
40	T	T	T	T	T	T	71	W	W	W	W	W	W
41	T	T	T	T	T	T	72	W	W	W	W	W	T
42	T	T	T	T	T	T	73	W	W	T	T	T	T
43	T	T	T	T	T	T	74	T	T	T	T	T	T
44	T	T	T	T	T	T	75	T	T	T	T	T	T
45	T	T	T	T	T	T	76	W	W	W	W	W	W
46	T	T	T	T	T	T	77	W	W	W	W	W	W
47	T	T	T	T	T	T	78	W	W	W	W	W	W
48	T	T	T	T	T	T	79	W	W	W	T	T	T
49	T	T	T	T	T	T	80	T	T	T	T	T	T
50	T	T	T	T	T	T	81	W	W	W	W	W	W
51	T	T	T	T	T	T	82	W	W	W	W	W	W
52	T	T	T	T	T	T	83	W	W	W	W	W	W
53	T	T	T	T	T	T	84	W	W	W	W	W	W
54	T	T	T	T	T	T	85	T	T	T	T	T	T
55	T	T	T	T	T	T	86	W	W	W	W	W	W
56	T	T	T	T	T	T	87	W	W	W	W	W	W
57	T	T	T	T	T	T	88	W	W	W	W	W	W
58	T	T	T	T	T	T	89	W	W	W	W	W	W
59	T	T	T	T	T	T	90	W	W	W	W	W	W
60	T	T	T	T	T	T	91	W	W	W	W	W	W
61	T	T	T	T	T	T	92	W	W	W	W	W	W
62	T	T	T	T	T	T	93	W	W	W	W	W	W
63	T	T	T	T	T	T	94	W	W	W	W	W	W
64	T	T	T	T	T	T	95	W	W	W	W	W	W
65	T	T	T	T	T	T	96	W	W	W	W	W	W
66	W	W	T	T	T	T	97	W	W	W	W	W	W
67	W	T	T	T	T	T	98	W	W	W	W	W	W
68	T	T	T	T	T	T	99	W	W	W	W	W	W
69	T	T	T	T	T	T	100	T	T	T	T	T	T
70	T	T	T	T	T	T							

age 40–80 are *treat* while *wait* is the most common decision after age 80. The results suggest that an average regression probability of 20 % allows treatment and waiting decisions to be associated with specific age ranges.

5 Conclusions and Future Work

We formulate an MDP model to determine the optimal treatment policies for breast cancer treatment in the presence of two proven medical facts: cancer regression and type II error in mammography results. We solve our model to optimality and obtain results that give an insight about the complexity of this disease. Our study suggests that optimal treatment policies for breast cancer might be different from the common recommendations. Specifically, we show that when in situ breast cancer is diagnosed, the quality of life may be negatively affected if treatment is always recommended. Our results show that the optimal decisions between age 40–100 should be: to *wait* 40% of the times (mainly after age 80) and to *treat* 60 % of the times. This contrasts to the current policy that always suggests treatment (100 % of the times). The sensitivity analyses on cancer regression suggest that there is little variation in the optimal treatment decisions when different cancer progression probabilities are considered; regardless of how small the regression rate is, more waiting decisions (on average 40 % of the times) are preferred.

In addition, we find a trade-off between overtreatment and late cancer detection when we analyze different scenarios regarding cancer regression and type II error in mammography results; we consider both factors and our results show that treatment may be optimal only for patients younger than 80 if in situ cancer is diagnosed. The decision not to treat may benefit patients older than 80 given the fact that these women may have comorbidities, and treatment may decrease their qualify of life significantly.

The contributions of our work include the handling of uncertainty in the real cancer state of a patient through the post-diagnosis cancer state and the modifications proposed to incorporate the observed cancer state in a simple and intuitive way into a dynamic decision model. Also, to the best of our knowledge, the study proposed by Zhang and Ivy [41] is the first to consider no treatment decisions for cancer-diagnosed patients and our study contributes to this work by adding factors that may influence treatment policies. Finally, our study also contributes to the literature of analytical studies for medical decision modeling of breast cancer that considers disease regression. Although the biological reasons behind regression are not fully known yet, this work takes a first step to examine the impact of regression on breast cancer treatment decisions. It shows how the decisions vary when different probabilities of cancer regression are incorporated. However, as there is no study on the relations between cancer progression and regression, the way we extract the regression probability in the Markov model can create a bias.

An important step following this work is to improve both the transition probability matrices used to describe the natural history of breast cancer and the estimations regarding quality of life. As discussed in Sect. 3, the transition probability matrices required by our approach are difficult to estimate and the information available in the literature lacks a more detailed differentiation regarding age and population dependency. A good contribution would be to estimate the natural history of breast

cancer using shorter age ranges and to provide more accurate information regarding sensitivity and specificity depending on the age and cancer state. Likewise, it is important to include type I error in mammography results in order to obtain more realistic results.

References

1. Arias, E.: United States life tables, 2008. Natl. Vital Stat. Rep. **61**(3), 14–15 (2012)
2. Ayer, T., Alagoz, O., Stout, N.: A POMDP approach to personalize mammography screening decisions. Oper. Res. **60**(5), 1019–1034 (2011)
3. Beam, C., Conant, E., Sickles, E.: Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. J. Natl. Cancer Inst. **95**(4), 282–290 (2003)
4. Beam, C., Sullivan, D., Layde, P. Effect of human variability on independent double reading in screening mammography. Acad. Radiol. **3**(11), 891–897 (1996)
5. Bonomi, A.E., Boudreau, D.M., Fishman, P.A., Ludman, E., Mohelnitzky, A., Cannon, E. A., Seger, D.: Quality of life valuations of mammography screening. Qual. Life Res. **17**(5), 801–814 (2008)
6. Brooks, R., Jendteg, S., Lindgren, B., Persson, U., Björk, S.: Euroqol©: health-related quality of life measurement. Results of the swedish questionnaire exercise. Health Policy **18**(1), 37–48 (1991)
7. Burnside, E., Trentham-Dietz, A., Kelcz, F., Collins, J.: An example of breast cancer regression on imaging. Radiol. Case Rep. **1**(2), 27–37 (2006)
8. Cao, A.-Y., Hu, Z., Shao, Z.-M.: Mutation screening of breast cancer susceptibility genes in chinese high-risk families: the results will develop the genetic testing strategy in china. Breast Cancer Res. Treat. **120**(1), 271–272 (2010)
9. Carlson, R.W., Anderson, B.O., Burstein, H.J., et al.: NCCN breast cancer clinical practice guidelines in oncology. J Natl Compr Canc Netw. 3, 238–289 (2005)
10. Chen, H., Duffy, S., Tabar, L.: A markov chain method to estimate the tumour progression rate from preclinical to clinical phase, sensitivity and positive predictive value for mammography in breast cancer screening. The Statistician **45**(3), 307–317 (1996)
11. Chhatwal, J., Alagoz, O., Burnside, E.. Optimal breast biopsy decision-making based on mammographic features and demographic factors. Oper. Res. **58**(6), 1577–1591 (2010)
12. Elmore, J., Carney, P.: Does practice make perfect when interpreting mammography? J. Natl. Cancer Inst. **94**(5), 321–323 (2002)
13. Fracheboud, J., Groenewoud, J., Boer, R., Draisma, G., de Bruijn, A., Verbeek, A., de Koning, H.: Seventy-five years is an appropriate upper age limit for population-based mammography screening. Int. J. Cancer **118**(8), 2020–2025 (2006)
14. Gur, D., Wallace, L., Klym, A., Hardesty, L., Abrams, G., Shah, R., Sumkin, J.: Trends in recall, biopsy, and positive biopsy rates for screening mammography in an academic practice. Radiology, **235**(2), 396–401 (2005)
15. Harris, J., Lippman, M., Morrow, M., Osborne, C., Fund, R.: Diseases of the breast. Lippincott Williams & Wilkins, Philadelphia, PA (2000)
16. Jemal, A., Bray, F., Center, M., Ferlay, J., Ward, E., Forman, D.: Global cancer statistics. CA Cancer J. Clin. **61**(2), 69–90 (2011)
17. Jeon, Y. W., Choi, J.E., Park, H.K., Kim, K.S., Lee, J.Y., Suh, Y.J.. Impact of local surgical treatment on survival in young women with t1 breast cancer: Long-term results of a population-based cohort. Breast Cancer Res. Treat. **138**(2), 475–84 (2013)
18. Karssemeijer, N., Otten, J., Verbeek, A., Groenewoud, J., de Koning, H., Hendriks, J., Holland, R.: Computer-aided detection versus independent double reading of masses on mammograms. Radiology **227**(1), 192–200 (2003)

19. Kerlikowske, K., Grady, D., Barclay, J., Ernster, V., Frankel, S., Ominsky, S., Sickles, E.: Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. J. Natl Cancer Inst. **90**(23), 1801–1809 (1998)
20. Kerlikowske, K., Grady, D., Rubin, S., Sandrock, C., Ernster, V.: Efficacy of screening mammography. JAMA **273**(2), 149–154 (1995)
21. Larsen, S., Rose, C.: Spontaneous remission of breast cancer. A literature review. Ugeskrift Laeger **161**(26), 4001 (1999)
22. Lewison, E.: Spontaneous regression of breast cancer. Natl Cancer Inst. Monogr. **44**, 23 (1976)
23. Magnus, M.C., Ping, M., Shen, M.M., Bourgeois, J., Magnus, J.H.: Effectiveness of mammography screening in reducing breast cancer mortality in women aged 39–49 years: A meta-analysis. J. Wom. Heal. **20**(6), 845–852 (2011)
24. Maillart, L., Ivy, J., Ransom, S., Diehl, K.: Assessing dynamic breast cancer screening policies. Oper. Res. **56**(6), 1411–1427 (2008)
25. Michaelson, J., Halpern, E., Kopans, D.: Breast cancer: computer simulation method for estimating optimal intervals for screening. Radiology **212**(2), 551–560 (1999)
26. Moore, S.G., Shenoy, P.J., Fanucchi, L., Tumeh, J.W., Flowers, C.R.: Cost-effectiveness of MRI compared to mammography for breast cancer screening in a high risk population. BMC Health Serv. Res. **9**(1), 9 (2009)
27. Osler W. The medical aspects of carcinoma of the breast, with a note on the spontaneous disappearance of secondary growth. Am Med. 17–19 (1901)
28. Phillips, C., Thompson, G.: *What Is a QALY?* vol. 1. Hayward Medical Communications Hayward Medical Communications, London (1998)
29. Pisano, E.D., Gatsonis, C., Hendrick, E., Yaffe, M., Baum, J.K., Acharyya, S., Conant, E.F., Fajardo, L.L., Bassett, L., D'Orsi, C., et al.: Diagnostic performance of digital versus film mammography for breast-cancer screening. New Engl. J. Med. **353**(17), 1773–1783 (2005)
30. Puterman, M. L.: *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, Hoboken, NJ (2005)
31. Ruiterkamp, J., Voogd, A.C., Bosscha, K., Tjan-Heijnen, V.C., Ernst, M.F.: Impact of breast surgery on survival in patients with distant metastases at initial presentation: a systematic review of the literature. Breast Cancer Res. Treat. **120**(1), 9–16 (2010)
32. Schaefer, A.J., Bailey, M.D., Shechter, S.M., Roberts, M.S.: Modelling medical treatment using Markov decision processes. In: Brandeau ML, Sainfort F, Pierskalla WP, editors. Operations research and health care: a handbook of methods and applications. Kluwer, Boston (MA) 593–612(2004)
33. Shen, Y., Yang, Y., Inoue, L.Y., Munsell, M.F., Miller, A.B., Berry, D.A. (2005): Role of detection method in predicting breast cancer survival: analysis of randomized screening trials. J. Natl Cancer Inst. **97**(16), 1195–1203 (2004)
34. Sickles, E., Wolverton, D., Dee, K.: Performance parameters for screening and diagnostic mammography: Specialist and general radiologists. Radiology **224**(3), 861–869 (2002)
35. Smith, R., Cokkinides, V., von Eschenbach, A., Levin, B., Cohen, C., Runowicz, C., Sener, S., Saslow, D., Eyre, H.: American cancer society guidelines for the early detection of cancer. CA Cancer J. Clin. **52**(1), 8–22 (2002)
36. Smith, R.A., Cokkinides, V., Brooks, D., Saslow, D., Brawley, O.W.: Cancer screening in the united states, 2010: a review of current American Cancer Society guidelines and issues in cancer screening. CA Cancer J. Clin. **60**(2), 99–119 (2010)
37. Stout, N., Rosenberg, M., Trentham-Dietz, A., Smith, M., Robinson, S., Fryback, D.: Retrospective cost-effectiveness analysis of screening mammography. J. Natl Cancer Inst. **98**(11), 774–782 (2006)
38. Tosteson, A.N., Stout, N.K., Fryback, D.G., Acharyya, S., Herman, B., Hannah, L., Pisano, E., et al.: Cost-effectiveness of digital mammography breast cancer screening: results from acrin dmist. Ann. Intern. Med. **148**(1), 1 (2008)

39. Zahl, P., Mæhlen, J., Welch, H.: The natural history of invasive breast cancers detected by screening mammography. Arch. Intern. Med. **168**(21), 2311 (2008)
40. Zhang, S., Ivy, J.: Analytic modeling of breast cancer spontaneous regression. In: Proceedings of Industrial and Systems Engineering Research Conference (2012a)
41. Zhang, S., Ivy, J.: Optimal decision making in the presence of breast cancer regression. In: Proceedings of Industrial and Systems Engineering Research Conference (2012b)

On the Iterative Steering of a Rolling Robot Actuated by Internal Rotors

Akihiro Morinaga, Mikhail Svinin and Motoji Yamamoto

Abstract This chapter deals with a motion planning problem for a spherical rolling robot actuated by two internal rotors that are placed on orthogonal axes. The mathematical model of the robot, represented by a driftless control system, contains a physical singularity corresponding to the motion of the contact point along the equatorial line in the plane of the two rotors. It is shown that steering through the singularity by finding a globally regular valid basis is not applicable to the system under consideration. The solution of the motion planning problem employs the nilpotent approximation of the originally non-nilpotent robot dynamics, and is based on an iterative steering algorithm. At each iteration, the control inputs are constructed with the use of geometric phases. To solve the state-to-state transfer problem, a globally convergent steering algorithm with adjustable step size is implemented and tested under simulation. It is shown that its steering efficiency is not superior to the algorithm with constant iteration step size.

Keywords Motion planning · Rolling constraints · Non-holonomic systems

1 Introduction

In recent years, there is a growing interest in robotics research to spherical rolling robots: a rolling robot actuated by internal rotors. Under a proper placement of the rotors the center of mass of the robot is at the geometric center of the sphere and, as a result, the gravity terms do not enter the motion equations [2, 5].

A. Morinaga (✉) · M. Svinin · M. Yamamoto
Mechanical Engineering Department, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan
e-mail: morinaga@ctrl.mech.kyushu-u.ac.jp

M. Svinin
e-mail: svinin@mech.kyushu-u.ac.jp

M. Yamamoto
e-mail: yama@mech.kyushu-u.ac.jp

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_14

The mathematical model of rolling robot with two rotors inherits the basic properties of that for the ball–plate system [1, 8, 11]. Since it is not differentially flat, not nilpotent, and cannot be represented in a chained form, it belongs to a special class of nonholonomic systems, the class for which conventional planning techniques are not directly applicable.

One possible approach to control generic nonholonomic systems is to use iterative steering based on the nilpotent approximation of the the non-nilpotent system dynamics. Since the control problem for the nilpotent system can be solved exactly, the control inputs found for the nilpotent system can be used for iterative steering of the original non-nilpotent system. The idea was first proposed in [6] and later developed in [11], where an iterative algorithm with constant step size was developed and applied to steering of the ball–plate system. In [9], this approach was used in motion planning of a rolling robot with two rotors.

The further development in this research direction covered: the taking into consideration of possible singularities of the mathematical model [13, 14] and the development of globally convergent algorithms with adjustable step size [3, 4]. In this technical note, we address these issues using the mathematical model of the rolling robot with two rotors. Specifically, we are interested in the applicability of nonhomogeneous nilpotent approximation and in the efficiency of the steering algorithm with adjustable step size.

The chapter is organized as follows. In Sect. 2, a summary of the mathematical model established in [12] and [9] is given. In Sect. 3, an implementation of the globally convergent algorithm [3, 4] with adjustable step size is described. The steering algorithm is tested under simulation in Sect. 4. Finally, conclusions are drawn in Sect. 5.

2 Mathematical Model

The actuation scheme of the rolling robot is shown in Fig. 1. Define the state vector $\mathbf{x} = \{x, y, \vartheta, \varphi, \psi\}^{\mathrm{T}}$, where x, y are the coordinates of the center of the sphere and ϑ, φ, ψ are special Euler angles describing the orientation of the sphere.

The mathematical model of the rolling robot can be represented by the following five states-2 inputs driftless affine control system [9]:

$$\dot{\mathbf{x}} = \mathbf{f}_1(\mathbf{x})\dot{q}_1 + \mathbf{f}_2(\mathbf{x})\dot{q}_2, \tag{1}$$

where the rates of the rotors angles are considered as the control inputs, and the vector fields $\mathbf{h}_1$ and $\mathbf{h}_2$ are defined as:

$$\mathbf{f}_1 = \gamma \begin{bmatrix} \sin\vartheta \tan\varphi \\ \cos\vartheta \\ \sin\vartheta \dfrac{\sin^2\varphi + \kappa\cos^2\varphi}{\cos\varphi} \\ R(\cos\vartheta\sin\psi - \sin\vartheta\sin\varphi\cos\psi) \\ R(\cos\vartheta\cos\psi + \sin\vartheta\sin\varphi\sin\psi) \end{bmatrix}, \mathbf{f}_2 = \gamma \begin{bmatrix} 1 \\ 0 \\ (1-\kappa)\sin\varphi \\ -R\cos\varphi\cos\psi \\ R\cos\varphi\sin\psi \end{bmatrix}. \quad (2)$$

The dimensionless constants κ and γ are defined as:

$$\kappa = 1 + \frac{MR^2}{\frac{2}{3}m_o R^2 + 2J_p + J_r}, \quad \gamma = \frac{J_r}{MR^2}\frac{\kappa - 1}{\kappa}. \quad (3)$$

where R is the radius of the sphere, m_o is the mass of the spherical shell, M is the total mass of the composite system (the shell and the rotors), J_r and J_p are the inertia moments of the single rotor about,respectively, the axis of rotation and the plane orthogonal to the axis of rotation.

The motion planning for the robot under consideration consists of finding a trajectory $\mathbf{x}(t), t \in [0, T]$, given the start state $\mathbf{x}(0) = \mathbf{x}_s$ and the final state $\mathbf{x}(T) = \mathbf{x}_f$.

Let $L(\mathbf{f}_1, \mathbf{f}_2)$ be the Lie algebra generated by the vector fields $\mathbf{f}_1$ and $\mathbf{f}_2$. The first eight elements of the P. Hall basis of $L(\mathbf{f}_1, \mathbf{f}_2)$ are $\mathbf{f}_1$, $\mathbf{f}_2$, $\mathbf{f}_3 = [\mathbf{f}_1, \mathbf{f}_2]$, $\mathbf{f}_4 = [\mathbf{f}_1, \mathbf{f}_3]$, $\mathbf{f}_5 = [\mathbf{f}_2, \mathbf{f}_3]$, $\mathbf{f}_6 = [\mathbf{f}_1, \mathbf{f}_4]$, $\mathbf{f}_7 = [\mathbf{f}_2, \mathbf{f}_4]$, $\mathbf{f}_8 = [\mathbf{f}_2, \mathbf{f}_5]$, where $[\cdot, \cdot]$ stands for the Lie brackets of two vector fields. Define the distributions $B_{ij} = \{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, \mathbf{f}_i, \mathbf{f}_j\}$, $i = 4, 5$, $j = 5, 6, 7, 8$.

One can find [12] that the distribution B_{45} has full rank outside of the singularity set $\vartheta = \pm\pi/2$, where the contact point lies on the great circle in the equatorial plane of the rotors axes (red line in Fig. 1). However, the distribution B_{46} has full rank at $\vartheta = \pm\pi/2$ and therefore the system (1) is controllable.

Structurally, the driftless control system, describing the rolling robot with two rotors, is similar to that describing a two trailer vehicle [14]. The degree on nonholonomy and the grow vector outside the singular set are, 3 and $\{2, 3, 5\}$, respectively, while on the singularity set they become 4 and $\{2, 3, 4, 5\}$. However, the nature of the singularity is different. While for the two-trailer system it was possible to find a globally valid P. Hall basis (the basis B_{46} in [13, 14]), for the rolling system such a basis does not appear to exist. For example, the bases B_{ij}, $i = 4, 5$, $j = 6, 7, 8$, are not valid on the blue lines shown in Fig. 1. We have checked P. Hall bases with the Lie brackets up to the length 6 and could not find a globally valid one. The nonexistence of the globally valid basis is conjectured now; a formal proof of this statement constitutes the subject of future research.

Assuming that globally valid basis does not exist, one can still employ the technique of nonhomogeneous nilpotent approximations [13] with multiple approximations. A computational procedure, implementing such a technique, is computationally expensive. In addition, to synthesize the control actions, one would need to resort to generic techniques for steering of nilpotent systems, and this would also add to the increase of the computational time. In this situation, the motion

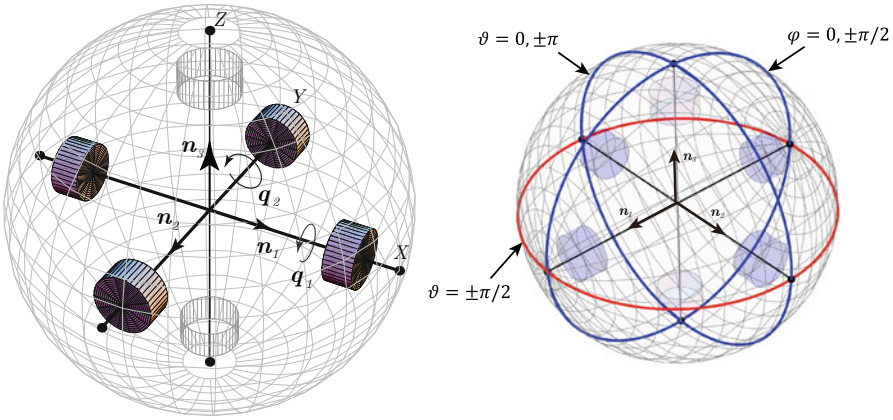


Fig. 1 Rolling system with orthogonal placement of rotors (*left*). Singular sets of different distributions (*right*). The *red* line represents the physical singularity at which $\det B_{45} = 0$; on the *blue* line corresponding to $\varphi = 0, \pm\pi$ one has $\det B_{47} = \det B_{56} = \det B_{58} = 0$; on the *blue* line corresponding to $\vartheta = 0, \pm\pi$ one has $\det B_{46} = \det B_{48} = \det B_{57} = 0$

planning strategy with decomposition into trivial and nontrivial maneuvers [9] is a reasonable alternative. In this strategy, the nilpotent approximation is conducted only at the south pole of the sphere $\vartheta_0 = \varphi_0 = 0$, where, compared to the other contact points, the calculations are considerably simpler.

Before approximating the original system (1), one converts it to a triangular form. This can be done with the use of the following input transformation:

$$\begin{bmatrix} \dot{q}_1 \\ \dot{q}_2 \end{bmatrix} = \frac{1}{\gamma} \begin{bmatrix} 0 & \sec\vartheta \\ 1 & -\tan\vartheta\tan\varphi \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \tag{4}$$

where u_1 and u_2 are the new control inputs. This results in the following representation:

$$\dot{\mathbf{x}} = \mathbf{g}_1(\mathbf{x})u_1 + \mathbf{g}_2(\mathbf{x})u_2, \tag{5}$$

where

$$\mathbf{g}_1 = \begin{bmatrix} 1 \\ 0 \\ (1-\kappa)\sin\varphi \\ -R\cos\varphi\cos\psi \\ R\cos\varphi\sin\psi \end{bmatrix}, \mathbf{g}_2 = \begin{bmatrix} 0 \\ 1 \\ \kappa\sec\varphi\tan\vartheta \\ R\sin\psi \\ R\cos\psi \end{bmatrix}. \tag{6}$$

In the neighborhood of the south pole of the sphere, where $\vartheta_0 = \varphi_0 = 0$, the transformation from the original to the privileged coordinates is constructed as [9]:

$$z_1 = \vartheta, \tag{7}$$

$$z_2 = \varphi, \tag{8}$$

$$z_3 = \frac{1}{2\kappa - 1}\psi, \tag{9}$$

$$z_4 = \frac{1}{3\kappa - 1}\varphi + \frac{\sin\psi_0}{R(1-3\kappa)}x + \frac{\cos\psi_0}{R(1-3\kappa)}y, \tag{10}$$

$$z_5 = \frac{1}{3\kappa - 1}\vartheta + \frac{\cos\psi_0}{R(1-3\kappa)}x - \frac{\sin\psi_0}{R(1-3\kappa)}y. \tag{11}$$

and the nilpotent approximation of the system (5) is obtained as:

$$\dot{\mathbf{z}} = \tilde{\mathbf{g}}_1(\mathbf{z})u_1 + \tilde{\mathbf{g}}_2(\mathbf{z})u_2, \tag{12}$$

where the vector fields $\tilde{\mathbf{g}}_1$ and $\tilde{\mathbf{g}}_2$ defined as:

$$\tilde{\mathbf{g}}_1 = \begin{bmatrix} 1 \\ 0 \\ \dfrac{1-\kappa}{2\kappa-1}z_2 \\ \cos\psi_0 \dfrac{2\kappa-1}{1-3\kappa}z_3 + \sin\psi_0 \dfrac{1}{2(1-3\kappa)}z_2^2 \\ \sin\psi_0 \dfrac{1-2\kappa}{1-3\kappa}z_3 + \cos\psi_0 \dfrac{1}{2(1-3\kappa)}z_2^2 \end{bmatrix}, \quad \tilde{\mathbf{g}}_2 = \begin{bmatrix} 0 \\ 1 \\ \dfrac{\kappa}{2\kappa-1}z_1 \\ \sin\psi_0 \dfrac{2\kappa-1}{1-3\kappa}z_3 \\ \cos\psi_0 \dfrac{2\kappa-1}{1-3\kappa}z_3 \end{bmatrix}. \tag{13}$$

The approximate system is nilpotent, controllable, and has the same degree of nonholonomy and the same grow vector as the original system [11, 13]. Also, the approximate system is in triangular form. The latter property facilitates the integration of the approximate dynamics in closed form under some well-defined control inputs.

3 Iterative Steering

Having constructed the nilpotent approximation (12), one can use it for the iterative steering of the original system (5) from the initial $\mathbf{x}_s$ to a desired state $\mathbf{x}_f$. For this purpose, in [9] we used an algorithm with constant iteration step, implemented in the spirit of [11]. In this chapter, we study the globally convergent algorithm with variable iteration step [3, 4].

Define the pseudonorm of the vector of privileged coordinates $\mathbf{z}$ at point $\bar{\mathbf{x}}$:

$$\|\mathbf{z}\|_{\bar{\mathbf{x}}} = |z_1|^{1/w_1} + \cdots + |z_n|^{1/w_n} \tag{14}$$

where the weights $w_1, w_2, \ldots, w_n$ are defined as follows. Let $L^s(\bar{\mathbf{x}})$ be the vector space generated at $\bar{\mathbf{x}}$ by the Lie brackets of $\mathbf{g}_1$ and $\mathbf{g}_2$ of length $\leq s$, $s = 1, 2, \ldots$ and $n_s(\bar{\mathbf{x}}) = \dim L^s(\bar{\mathbf{x}})$, $s = 1, \ldots, r$, where r is smallest integer such that $\dim L^r(\bar{\mathbf{x}}) = 5$. The weight w_i of the coordinates x_i is defined by setting $w_j = s$ if $n_{s-1} < j \leq n_s$, with $n_s = n_s(\bar{\mathbf{x}})$ and $n_0 = 0$.

A formalized description of the algorithm with variable iteration step can be summarized as follows:

1. Set $i = j = k = 0$, $t_j = 0$ and let initial states and parameter be $\mathbf{x}_{0,0} = \mathbf{x}_s$, $\eta_0 = \|\mathbf{z}(\mathbf{x}_s)\|_{\mathbf{x}_f}$.
2. Set $k = k + 1$ and $t_{i+1} = t_i + \Delta T$, where ΔT is the movement time allocated for one iteration. Define the subgoal $\mathbf{x}^d_{j,k} = \mathbf{x}_{j,k} + \mathbf{H}_k(\mathbf{x}_f - \mathbf{x}_{j,k})$, where $\mathbf{x}_{j,k}$ is initial state of each iteration, $\mathbf{H}_k = \mathrm{diag}\{h^1_k, \cdots, h^5_k\}$ and h^m_k are a sufficiently small number defined as below,

$$h^m_k = \max\left\{0, \left(1 - \frac{k\eta_j}{\|\mathbf{z}(\mathbf{x}_{j,0})\|_{\mathbf{x}_f}}\right)^{w_m}\right\}, \; m = 1, \ldots, 5. \tag{15}$$

   where η_j is the adjustable step size and w_m is the weight of the m-th coordinate.
3. Compute from (7) to (11) the image of the subgoal in the privileged coordinates, $\mathbf{z}^d_i$, and construct a controller $\mathbf{u}_i$ that steers the approximate nilpotent system (12) from the origin to $\mathbf{z}^d_i$ in the time interval $t \in [t_i, t_{i+1}]$. This control problem can be solved exactly, and the control law found for the nilpotent system is then applied to the original system (5).
4. If $\|\mathbf{z}(\mathbf{x}(t_i + 1))\|_{\mathbf{x}^d_{i,j}} > \frac{1}{2}\|\mathbf{z}(\mathbf{x}(t_i))\|_{\mathbf{x}^d_{i,j}}$, reduce step size by $\eta_{j+1} = \eta_j/2$ and set $j = j + 1$, $k = 0$ and $\mathbf{x}_{j,0} = \mathbf{x}(t_{i+1})$.
5. Set $i = i + 1$ and return to step 2. Repeat iterative process until state error $\|\mathbf{z}(\mathbf{x}(t_{i+1}))\|_{\mathbf{x}_f}$ becomes smaller than given tolerance e.

As in [9], the particular structure of the control law for steering the nilpotent system (12) is constructed in the spirit of the geometric phase approach [7, 10] because it results to simple calculations and has a clear geometric interpretation. As in [11], the nontrivial maneuver in our construction is divided into two parts corresponding to attaining, respectively, the desired orientation and translation. More specifically, reconfiguring the initial state $\mathbf{x}_s = [0, 0, \psi_s, x_s, y_s]$ to the final one $\mathbf{x}_f = [0, 0, \psi_f, x_f, y_f]$ is described as follows.

3.1 Orientation Part

In the orientation part, we steer ψ to the desired values without changing ϑ, φ and regardless of the values of x and y. The computational procedure can be summarized as follows:

- Set $i = 0$, $t_i = 0$, $\mathbf{x}_i \triangleq [\vartheta_i, \varphi_i, \psi_i, x_i, y_i] = [0, 0, \psi_s, x_s, y_s]$.
- For i-th iteration, set $t_{i+1} = t_i + \Delta T$ and define the subgoal $\psi^d_{j,k} = \psi_{j,k} + h^3_k(\psi_f - \psi_{j,k})$ for $\psi_{j,k} = \psi(t_i)$, where h^3_k is defined by (15). Then compute by (9) its image in the privileged coordinates:

$$\Psi^d_i = h^3_k(\psi_f - \psi_{j,k})/(2\kappa - 1), \tag{16}$$

where $\kappa > 1$ is the inertia ratio given by (3). Let $\omega = 2\pi/\Delta T$. Define the control law by

$$u_1(t) = r_i \cos \sigma_i \omega t, \tag{17}$$

$$u_2(t) = r_i \sin \sigma_i \omega t, \tag{18}$$

where $t \in [t_i, t_{i+1}]$ and $\sigma_i = \text{sign}(\psi_f - \psi_{j,k})$. Geometrically, in the space of the contact coordinates ϑ and φ this control law traces a circle of radius r_i in the direction defined by σ_i:

$$\vartheta(t) = \frac{r_i}{\sigma_i \omega} \sin \sigma_i \omega t, \tag{19}$$

$$\varphi(t) = \frac{r_i}{\sigma_i \omega}(1 - \cos \sigma_i \omega t). \tag{20}$$

Therefore, by the end of the iteration, the contact coordinates ϑ and φ remain unchanged.
By direct integration of the approximate system (12,13) with the control (17,18), it can be shown that

$$z_1(t_{i+1}) = z_2(t_{i+1}) = 0,\ z_3(t_{i+1}) = \sigma_i \pi r_i^2/\omega^2. \tag{21}$$

The free parameter r_i is defined from the condition $z_3(t_{i+1}) = \Psi^d_i$, which results in

$$r_i = \omega \sqrt{|\Psi^d_i|/\pi}. \tag{22}$$

Thus, the control law defined by (17, 18) steers the privileged coordinates z_1, z_2, and z_3 from the origin to, respectively, $0, 0$, and Ψ^d_i. By checking the first three equations of the system (5, 6), one can see that in the original coordinates we have $\vartheta(t_{i+1}) = \varphi(t_{i+1}) = 0$ but $\psi(t_{i+1})$ does not necessarily reach ψ^d_i.
- Set $\mathbf{x}_i = \mathbf{x}_{i+1}$, increase the counter $i = i + 1$ and repeat the calculations until $\psi(t_{i+1})$ reaches a given vicinity of ψ_f.

It should be noted that, since the control law (17, 18) defines a circle in the space of ϑ and φ, one has $|\vartheta(t)| \le r_i/\omega$ and $|\varphi(t)| \le r_i/\omega$. To guarantee that the contact point does not leave the lower hemisphere, one must have

$$\frac{r_i}{\omega} = \sqrt{\frac{h^3_k\,|\psi_f - \psi_{j,k}|}{(2\kappa - 1)\pi}} \le \sqrt{\frac{h^3_k}{2\kappa - 1}} < \pi/2, \tag{23}$$

which gives the following estimate:

$$h_k^3 < (2\kappa - 1)\pi^2/4. \tag{24}$$

Thus, since the inertia ratio $\kappa > 1$, for any $h_k^3 \in [0, 1]$ the control law (17, 18) keeps the contact point away from the singular set $\vartheta = \pm\pi/2$.

3.2 Translation Part

In the translation part of the maneuver, we steer x, y to the desired values x_f, y_f without changing (in the final configuration) $\vartheta = \varphi = 0$ and $\psi = \psi_f$. The computational procedure can be summarized as follows:

- Set $i = 0$, $t_i = 0$, $\mathbf{x}_i = [0, 0, \psi_f, \bar{x}_s, \bar{y}_s]$, where $\bar{x}_s, \bar{y}_s$ are the coordinates of the contact point in the plane attained by the end of the first part of the maneuver.
- Set $t_{i+1} = t_i + 2\Delta T$. Define the subgoal in the contact plane $x_{j,k}^d = x_{j,k} + h_k^4(x_f - x_{j,k})$ and $y_{j,k}^d = y_{j,k} + h_k^5(y_f - y_{j,k})$ for $x_{i,j} = x(t_i)$ and $y_{i,j} = y(t_i)$ and compute the image of the subgoal in the privileged coordinates by (10, 11):

$$X_i^d = \frac{h_k}{R(1-3\kappa)}\left\{\sin\psi_f(x_f - x_{j,k}) + \cos\psi_f(y_f - y_{j,k})\right\}, \tag{25}$$

$$Y_i^d = \frac{h_k}{R(1-3\kappa)}\left\{\cos\psi_f(x_f - x_{j,k}) - \sin\psi_f(y_f - y_{j,k})\right\}. \tag{26}$$

where $h_k = h_k^4 = h_k^5$. Let $\omega = 2\pi/\Delta T$. Define the control law by

$$u_1(t) = r_i \cos(\theta_i + \omega t), \tag{27}$$

$$u_2(t) = r_i \sin(\theta_i + \omega t), \tag{28}$$

for $t \in [2(i-1)\Delta T, (2i-1)\Delta T]$ and

$$u_1(t) = r_i \cos(\theta_i - \omega t), \tag{29}$$

$$u_2(t) = r_i \sin(\theta_i - \omega t), \tag{30}$$

for $t \in [(2i-1)\Delta T, 2i\Delta T]$. Geometrically, this control law defines two symmetric circles in the space of the contact coordinates ϑ and φ:

$$\vartheta(t) = \frac{r_i}{\omega}\{\sin(\theta_i + \omega t) - \sin\theta_i\}, \tag{31}$$

$$\varphi(t) = \frac{r_i}{\omega}\{\cos\theta_i - \sin(\theta_i + \omega t)\}, \tag{32}$$

for $t \in [2(i-1)\Delta T, (2i-1)\Delta T]$ and

$$\vartheta(t) = \frac{r_i}{\omega}\{\sin\theta_i - \sin(\theta + \omega t)\}, \tag{33}$$

$$\varphi(t) = \frac{r_i}{\omega}\{\sin(\theta_i + \omega t) - \cos\theta\}, \tag{34}$$

for $t \in [(2i-1)\Delta T, 2i\Delta T]$. Therefore, the contact variables ϑ and φ are zero at $t = t_i + (2i-1)\Delta T$, and $t = t_i + 2i\Delta T$. Next, since one circle is traced in clockwise direction and the other in the counterclockwise direction, by the end of the iteration the angle ψ remains unchanged.

By direct integration of the approximate system (12, 13) with the control (27, 28) and (29, 30), it can be shown that

$$z_1(t_{i+1}) = z_2(t_{i+1}) = z_3(t_{i+1}) = 0, \tag{35}$$

$$z_4(t_{i+1}) = 2\pi r_i^3 \sin(\psi_f - \theta_i)/\omega^3, \tag{36}$$

$$z_5(t_{i+1}) = 2\pi r_i^3 \cos(\psi_f - \theta_i)/\omega^3. \tag{37}$$

The free parameters r_i and θ_i are defined from the conditions $z_4(t_{i+1}) = X_i^d$ and $z_5(t_{i+1}) = Y_i^d$, which results to

$$r_i = \omega\sqrt[6]{\frac{(X_i^d)^2 + (Y_i^d)^2}{4\pi^2}}, \tag{38}$$

$$\theta_i = \psi_f - \arctan(X_i^d / Y_i^d). \tag{39}$$

Thus, the control law defined by (27, 28) and (29, 30), steers the state variables of the approximate nilpotent system (12) from the origin to, respectively, $[0, 0, 0, X_i^d, Y_i^d]$. By the geometric construction, for the original state variables we have $\vartheta(t_{i+1}) = \varphi(t_{i+1}) = 0$ and $\psi(t_{i+1}) = \psi_f$, but $x(t_{i+1})$ and $y(t_{i+1})$ do not necessarily reach $x_{j,k}^d$ and $y_{j,k}^d$.

- Set $\mathbf{x}_i = \mathbf{x}_{i+1}$, increase the counter $i = i + 1$ and repeat the calculations until the coordinates of the contact point in the plane $[x(t_{i+1}), y(t_{i+1})]$ reach a given vicinity of $[x_f, y_f]$.

Under the control law (27, 28) and (29, 30), the contact point in the space of the contact coordinates ϑ and φ is always on the circle of radius r_i/ω, with $|\vartheta(t)| \leq r_i/\omega$ and $|\varphi(t)| \leq r_i/\omega$. To ensure that the contact point does not leave the lower hemisphere, one needs to have

$$\frac{r_i}{\omega} = \sqrt[6]{\frac{(X_i^d)^2 + (Y_i^d)^2}{4\pi^2}} < \frac{\pi}{2}. \tag{40}$$

By transforming this inequality with the use of (25, 26), one obtains the following estimate:

$$h_k < \bar{h}_k = \frac{\pi^4 R(3\kappa - 1)}{32\sqrt{(x_f - x_{j,k})^2 + (y_f - y_{j,k})^2}}. \tag{41}$$

Thus, if h_k is selected in accordance with (41) the trajectory of the contact point on the sphere is kept below the singular set $\vartheta = \pm\pi/2$. If h_k defined by (15) happens to be larger than $\bar{h}_k$, one needs to replace $h_k = \bar{h}_k - \epsilon$, where ϵ is a small constant.

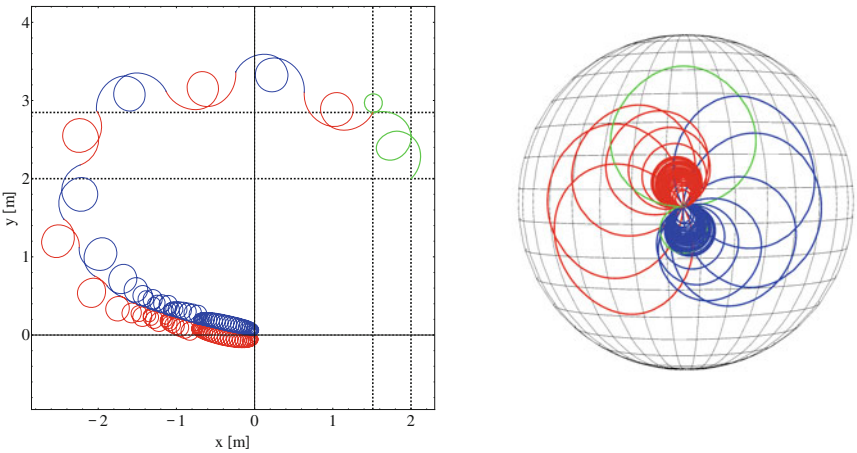


Fig. 2 Trajectory of the contact point on the plane (*top*) and on the sphere (*bottom*) during the nontrivial maneuver for $\kappa = 2.5$. The first part of the maneuver is shown in *green* color, while the second part is shown in *red* and *blue* colors

4 Simulation

The performance and effectiveness of the global steering algorithm are tested under simulation. In the simulation example, as in [9], we drive the system from the initial state $\mathbf{x}_0 = \{0, 0, 3\pi/4, 2.0, 2.0\}$ to the final state $\mathbf{x}_f = \{0, 0, 0, 0, 0\}$, for $\kappa = 2.5$ and $\kappa = 10$. The units of all the dimensional quantities are specified in the International System of Units (SI) system.

For $\kappa = 2.5$, the simulation results (the trajectories of the contact point on the plane and on the sphere) are shown in Fig. 2. In the orientation part of the maneuver, where we drive ψ to zero regardless of the attained x and y, the error between the current and target values of ψ becomes less than 0.001 after three iterations. By the end of the orientation part of the maneuver the values of x and y change, respectively, from 2.0 to 1.51 and from 2.0 to 2.85. In the translation part of the maneuver, where x and y are driven to the origin, it takes 53 iterations for the error between the current and target position of the contact point on the plane, defined by the Euclidian distance $\sqrt{(x(t) - x_f)^2 + (y(t) - y_f)^2}$, to become less than 0.001.

For $\kappa = 10$, the simulation results are shown in Fig. 3. In the orientation part of the maneuver, the error between the current and target values of ψ becomes less than 0.001 after two iterations. By the end of the orientation part of the maneuver the values of x and y change, respectively, from 2.0 to 1.84 and from 2.0 to 2.38. In the translation part of the maneuver, it takes 64 iterations for the error between the current and target position of the contact point on the plane to become less than 0.001.

It is interesting to compare the performance of the iterative algorithm with adjustable step size with that with the constant step size. For the exactly same simulation

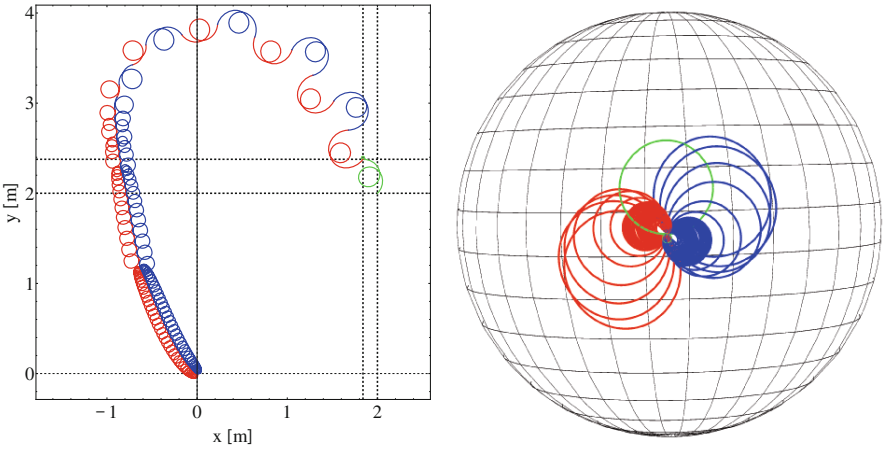


Fig. 3 Trajectory of the contact point on the plane (*top*) and on the sphere (*bottom*) during the nontrivial maneuver for $\kappa = 10$. The first part of the maneuver is shown in *green* color, while the second part is shown in *red* and *blue* colors

example in [9], for $\kappa = 2.5$ there were five iterations in the orientation part and ten iterations in the translation part. As to $\kappa = 10$, there were 2 iterations in the orientation part and 29 iterations in the translation part. So, it appears that in the translation part the algorithm with the constant step can be more efficient in terms of the number of iterations necessary to reach the goal state.

It should be noted that the efficiency of the steering algorithms under consideration depends not only on κ but also on the initial state $\mathbf{x}_0$. To illustrate this point, we conduct another simulation where we set $\mathbf{x}_0 = \{0, 0, -\pi/2, 2.0, 2.0\}$ while keeping the target state at the origin. Simulation results for the algorithm with adjustable η and with constant $\eta = 0.6$ are shown in, respectively, Figs. 4 and 5. The algorithm with adjustable η produces 3 iterations in the orientation part and 12 iterations in the translation part, while the algorithm with constant η does 3 in the orientation part and 9 iterations in the translation part.

It should be pointed out that the algorithm with the adjustable step size converges globally while in the algorithm with the constant step size η does not converge for any $\eta \in [0, 1]$, so η should be set as a sufficiently small number. Setting η is done manually, and the smaller the η the larger the number of iterations. Finding optimal value of $\eta = \eta^\star$, resulting to minimal number of iterations, is a very tedious and computationally involving procedure that can be done only by tuning of η. If, instead of tuning, one would juts set η within a safe margin it would decrease the steering performance. For example, if in the last simulation one would set $\eta = 0.1$, the number of iterations in the orientation part would increase to 15 and the number of iterations in the translation part would increase to 76. The corresponding simulation results for this case are shown in Fig. 6.

On the other hand, the algorithm with the adjustable step size does not require any tuning and therefore is more convenient in use. However, the adjustment strategy

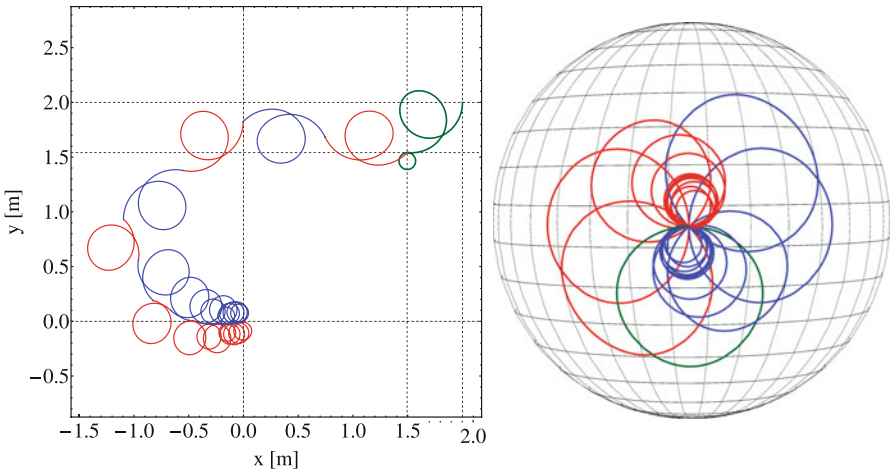


Fig. 4 Trajectory of the contact point on the plane (*top*) and on the sphere (*bottom*) during the nontrivial maneuver for $\kappa = 2.5$. The first part of the maneuver is shown in *green* color, while the second part is shown in *red* and *blue* colors

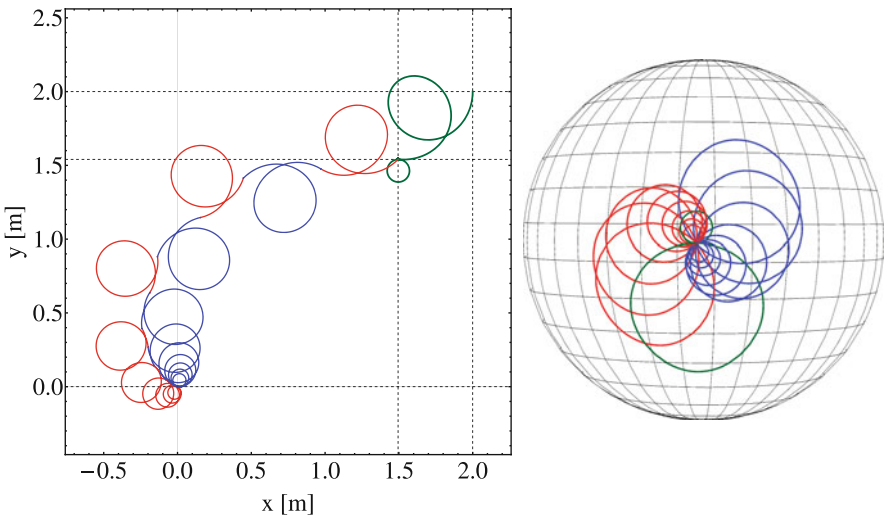


Fig. 5 Trajectory of the contact point on the plane (*top*) and on the sphere (*bottom*) during the nontrivial maneuver for $\kappa = 2.5$ produced by the algorithm with constant $\eta = 0.6$. The first part of the maneuver is shown in *green* color, while the second part is shown in *red* and *blue* colors

does not necessarily results to efficient steering. While the thorough investigation of the steering algorithms remains the subject of future work, it can be conjectured that if, instead of setting $\eta_0 = \|\mathbf{z}(\mathbf{x}_s)\|_{\mathbf{x}_f}$, we select $\eta_0 = \eta^\star$ the performance (the convergence rate) of the two steering algorithms under comparison will be nearly the same.

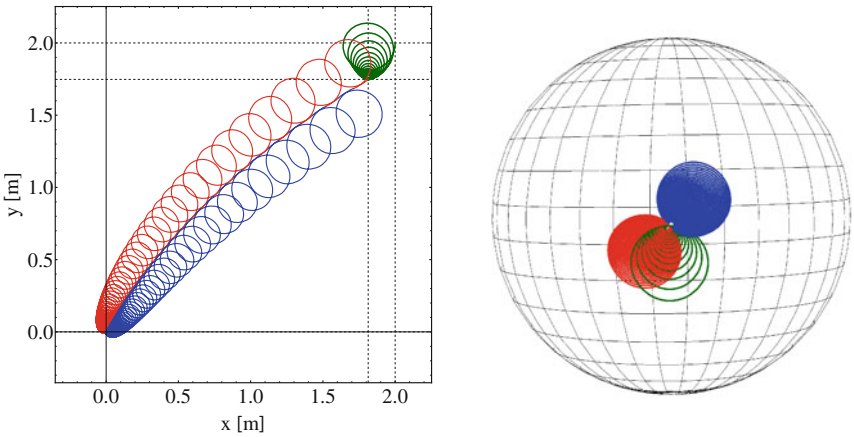


Fig. 6 Trajectory of the contact point on the plane (*top*) and on the sphere (*bottom*) during the nontrivial maneuver for $\kappa = 2.5$ produced by the algorithm with constant $\eta = 0.1$. The first part of the maneuver is shown in *green* color, while the second part is shown in *red* and *blue* colors

5 Conclusions

Motion planning for a spherical rolling robot, actuated by two internal rotors that are placed on orthogonal axes has been studied in this chapter. The mathematical model of the robot, represented by a driftless control system, contains a physical singularity corresponding to the motion of the contact point along the equatorial line in the plane of the two rotors. It has been shown that the technique of steering through the singularity by finding a globally regular valid basis is not applicable to the driftless control system. To solve the state-to-state transfer problem, a globally convergent steering algorithm has been implemented and tested. It has been shown in simulation example that its convergence rate is not always superior to the algorithm with constant iteration step size.

References

1. Alouges, F., Chitour, Y., Long, R.: A motion-planning algorithm for the rolling-body problem. IEEE Trans. Robot. **26**(5), 827–836 (2010)
2. Bhattacharya, S., Agrawal, S.K.: Spherical rolling robot: A design and motion planning studies. IEEE Trans. Robot. Autom. **16**(6), 835–839 (2000)
3. Chitour, Y., Jean, F., Long, R.. A global steering method for nonholonomic systems. J. Differ. Equations **254**, 1903–1956 (2013)
4. Jean, F., Oriolo, G., Vendittelli, M.: A global convergent steering algorithm for regular non-holonomic systems. In: Proc. IEEE Int. Conference on Decision and Control, pp. 7514–7519. Seville, Spain, 12–15 Dec 2005
5. Joshi, V.A., Banavar, R.N.: Motion analysis of a spherical mobile robot. Robotica **27**(3), 343–353 (2009)

6. Lafferriere, G., Sussmann, H.J.: A differential geometric approach to motion planning. In: Li, Z., Canny, J.F. (eds.) Nonholonomic motion planning, pp. 235–270. Kluwer, Norwell, MA (1993)
7. Li, Z., Canny, J.: Motion of two rigid bodies with rolling constraint. IEEE Trans. Robot. Autom. **6**(1), 62–72 (1990)
8. Marigo, A., Bicchi, A.: Rolling bodies with regular surface: Controllability theory and applications. IEEE Trans. Autom. Control **45**(9), 1586–1599 2000
9. Morinaga, A., Svinin, M., Yamamoto, M.: A motion planning strategy for a spherical rolling robot driven by two internal rotors. IEEE Trans. Robot. **30**(4) (2014, in press)
10. Mukherjee, R., Minor, M.A., Pukrushpan, J.T.: Motion planning for a spherical mobile robot: Revisiting the classical ball-plate problem. ASME J. Dyn. Syst. Meas. Control **124**(4), 502–511 (2002)
11. Oriolo, G., Vendittelli, M.: A framework for the stabilization of general nonholonomic systems with an application to the plate-ball mechanism. IEEE Trans. Robot. **21**(2), 162–175 (2005)
12. Svinin, M., Morinaga, A., Yamamoto, M.: On the dynamic model and motion planning for a spherical rolling robot actuated by orthogonal internal rotors. Regul. Chaotic Dyn. **18**(1–2), 126–143 (2013)
13. Vendittelli, M., Oriolo, G., Jean, F., Laumond, J.-P.: Nonhomogeneous nilpotent approximations for nonholonomic systems with singularities. IEEE Trans. Autom. Control **49**(2), 261–266 (2004)
14. Vendittelli, M., Oriolo, G., Jean, F., Laumond, J.-P.: Steering nonholonomic systems via nilpotent approximations: The general two-trailer system. In: Proc. IEEE Int. Conference on Robotics and Automation, vol. 2, pp. 823–829. Detroit, Michigan, 10–15 May 1999

Odontological Information Along Cone Splines

Cindy González and Marco Paluszny

Abstract Developable surfaces are a subset of ruled surfaces, which can be mapped onto a plane without deformation. Due to this property, they have considerable relevance in several applications. In the medical area, regarding information visualization along sections of organs, they could be useful in clinical diagnosis. They have also industrial applications, including footwear and clothing industries, where three-dimensional (3D) designs are made from flat materials.

In this research, we consider the issue of approximating developable surfaces with segments of circular cones, with the aim of constructing splines that model interesting surfaces. Our emphasis will be in the odontological area. We present examples of "panoramic views" of curved sections of human jaw which contain information about all the dental pieces. Moreover, the process allows for the simultaneous display of these pieces in a flat surface, without metric distortion.

Keywords Developable surface · DICOM volume · Segment of circular cone · Rational Bézier curve

1 Problem of Visualization from Volumes

The goal of the Visible Human Project [8] is to make detailed information about human anatomy accessible to the scientific community. It is a database that consists of 1871 horizontal plane sections from which "photos" of oblique slices—that can be extracted from the data volume—can be reconstructed. This process is mathematically well-known: it involves the trilinear interpolation [9]. In fact, the Visible

C. González (✉)
Université de Valenciennes et du Hainaut-Cambrésis, Le Mont Houy, LAMAV, 59313 Valenciennes cedex 9, France
e-mail: Cindy.Gonzalez@univ-valenciennes.fr

M. Paluszny
Escuela de Matemáticas, Universidad Nacional de Colombia, sede Medellín Calle 59A No.63-20, Medellín, Colombia
e-mail: mpalusznyk@unal.edu.co

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_15

Human portal offers a computational tool for extracting information along plane slices of arbitrary three-dimensional (3D) position [4].

A more interesting problem is the information extraction along curved slices, for instance, along an artery (in order to determine calcifications or other malformations), or along a jaw bone (with the purpose of making visible the information of several contiguous). The deployment of such information is potentially useful in surgical planning. In general, when a surface is displayed in a plane, some of the areas of the slice need to be stretched and this generates a distortion problem: The shape and/or size of the original organ, along a particular section, can differ from the shape and/or size when it is displayed in a plane screen. It is the same problem of deformation that arises when building maps.

1.1 Preliminary Aspects

Developable surfaces are special case of ruled surfaces, which can be unfolded or developed onto a plane without stretching or tearing. Mathematically speaking, developable surfaces are surfaces characterized by the property of possessing the same tangent plane at all points of the same ruling.

The ruled surfaces are defined by

$$\mathbf{s}(u, v) = l(u) + v\mathbf{e}(u),$$

where $l(u)$ is a curve on the surface, called the directrix and $\mathbf{e}(u)$ are unit vectors of the generator lines.

Among developable surfaces are conical surfaces, cylindrical surfaces, and tangent surfaces, we shall focus on circular cones. Due to the property can be isometrically mapped into the plane, the developable surfaces are interesting for visualization purposes.

Particularly, in the medical area, the problem of flattening a surface without stretching was considered by Saroul in his PhD thesis [7] and in a series of articles in scientific journals [3, 6]. Saroul minimizes the deformation in an area specified by the user at the expense of other areas (which might be less relevant or interesting) where the deformation is not controlled. Saroul's proposal has an inherently local nature [7]. In a recent article, Figuereido and Hersch [3] propose the extraction of information about 3D volumes contained in cylinders built on plane curves. The aforementioned can be generalized to developable surfaces whose construction, from the computer aided geometric design point of view, has been studied by Aumann [1]. The main difficulty with Aumann's method is the numerical nature regarding the flat presentation of the surface. Such difficulty disappears if we use cylinders and cones, instead of general developable surfaces. In this work, we use techniques developed by Pottmann and Leopoldseder [5], based on an article by Fuhs and Stachel [2], for constructing curved slices with cone splines. Cone splines are splines constructed by joining segments of cones with tangent continuity along common generators. For this construction we must take care to exclude cone vertices. An illustrative example is shown using a human jaw bone, and the slice is constructed with segments of circular cones.

2 Circular Cones with Prescribed Contacts

Leopoldseder and Pottmann presented in [5] an algorithm for the approximation of a given developable surface Γ by a cone spline surface and it depends on an adequate choice of generators on this surface. The authors leave the problem of finding the best choice of generators as an open question. The algorithm is described in Sects. 2.1 and 2.2.

Our problem is different, we have "in principle"[1] a fixed sequence of line/plane pairs as given contacts, one for each tooth. This sequence is extracted from the maxillary tomography. We use the technique from [5] to build a cone spline with these or most of these planes/lines. The spline exhibits all neighboring teeth in a given jaw region.

2.1 Contact Elements

Given a developable surface, we select a sequence of generators e_i of Γ and calculate its tangent planes τ_i. We refer to (e_i, τ_i) as *contact elements*. From a sequence of contact elements, we want to build cone segments to model surfaces that interpolate the given information. We will interpolate a couple of consecutive contact elements with two circular cone segments, which have the same tangent plane along a common generator.

For each (e_i, τ_i), $i = 1, 2$, there is an orthonormal basis coherent with Γ. For each contact element (e_i, τ_i) abusing notation, we use $\mathbf{e}_i$ to refer to the unit director along the generator e_i. Let $(\mathbf{e}_i, \mathbf{p}_i)$ be the orthonormal basis for τ_i such that $\mathbf{n}_i = \mathbf{e}_i \times \mathbf{p}_i$ is the unit normal of Γ.

2.2 General Algorithm

Given two consecutive contact elements (e_1, τ_1) and (e_2, τ_2), we want to find two cones of revolution Δ_1, Δ_2, with different vertices $\mathbf{v}_1, \mathbf{v}_2$, that have a common generator and the same tangent plane along this generator. Each Δ_i must also contain the generator e_i, and its tangent plane along e_i must match up with τ_i. The axes of a pair of cones in this position either intersect in a point $\mathbf{m}$ or are parallel. We consider the case in which the axes intersect in $\mathbf{m}$ (see Fig. 1).

[1] Actually for multiple root teeth, there might be more than one plane choices. These stem from the fact the best plane for clinical inspection could approximate any of the root pairs.

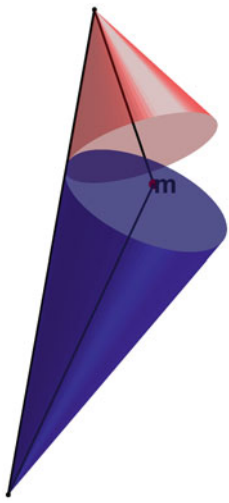


Fig. 1 Circular cones with a comun generator and their axes intersect at a point **m**

It can be shown [5] that the pair of cones described in the previous paragraph have an inscribed sphere Σ, whose center is the point **m**. This sphere touches both cones along two circles: c_1 and c_2 (see Fig. 2). The sphere Σ is determined from the two consecutive contact elements (e_i, τ_i). If $\mathbf{m}_1$ and $\mathbf{m}_2$ are points of the generators e_1 and e_2, the point **m** is the intersection of the normal planes:

$$\gamma_i : (\mathbf{x} - \mathbf{m}_i) \cdot \mathbf{p}_i = 0, \; i = 1, 2, \tag{1}$$

with the bisector plane of the two tangent planes:

$$\sigma : \mathbf{x} \cdot (\mathbf{n}_1 - \mathbf{n}_2) - \mathbf{m}_1 \cdot \mathbf{n}_1 + \mathbf{m}_2 \cdot \mathbf{n}_1 = 0. \tag{2}$$

Each one of the cones Δ_1, Δ_2, will touch the sphere Σ along a circle, and these circles will be tangentially connected at the point **c**. The circles c_1 and c_2 allow us to construct a biarc which connects, with tangential continuity in **c**, a segment of the circle c_1 with a segment of the circle c_2 (see Fig. 3). This joins generator e_1 with generator e_2.

Given a set of $n + 1$ control points $\mathbf{b}_0, \ldots, \mathbf{b}_n$, each one associated with a scalar ω_i called weight, a degree-n rational Bézier curve is defined by

$$c(t) = \frac{\omega_0 B_0^n(t)\mathbf{b}_0 + \ldots + \omega_n B_n^n(t)\mathbf{b}_n}{\omega_0 B_0^n(t) + \ldots + \omega_n B_n^n(t)},$$

where $B_i^n(t)$ are well-known Bernstein polynomials.

For a rational Bézier representation for the biarc, we will denote its control points by $\mathbf{a}_1$, $\mathbf{b}_1$, $\mathbf{c}$, $\mathbf{b}_2$, $\mathbf{a}_2$, (Fig. 3). Let $\mathbf{b}_1 = \mathbf{a}_1 + \lambda_1 \mathbf{p}_1$ and $\mathbf{b}_2 = \mathbf{a}_2 - \lambda_2 \mathbf{p}_2$.

The point $\mathbf{a}_i$ is the intersection point between the sphere Σ and the generator e_i, such that the vector $\mathbf{p}_i$ is the vector tangent to the spherical biarc at this point.

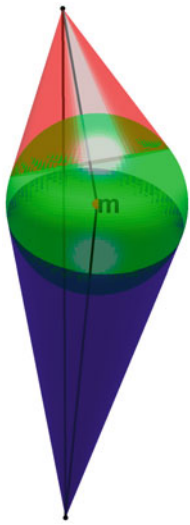


Fig. 2 Two-sphere Σ inscribed in the circular cones Δ_1 and Δ_2

The control polygon of a circle has the shape of an isosceles triangle; the segment of the *internal control points,* $\mathbf{b}_1, \mathbf{b}_2$, satisfy the condition:

$$\|\mathbf{b}_2 - \mathbf{b}_1\|^2 = (\lambda_1 + \lambda_2)^2. \tag{3}$$

This is equivalent to

$$(\mathbf{a}_2 - \mathbf{a}_1)^2 - 2\lambda_1(\mathbf{a}_2 - \mathbf{a}_1) \cdot \mathbf{p}_1 - 2\lambda_2(\mathbf{a}_2 - \mathbf{a}_1) \cdot \mathbf{p}_2 + 2\lambda_1\lambda_2(\mathbf{e}_1 \cdot \mathbf{e}_2 - 1) = 0. \tag{4}$$

Thus, if we choose λ_1 using Eq. 4 we can calculate λ_2. For the construction of the biarc, the contact point $\mathbf{c}$ is given by

$$\mathbf{c} = \frac{\lambda_2\mathbf{b}_1 + \lambda_1\mathbf{b}_2}{\lambda_1 + \lambda_2}. \tag{5}$$

So, setting the weights at the end points of the two arcs to 1, we can express the Bézier rational quadratic form of each circular arc as follows:

$$c_1(t) = \frac{\mathbf{a}_1(1-t)^2 + \omega_{11}\mathbf{b}_1 2t(1-t) + \mathbf{c}t^2}{(1-t)^2 + \omega_{11}2t(1-t) + t^2}, \tag{6}$$

$$c_2(t) = \frac{\mathbf{c}(1-t)^2 + \omega_{12}\mathbf{b}_2 2t(1-t) + \mathbf{a}_2 t^2}{(1-t)^2 + \omega_{12}2t(1-t) + t^2}, \tag{7}$$

where the weights ω_{1i} associated to the internal control points are given by:

$$\omega_{1i} = \frac{|(\mathbf{b}_i - \mathbf{a}_i)(\mathbf{c} - \mathbf{a}_i)|}{\|\mathbf{b}_i - \mathbf{a}_i\| \, \|\mathbf{c} - \mathbf{a}_i\|}. \tag{8}$$

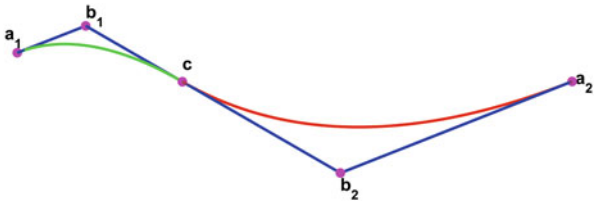


Fig. 3 Biarc with its control polygon

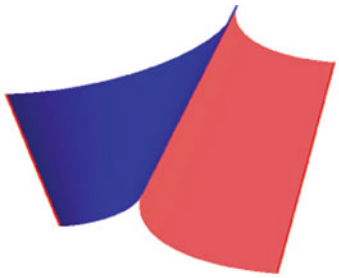


Fig. 4 Example of a cone pair with sharp edges

If λ_i is positive, the arc with control points $\mathbf{a}_i$, $\mathbf{b}_i$, $\mathbf{c}$ and weights $\omega_{1i} > 0$ are used. On the contrary, the complementary arc and a negative weight ω_{1i} may be used.

Finally, once found the biarc as was described above, we can compute the vertices $\mathbf{v}_i$ of the two cone segments. These are calculated as the intersection of the tangent plane to the sphere Σ at point $\mathbf{c}$ with generators e_i. The axes of the cone pair are the lines that pass through $\mathbf{v}_i$ and center $\mathbf{m}$ of Σ, respectively.

In order to calculate a cone pair within the one parameter set of solutions, we may choose λ_1 and compute λ_2 as explained above. The parameter λ_1 allows for the adjustment of each pair of cone segments according to visualization needs.

Since sometimes the cones of a pair lie locally on different sides of their common tangent plane, this might lead either to sharp edges in the cone spline (Fig. 4) or to an s-shaped cone pair (Fig. 5). In the first case the complementary arc has to be used, which is obtained by changing the sign of the weight of the control point $\mathbf{b}_i$. This guarantees a smooth surface but produces sharp edges in the biarc.

In the last case, the solution that works well is "to jump over"[2] the offending contact element.

3 Curved Slices Constructed with Segments of Circular Cones

Medical images are produced with various techniques. Especially interesting are those yielding sequences of parallel slices of a 3D volume. Examples of these are sequences of Digital Imaging and Communications in Medicine (DICOM) files

[2] If we have two adjacent contact elements (e_i, τ_i) and (e_{i+1}, τ_{i+1}), "jumping over a contact element" means omitting (e_{i+1}, τ_{i+1}) and considering instead (e_i, τ_i) and (e_{i+2}, τ_{i+2}).

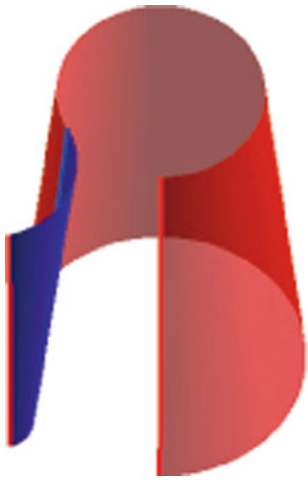


Fig. 5 Example of an s-shaped cone pair

generated by computed tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI). Figure 6 shows two slices, of RMI and CT, see https://mri.radiology.uiowa.edu/visible_human_datasets.html for additional examples.

DICOM format includes a file with image and patient data, also it contains a network communications protocol that allows its exchange with the data and quality necessary for clinical use. In this work, we used the Matlab platform, which allows for DICOM file reading.

In this section, we consider a medical volume, which stores information about the internal and external structure of the 16 dental pieces of a human upper jaw bone. The volume is assembled from a sequence of DICOM files acquired through computer axial tomography.

For each tooth, we find a contact element displaying the information of clinical interest. Given the segmentation of a tooth the choice of the plane depends on the position of the roots that need to be visualized.

The line joins the center point of the tooth enamel surface with a midpoint of the two chosen roots.

We will apply the aforementioned construction of circular cones from biarcs to modeling a curved slice that contains information about the internal structure of a sequence of neighboring teeth in a jaw bone. Figure 7 shows the selected plane of a tooth which has been texturized with the corresponding information extracted from the volume, using the trilinear interpolation process [9].

Figure 8 illustrates pairs of segments of cones for different values of λ_1, which were built for the same pair of contiguous contact elements.

Figure 9a illustrates a curved slice constructed with a sequence of segments of cones, manually choosing parameter values λ_1 to allow a good overall view of the teeth of the upper jaw bone. Figure 9b displays its flattened version.

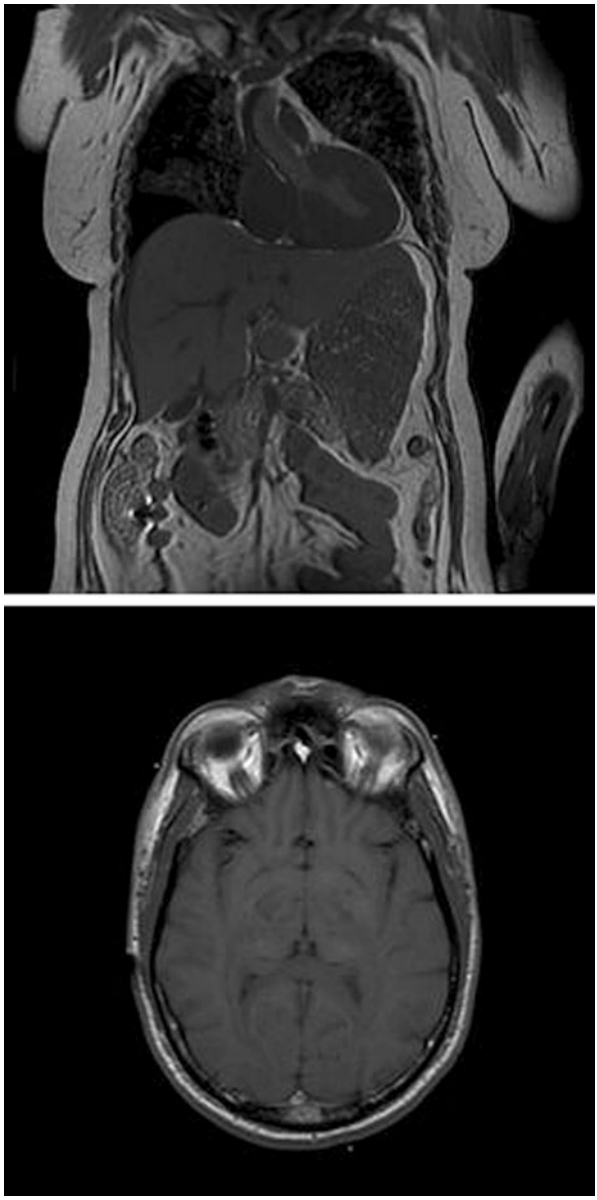


Fig. 6 Risk of malignancy index (RMI) slice of the body of a man and computed tomography (CT) slice of the head of a woman

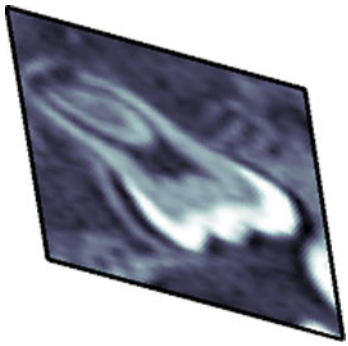


Fig. 7 Texturized plane of a dental piece from the upper jaw bone

4 Analysis and Conclusions

We present a technique for extracting information along curved slices (which can be flattened without deformation) from dental data volumes. Figure 10 illustrates three views of the curved slice of Fig. 9a, which also shows the generators used to construct the cone spline.

The illustrations in Figs. 11 and 12 show the information relating to the same dental pieces, using the technique of Aumann and cone splines, respectively. Under visual inspection both techniques yield results of comparable quality. The main advantage of the approach with circular cones is its mathematical simplicity as compared to developable surfaces technique of Aumann [1]: the flattening process is simpler for circular cones than for general developable surfaces. One limitation of the cone splines technique of approximation with circular cones is that special care is necessary to avoid the possible presence of cone vertices within the approximating curved slice, which would then be singular points of the cone spline. Singularities might also pop up in the case of developable surfaces, but in this case they are easier to avoid because this family of surfaces is larger. One way to handle the problem is to allow noncircular cones or cylinders when there is not a good approximation with circular cones. Another, easier and possibly sufficient way in many applications is to jump a generator where the cones constructed are not acceptable (see Fig. 10).

Paluszny [6] uses a technique of a generalization of the Aumann [1] to construct a developable surface cutting the 16 teeth of the upper human jaw bone.

Let $\mathbf{b}_0, \ldots, \mathbf{b}_n$ be the contact points of the polynomial Bézier curve $\mathbf{b}(t)$.

Aumann's algorithm produces a Bézier curve $\mathbf{c}(t)$, with control points $\mathbf{c}_0, \ldots, \mathbf{c}_n$ such that the ruled surface obtained by joining $\mathbf{b}(t)$ and $\mathbf{c}(t)$ is developable.

The point $\mathbf{c}_0$ is arbitrary but $\mathbf{b}_0$, $\mathbf{b}_1$, and $\mathbf{c}_0$ are not collinear and for $i = 0, \ldots n-1$,

$$\mathbf{c}_{i+1} = \mathbf{b}_i + \lambda(\mathbf{b}_{i+1} - \mathbf{b}_i) + \mu(\mathbf{c}_i - \mathbf{b}_i),$$

where λ and μ are arbitrary parameters.

In [6], one of the authors extends Aumann's construction for polynomial Bézier curves to Catmull–Rom–Overhauser interpolary tangent continous splines.

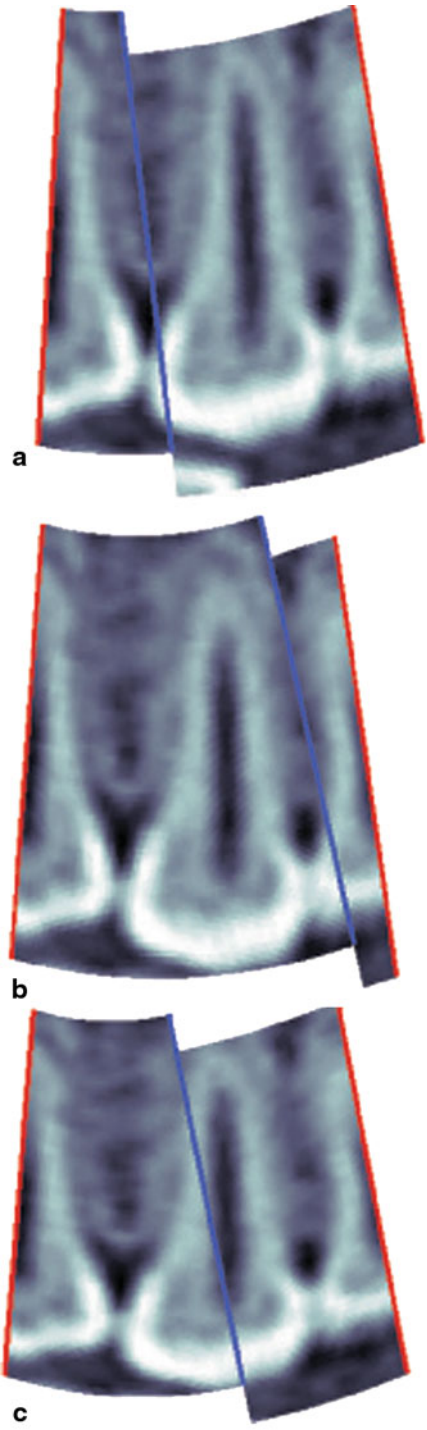


Fig. 8 Pairs of segments of cones, which correspond to values $\lambda_1 = 10, 16,$ and 23, respectively

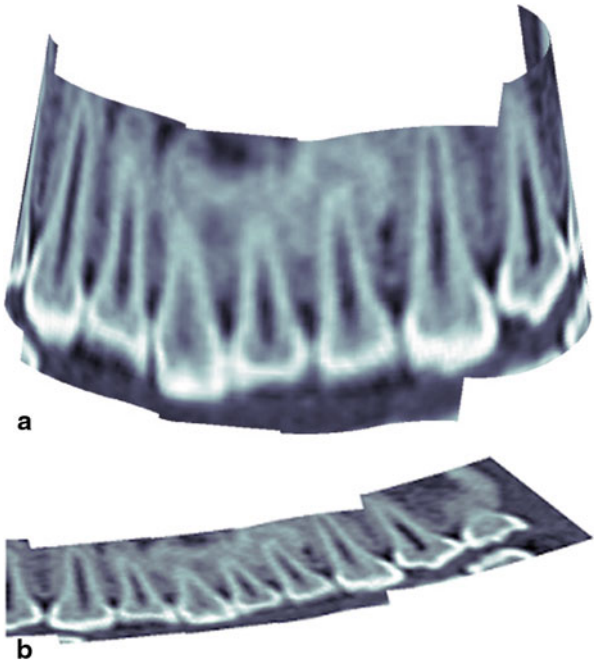


Fig. 9 **a** Three-dimensional (3D) view of the construction of a curved slice with segments of cone splines for some dental pieces of the upper jaw, **b** view of development without distortion on the slice

The main differences between Aumann-based approach and the technique of cone splines are:

- Aumann's developable is built along a prescribed interpolatory curve and depends globally on five parameters (the positions of $\mathbf{c}_0$, λ, and μ). The cone spline interpolates a sequence of contact elements (plane/line) and each segment can be adjusted independently while preserving the tangent continuity property at common generators. In other words: provides local control.
- Both surfaces can be unfolded isometrically onto the plane. This process is simpler for cone splines.
- The cone spline may contain segments which are not faithful to the teeth sequence, hence the "jumping over" technique will provide a solution. Aumann technique does not allow this solution of such a problem because given the Bézier curve $\mathbf{b}(t)$, the point $\mathbf{c}_0$ and the parameters λ and μ, the developable surface is completely determined.

Within the medical field other possible fields of application of the technique include the construction of curved sections of veins (to study valve function) and arteries (for detecting calcifications). Within industry we envision applications in the field of study of fractures in volumes.

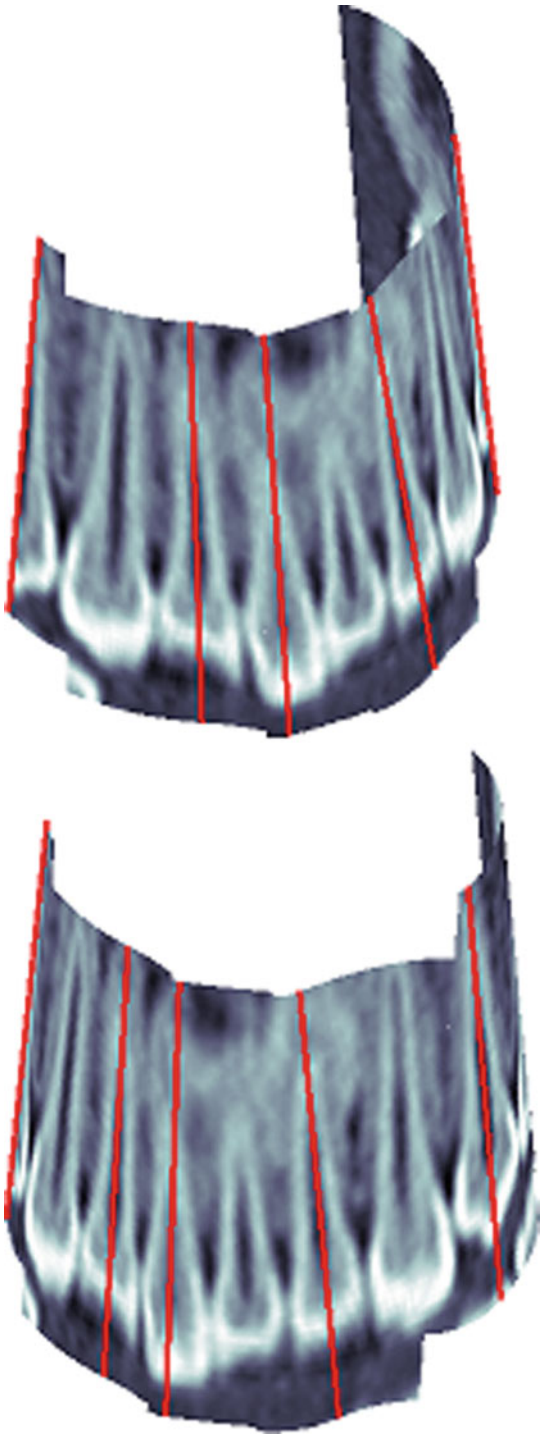


Fig. 10 Three-dimensional (3D) views of the curved slice constructed with cone splines "jumping over" some dental pieces of upper jaw bone

Fig. 11 Examples of teeth using the technique of Aumann

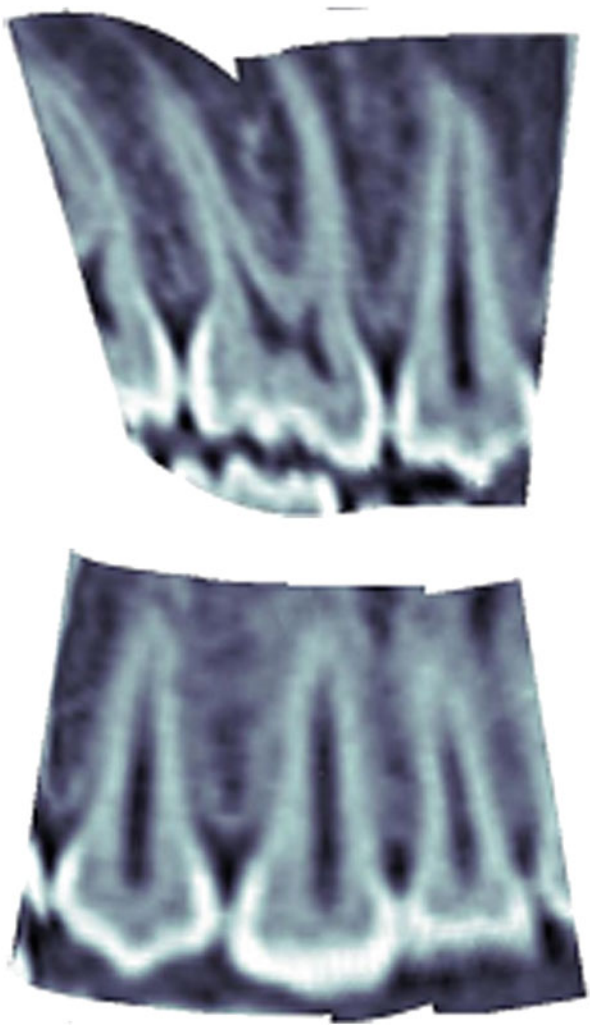


Fig. 12 Examples of dental pieces using segments of circular cones

Acknowledgement We wish to thank José F. Ramírez Huaca for Fig. 11.

References

1. Aumann, G.: A simple algorithm for designing developable Bézier surfaces. Comput. Aided Geom. Des. **20**, 601–619 (2003)
2. Fuhs, W., Stachel, H.: Circular pipe connections. Comput. Graph. **12**, 53–57 (1988)
3. Hersch, R., Figueiredo, O.: Parallel unfolding visualization of curved surfaces extracted from large three-dimensional volumes. J. Electr. Imaging **11**, 423–433 (2002)
4. Hersch, R.,Gerlach, S.: Exploring anatomic structures with EPFL's visible human web server. http://www.ibrarian.net/navon/page.jsp?paperid=13820841

5. Leopoldseder, S., Pottmann, H.: Approximation of developable surfaces with cone spline surfaces. Comput. Aided Des. **30**, 571–582 (1998)
6. Paluszny, M.: Between developable surfaces and circular cone splines: curved slices of 3D volumes. Proc. SPIE, vol. 7964. Medical Imaging 2011: Visualization, image-guided procedures, and modeling, 1 Mar 2011
7. Saroul, L.: Surface extraction and flattening for anatomical visualization. PhD Thesis, École Polytechnique Fédérale de Lausanne, Faculté Informatique et Communications (2006)
8. Wikipedia The Free Encyclopedia: Visible human project. http://en.wikipedia.org/wiki/Visible$_{-}$Human$_{-}$Project (2011). Accessed 18 Sept 2011
9. Wikipedia The Free Encyclopedia: Trilinear interpolation. http://en.wikipedia.org/wiki/Trilinear$_{-}$interpolation (2011). Accessed 5 July 2011

Modeling Cell Decisions in Bone Formation

Rodrigo Assar, Alejandro Maass, Joaquín Fernández, Ernesto Kofman and Martín A. Montecino

Abstract The process of bone formation involves several mechanisms, which can manifest dysfunctions such as osteoporosis in case of imbalances between them. In basic terms, osteo-adipo progenitors derive from the bone marrow, and depending on multiple stimulus signals, can stay in their progenitor state (preosteoblast) or can differentiate to form bone and fat tissue [3]. We point to model the dynamics of the cell decisions to differentiate from preosteoblasts to osteoblasts, considering stimulatory signals, and the important role of epigenetics. Given a cell, the presence of specific epigenetic marks favors the expression of biomarker genes and the posterior differentiation into osteoblasts. Starting with a group of marked cells, we model in silico the proliferation of such cells and the epigenetic inheritance. We consider a hybrid system [2, 8] in which each cell grows continuously over time until being ready to divide, and the success in division and epigenetic inheritance includes randomness. Stimulating the proliferation of marked cells, the model predicts the dynamics to increase the number of osteoblasts helping in testing medical treatments and production in vitro.

Keywords Cell decisions · Bone formation · Hybrid systems

R. Assar (✉)
ICBM Escuela de Medicina, Universidad de Chile, Santiago, Chile
e-mail: rodrigo.assar@gmail.com

A. Maass
Departamento de Ingeniería Matemática, Universidad de Chile, Santiago, Chile
e-mail: amaass@dim.uchile.cl

J. Fernández
Departamento de Control, FCEIA, Universidad Nacional de Rosario,
CIFASIS-CONICET, Rosario, Argentina
e-mail: joaquin.f.fernandez@gmail.com

E. Kofman
Departamento de Control, FCEIA, Universidad Nacional de Rosario,
CIFASIS-CONICET, Rosario, Argentina
e-mail: ekofman@gmail.com

M. A. Montecino
Centro de Investigaciones Biomédicas, Universidad Andrés Bello, Santiago, Chile
e-mail: mmontecino@unab.cl

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_16

1 Introduction

The aim of this work is to model the dynamics of bone formation by considering the cells decisions leading to bone formation. The process of bone tissue formation is given by the delicate interaction between bone formation and resorption. At the cellular level, at any time the cells have to decide whether to proliferate, to differentiate, or to perform apoptosis. Many stimulatory signals affect such cell decisions, and small variations in their conditions provoke important changes in the dynamics. With accurate models of the bone formation dynamics, we want to find and validate new treatments for bone mass disorders such as osteoporosis. We aim to obtain treatments with less limitations, less side effects, and more suited to patient-specific conditions.

Bone formation connects multiple-level processes, which go from tissues to cells and genes. Understanding the interaction of these processes and how to control them is an important contribution to find new treatments for bone mass disorders such as osteoporosis. Bone formation is the result of preosteoblasts differentiation into osteoblasts, which then turn into osteocytes and lining cells constituting the bone tissue. Osteoclasts are responsible for bone resorption through osteoblasts apoptosis.

Our approach is based on hybrid systems [8]. This modeling theory is adapted to integrate different type of stimulatory signals, considering continuous dynamics which interact with discrete changes. The resulting hybrid models are implemented by two frameworks: BioRica [2] and QSS solver [13].

Here, we consider two hybrid models. The first one is based on switching a gene regulatory network (GRN), which models the differentiation from progenitor cells (we call them preosteoblasts) into osteoblasts and adipocytes depending on the condition of stimulatory signals [3]. However, the dynamics of preosteoblasts is, in fact, more complex. An important element to consider is the cellular phenotype and culture conditions. Consequently, in the second model we consider a culture with cells predisposed to osteoblast differentiation. This higher predisposition is given by the presence of epigenetic marks [7, 15]. In addition, we consider that the population size affects the division rate due to space and nutrient limitations [6]. To our knowledge, this study is the first attempt to model bone formation using hybrid systems and including epigenetic inheritance.

1.1 The Problem: Cells Decisions in Bone Formation

In general, every cell has to decide between maintaining its stage, or changing by division, death, or differentiation. Going to one or another commitment depends on cell maturation, signals, and other environmental characteristics such as nutrients, and the cell phenotype (see Fig. 1). The proliferation and apoptosis rates of each cell lineage are regulated by many stimulatory signals. Preosteoblasts going to differentiate have to decide between osteoblasts or adipocytes depending on the condition of such stimulatory signals. The signals we consider are the activation of the *Wnt pathway* (favoring bone cells [14]), the increase in *homocysteine* (favoring preosteoblasts

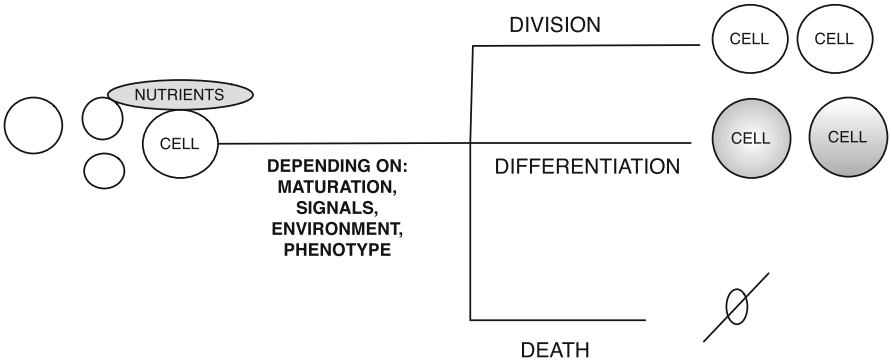


Fig. 1 The cell decisions. At any time, the cell cannot only stay in its stage but also it can change its stage deciding to divide, die (do apoptosis), or differentiate. The stage change event is affected by the conditions of maturation, signals, environment, and phenotype

and osteoblasts apoptosis [10]), and the stimulation of $PPAR\gamma$ (favoring adipocytes [5]).

Effects associated to environment and phenotypic characteristic of preosteoblasts are also considered. In particular, the predisposition of preosteoblasts to differentiate into osteoblasts is included through the presence of specific epigenetic marks [7, 15]. For example, in mammal cells, the presence of *H3K9Ac,* acetylation of the lysine 9 of the histone 3, or the tri-methylation *H3K4me3* in combination with *H3K27Ac* is strongly associated with active gene expression [7, 15] favoring the associated cell lineage.

2 Methods

Our approach is based on the hybrid systems modeling. The basic idea of hybrid systems is connecting continuous and discrete dynamics. Hybrid models consider state variables (continuous) and mode variables (discrete). The state variables evolve over time according to behavior laws, but at any time these laws are modified by *mode changes* [8].

Using hybrid systems in biology is supported in systems biology paradigm [12]. The system behaviors are the result of the interaction of the single models. Some examples of hybrid models in biomedicine are shown in [1] and were the main focus in recent conferences.[1]

[1] HSCB 2009: http://www.eziobartocci.com/hscb, HSB 2012: http://hsb2012.units.it and HSCC 2013: http://2013.hscc-conference.org

The hybrid systems approach allows modeling acclimatization [4]. That is, the dynamics of behavior changes of biological entities as adaptation response to environmental changes. In our case, stimulus signals are the changes in external factors which affect the dynamics of preosteoblasts, osteoblasts, and adipocytes.

Herein, we describe the cell decisions involved in bone tissue formation by two hybrid models. The first model [3] describes the differentiation from presoteoblasts into osteoblasts and adipocytes, and the second model focuses on the proliferation of preosteoblasts and the inheritance of epigenetic marks favoring bone formation.

We implement and simulate the resulting models with *BioRica*[2] [2] and *QSS solver*[3] [13]. Our implementation allows reusing systems biology markup language (SBML) models [9] and including stochastic transitions, with good computation times on stiff components.

2.1 First Model: Switching a Differentiation GRN

We consider that the cellular lineage are limited to preosteoblasts, osteoblasts, and adipocytes, whose dynamics we describe by three state variables: x_P, x_O, x_A. They correspond to the concentration of preosteoblasts, osteoblasts, and adipocytes, respectively. The mode variables connecting stimulus signals with the dynamics of cell lineages are the values of the parameters z_D, z_O, z_A, k_P, k_O. The first three modes describe if the differentiation, the osteoblast lineage, or the adipocyte lineage were stimulated. In particular, for modeling z_O we reuse a SBML model of the Wnt pathway activation. The mode coefficients k_P and k_O are the apoptosis rate of preosteoblasts and osteoblasts (Fig. 2). More details in [3].

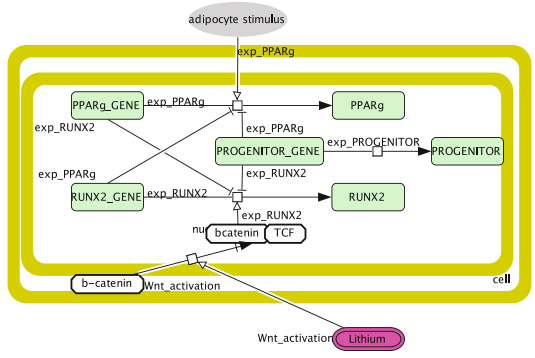
As shown in Fig. 2, the interaction between these three lineages is described by a GRN in which each lineage is associated with a specific biomarker gene (*PROGENITOR, RUNX2,* and *PPARγ*, respectively). The mode changes are triggered by four stimulatory signals: the activation of the differentiation, the activation of the Wnt pathway which stimulates the osteoblast lineage [11], the activation of *PPARγ*, and the increase of *homocysteine* [10] affecting the apoptosis rate of preosteoblasts and osteoblasts.

The pass from a generic cell to the population is given by associating biomarker genes with cell lineage concentrations. That is to say, the activation/inhibition relations in the GRN are translated into a system of ordinary differential equations for x_P, x_O, and x_A. Thus, we build a switched GRN in which the switches of the mode variables are controlled by deterministic and stochastic stimulatory signals with different levels of complexity. In Fig. 2, we show more details about the mode changes and system equations.

[2] http://biorica.gforge.inria.fr
[3] http://sourceforge.net/projects/qssengine/

OSTEO-ADIPO SWITCH MODEL Assar et al., 2012

Scheme:



GRN: RUNX2 associated to osteoblasts,
PPARg associated to adipocytes and PROGENITOR for progenitor lineage
+
Stimulus signals to differentiation and apoptosis

Mathematical model:

x_P,x_O,x_A state variables

z_D,z_O,z_A,
k_P, k_O mode coefficients

GRN modeled by ODEs:

$$\dot{x}_P(t) = \frac{a_P \cdot x_P^n + b_P}{m_P + z_D + c_{PP} \cdot x_P^n} - k_P \cdot x_P,$$

$$\dot{x}_O(t) = \frac{a_O \cdot x_O^n + b_O + z_O}{m_O + c_{OO} \cdot x_O^n + c_{OA} \cdot x_A^n + c_{OP} \cdot x_P^n} - k_O \cdot x_O,$$

$$\dot{x}_A(t) = \frac{a_A \cdot x_A^n + b_A + z_A}{m_A + c_{AA} \cdot x_A^n + c_{AO} \cdot x_O^n + c_{AP} \cdot x_P^n} - k_A \cdot x_A,$$

Switching mode coefficients:

$Z_D = time \geq T_D$, Mode activation to differentiate

$Z_O = 0.8 \bullet (bCatenin\_TCF \geq 1.1 \bullet normal\_bCatenin\_TCF)$, Wnt activation to osteoblasts (Kim et al., 2007)

$Z_A = 0.8 \bullet (time \geq T_A)$, Mode activation to adipocytes

$T_D \sim Exponential(0.01)$,

$T_A \sim Exponential(0.0005)$,

$T_A \geq T_D$

$T_{AP} \sim Exponential(0.0005)$

$k_P = 0.1 \bullet (time \geq T_{AP})$

$k_O = 0.3 \bullet (time \geq T_{AP})$ Mode activation to apoptosis: progenitors and osteoblasts

Fig. 2 First model: Scheme of the gene regulatory network (GRN) and the stimulus signals. Mathematical model: the system of ordinary differential equations and the mode changes that affect it. The dynamics of x_P, x_O, x_A (concentrations of each cell lineage) is switched by modes z_D, z_O, z_A, k_P, k_O that model stimulus signals to each commitment (differentiate, osteoblasts, adipocytes) and rates of death for progenitors and osteoblasts

2.2 Second Model: Considering Preosteoblast Cells as Agents

Now, we focus on modeling the dynamics of preosteoblasts proliferation. The previous approach assumes that every preosteoblast cell has the same division, differentiation, and death rate. However, it is known that not all the cells have the same behavior, in particular to decide differentiation lineages. We include this element considering the phenotype, through epigenetics, as a factor of lineage predisposition.

Epigenetic mechanisms operate at cell lineage key regulatory genes. Epigenetic marks correspond to patrons not in the DNA which imply states of the chromatin structure to favor transcription factor bindings and the consequent gene expression. Consequently, these marks contribute to control the expression of target genes, and with that, generate variably intermediate states in the predisposition to differentiate into a specific cell lineage. In particular for osteoblasts [7, 15], $RUNX2$ is considered as the target gene, whose expression is associated to osteoblast lineage.

As in our previous approach, the system is first modeled at the cellular level. However, in the pass from a cell to the population, we do not assume a generic cell behavior. We consider cells as agents separately modeled, and the population is described by the dynamics of all the individual cells and their interactions. As shown in Fig. 3, the dynamics of every cell is given by its cell cycle over time, but the maturation process is not the same for all the cells. The cell decisions are regulated by such maturation level M and the cell phenotype e (the configuration of the epigenetic marks). The cell phenotype is characterized as the presence of epigenetic marks in the cell which provoke predisposition to the osteoblast lineage ($e = 1$ if the mark is present, $e = 0$ if it is absent).

Within the cell cycle, after the Gap 1, the cell begins the synthesis coming to the reaching Gap 2 and after that undergoes mitosis (the action of dividing). We introduce the option to enlarge the division time by entering the Gap 0 [6], increasing this event probability in function of the number of cells . In addition, reaching a maturation high enough, at the mitosis time, the success of the division is also randomly modeled. Other possibilities we introduce are cell death, apoptosis, and aberrant result (cancer). We consider that, after the cell division, every daughter cell inherits the epigenetic mark separately and randomly. Thus, if the mother cell presents the mark, it can be inherited or not by each daughter cell. The case in which the inheritance is different between daughter cells is called *asymmetric division,* and it is considered as the cause of final differentiation between cells after stimulus signals [7, 15].

3 Results

In Fig. 4, we show the simulated dynamics of the concentrations of preosteoblasts, osteoblasts, and adipocytes, together with the effect of some stimulus signals over them. It is appreciated how this first model achieves to predict the positive effect that Wnt pathway activation has in the formation of osteoblasts. However, consistently

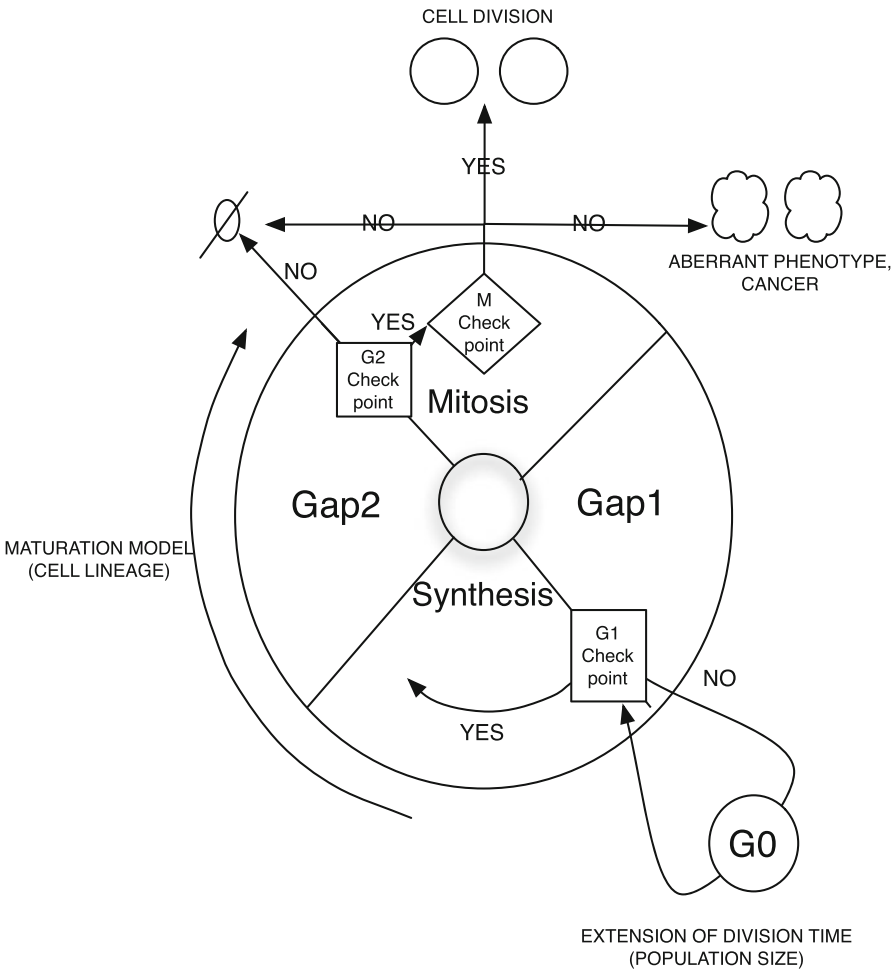


Fig. 3 The cell cycle and cell decisions. The cell maturation is regulated by the cell cycle and depends on the phenotype. The cell begins the synthesis after Gap 1, reaching Gap 2, and after that undergoes mitosis (division). The division time can be enlarged by entering the Gap 0, with bigger probability if the number of cells increases. In addition, reaching a maturation high enough, the success of the division is also randomly modeled. Other possibilities we introduce are cell death, apoptosis, and aberrant result (cancer)

with experimental results, this effect is too weak and nonpermanent. In opposition, the formation of adipocytes is highly sensitive to signals.

As we explained in the previous section, at the second model the dynamics is characterized by the number of cells (N) and the number of those cells with the epigenetic predisposition to the osteoblast lineage (E). In Fig. 5, we show an example of population dynamics depending of individual cell decisions to divide (symmetrically or not) and to die over time. We also show the effect of stimulating the proliferation

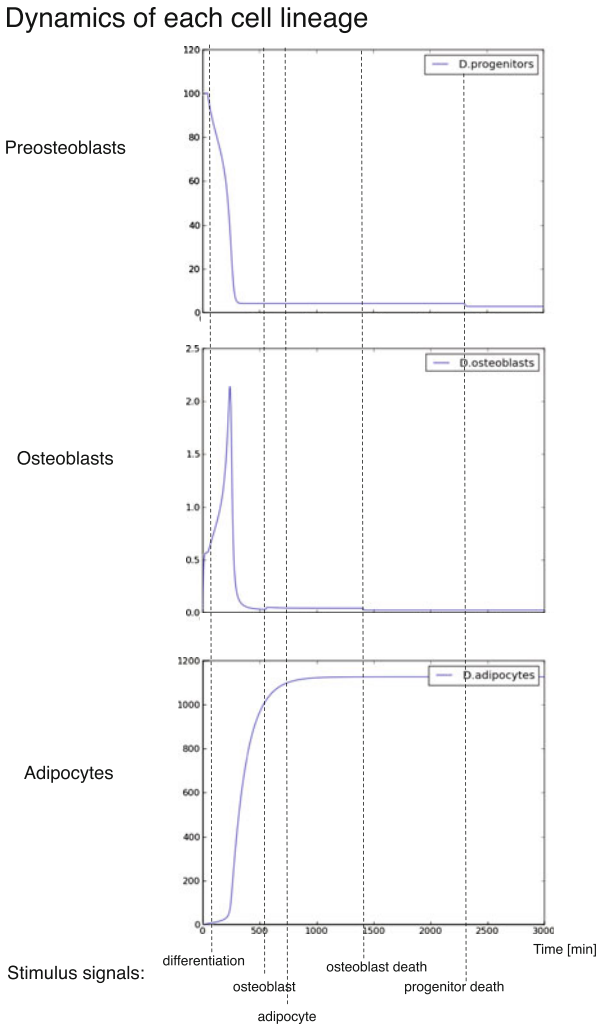


Fig. 4 Results of the first model: the dynamics of the concentrations of preosteoblasts (x_P), osteoblasts (x_O), and adipocytes (x_A). Depending on the stimulus signals the cells receive over time, the preosteoblast, osteoblast, or adipocyte lineages are favored. We show simulation results for different consecutive stimulus signals. The Wnt pathway activation in general allows a weak increment in the concentration of osteoblasts

of epigenetically marked cells by the Wnt pathway (according to [14]) at early and late times. As expected, the effect is better if the culture is stimulated early, at the phase of exponential growth (see Fig. 5).

4 Conclusions and Discussion

The approach by hybrid systems, as shown through this chapter, allows including stimulus signals and the interaction of different types of dynamics in bone formation. With that, the limitations of choosing only one kind of model (continuous or discrete,

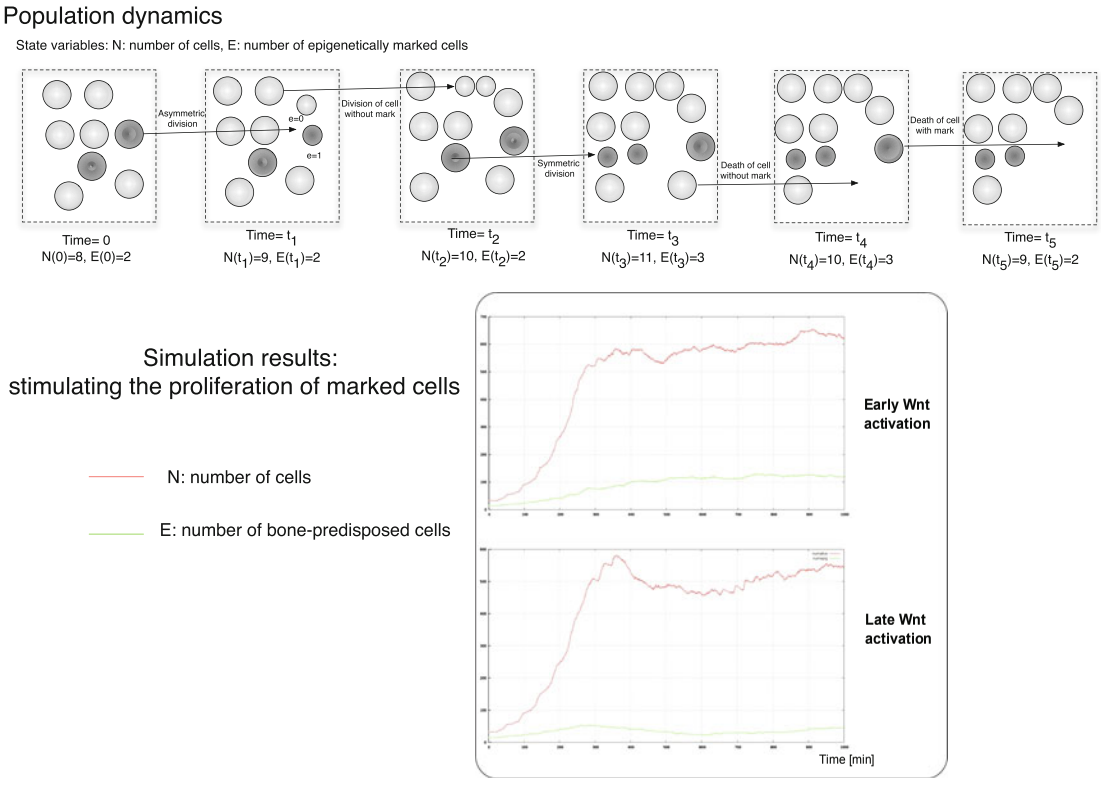


Fig. 5 Results of the second model: the dynamics of the total number of cells and those epigenetically marked. At the *top*: After a successful division of a marked cell, daughter cells can receive both the mark (symmetric division) or only one of them (asymmetric division). As a result, the number of cells N and the number of epigenetically marked cells E evolve over time depending on the decisions of every cell. At the *bottom*: Simulation results for early and late stimulation of the proliferation of epigenetically marked cells

determinist or stochastic or nondeterministic) are avoided. As we have shown, hybrid systems implemented by BioRica and QSS, allow building more realistic models and reusing SBML models.

Through the combination of stimulus signals, and including other factors, we expect to contribute to discover in silico how to increase significantly the formation of osteoblasts. Our final goal is having good models at each level and defining the mechanisms to connect them. With both models we introduced here, we point to find treatment with less side effects and more adapted to patient-specific conditions. Here, we provide insights on the cellular level and describe the two ways to pass to the cells' population level. However, we have a long way to go to obtain complete models for bone formation. It is necessary to integrate many levels, from tissues to genes, and many regulatory processes.

Our first model allows obtaining predictions on the concentration of each lineage (preosteoblast, osteoblast, and adipocyte) over time and the effect of combining stimulus signals. This model succeeds in predicting a weak and nonpermanent increase of osteoblasts by activating the Wnt pathway. Although only few signals were considered, the modeling scheme is flexible enough to include more stimuli. This first model of osteo-adipo differentiation fails in assuming mean cell behaviors in cell decisions without incorporating the phenotypic characteristics of the cells, which is covered by the second model.

The second model for preosteoblasts proliferation points to a more realistic description. With this model, we estimate the effect of low and short stimulus signals (the activation of the Wnt pathway in this case) to increase the number of epigenetically bone-predisposed cells. The model structure allows one to specify it for different phenotypic conditions. The initial predisposition to the formation of osteoblasts is chosen deciding the number of epigenetically marked cells. In addition, the model of cell maturation and parameters, such as the probability of success division and epigenetic inheritance, allows calibrating different culture conditions.

Acknowledgement This work was partially supported by ICBM, Fondecyt 3130762, and Project CIRIC-INRIA Chile.

References

1. Aihara, K., Suzuki, H.: Theory of hybrid dynamical systems and its applications to biological and medical systems. Philos. T. Roy. Soc. A **368**(1930), 4893–4914 (2010)
2. Assar, R., Sherman, D.J.: Implementing biological hybrid systems: allowing composition and avoiding stiffness. Appl. Math. Comput. **223**, 167–179 (2013)
3. Assar, R., Leisewitz, A.V., Garcia, A., Inestrosa, N.C., Montecino, A.M., Sherman, D.J.: Reusing and composing models of cell fate regulation of human bone precursor cells, June 2012. PMID: 22309764
4. Assar, R., Montecino, A.M., Maass, A., Sherman, D.J.: Modeling acclimatization by hybrid systems: condition changes alter biological system behavior models. Biosystems **121**, 43–53 (2014)

5. Chen, J.-R., Lazarenko, O.P., Wu, X., Tong, Y., Blackburn, M.L., Shankar, K., Badger, T.M., Ronis, M.J.J.: Obesity reduces bone density associated with activation of PPAR and suppression of wnt/-catenin in rapidly growing male rats. PLoS ONE, **5**(10), e13704 (2010)
6. Foster, D.A., Yellen, P., Xu, L., Saqcena, M.: Regulation of g1 cell cycle progression distinguishing the restriction point from a nutrient-sensing cell growth checkpoint(s). Genes Cancer **1**(11), 1124–1131 (2010)
7. Gordon, J.A.R., Hassan, M.Q., Koss, M., Montecino, M., Selleri, L., van Wijnen, A.J., Stein, J.L., Stein, G.S., Lian, J.B.: Epigenetic regulation of early osteogenesis and mineralized tissue formation by a HOXA10-PBX1-associated complex. Cells Tissues Organs **194**(2–4), 146–150 (2011). PMID: 21597276
8. Henzinger, T.A.: The theory of hybrid automata. In: Eleventh Annual IEEE Symposium on Logic in Computer Science, 1996. LICS '96. Proceedings, pp. 278–292. IEEE, July 1996
9. Hucka, M., Hucka, M., Bergmann, F., Hoops, S. Keating, S., Sahle, S., Wilkinson, D.: The systems biology markup language (SBML): language specification for level 3 version 1 core (Release 1 candidate). Nature Precedings, January 2010
10. Kim, D.J., Koh, J.-M., Lee, O., Kim, N.J., Lee, Y.-S., Kim, Y.S., Park, J.-Y., Lee, K.-U., Kim, G.S.: Homocysteine enhances apoptosis in human bone marrow stromal cells. Bone **39**(3), 582–590 (2006). PMID: 16644300
11. Kim, D., Rath, O., Kolch, W, Cho, K-H: A hidden oncogenic positive feedback loop caused by crosstalk between wnt and ERK pathways. Oncogene **26**(31), 4571–4579 (2007). PMID: 17237813
12. Kitano, H.: Computational systems biology. Nature **420**(6912), 206–210 (2002). PMID: 12432404
13. Kofman, E., Junco, S.: Quantized state systems. A DEVS approach for continuous systems simulation. Trans. SCS **18**(3), 123–132 (2001)
14. Krishnan, V., Bryant, H.U., Macdougald, O.A.: Regulation of bone mass by Wnt signaling. J. Clin. Invest. **116**(5), 1202–1209 (2006). PMID: 16670761
15. Zaidi, S.K., Young, D.W., Montecino, M., Lian, J.B., van Wijnen, A.J., Stein, J.L., Stein G.S.: Mitotic bookmarking of genes: a novel dimension to epigenetic control. Nat. Rev. Genet. **11**(8), 583–589 (2010). PMID: 20628351 PMCID: 3033599

Biodiversity and its Role on Diseases Transmission Cycles

Juan Manuel Cordovez and Camilo Sanabria

Abstract Recently, most notably after the appearance of the article by Keesing et al., *Impacts of biodiversity on the emergence and transmission of infectious diseases* in *Nature* in 2010, there has been a growing interest in understanding the relationship between biodiversity and epidemiology. On the one hand, regions with high biodiversity may be sources of new pathogens; on the other hand, biodiverse ecological communities may buffer the transmission by "diluting" the disease. Using mathematical epidemiology, we provide a framework to measure and interpret how a change in the abundance and richness of species of an ecosystem could affect the prevalence and the incidence of a disease. Moreover, we are able to quantify the effect of such a change on the incidence of an infectious disease in a specific species of the ecosystem, making this framework highly relevant for assessing the impact on humans.

Keywords Biodiversity · Virulence · Epidemiological network · Cycles

1 Introduction

With the current trends in land use, great changes in ecosystem biodiversity are expected [13, 22]. More precisely, as the habitat is degraded, species composition and abundance will change to accommodate for an environment that offers less shelters, hiding places, and varied diets. Furthermore, because species form an intricate network that supports life, including parasitic forms, alteration to this network might

J. M. Cordovez (✉)
Departamento de Ingeniería Biomédica, Universidad de los Andes,
Carrera 1 Este # 19A - 40, Bogotá D.C., 111711 Colombia
e-mail: jucordov@uniandes.edu.co.

C. Sanabria
Departamento de Matemáticas, Universidad de los Andes,
Carrera 1 # 18A - 12, Bogotá D.C., 111711 Colombia
e-mail: c.sanabria135@uniandes.edu.co.

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_17

have deeper impacts than the normal loss in biodiversity. Indeed, species exhibit different competence for parasite development and transmission; a shift in species composition can potentially alter the presence of the parasite in the network, measured in terms of number of species infected and parasite abundance [17, 18, 20]. Some studies have suggested that mid-size predators, usually more competent parasite hosts, replace top-level predators because the latter are highly sensitive to habitat disturbance [13].

The dilution effect hypothesis, suggested in the literature about 15 years ago, proposes that as species diversity increases the individual risk of any member of the network to become infected diminishes [13, 21–23]. If we take into consideration the multiple hosts vector-borne diseases that include humans, the dilution effect implies that as habitats become degraded the probability to have human cases of a wide sort of parasitic diseases increases. However, as sound this idea might be, the hypothesis has been proven hard to be validated experimentally [14, 15]. Field studies have the challenges of measuring biodiversity while controlling for species competence in degraded habitats that often do not provide enough resolution. Thus, it is not surprising that some reports have supported the hypothesis of an inverse relation between biodiversity and parasite infection risk [8] while others have suggested that the relation is not apparent [18].

If we consider the environment as a network where nodes correspond to species and the edges represent relationships between species (i.e., who eats who), it is clear that changes in the environment can remove some nodes (while may add others) and change the overall connection between the nodes. One can imagine a situation in which removing several nodes of the incompetent species produces an increase in abundance of a competent species and thus an increase in network infection (taken as the proportion of infected individuals in the network). But similarly a change in the environment that produces the disappearance of a competent species can produce the opposite effect. Thus, in theory, the dilution effect could take place only under certain circumstances [13].

In this study, we propose a theoretical approach to study dilution from an epidemiological perspective. We are interested in determining the key aspects of the network architecture that can lead to the dilution effect. To this end, we propose a general framework mathematical model of susceptible and infective subjects from different species that are interconnected. In the model, the species competence is captured by the force of infection that states the probability of a susceptible becoming infected after interaction with an infected individual. By computing the next-generation matrix (NGM) of the system and the associated basic reproductive number on a network scheme, we compute the virulence of each cycle (i.e., geometric mean of the NGM entries of all the nodes involved in every possible circuit) and based on this scale we identify critical cycles of disease transmission.

Mathematical epidemiological models are a specific type of dynamical population models. Those models aim at describing or predicting the average dynamics of the transmission of a disease among members of a particular population. The literature detailing with the methods and tools of mathematical epidemiology is plentiful, for example one could consider [1, 4, 5, 19]. An important quantity associated to

these models is the abovementioned basic reproductive number: a threshold value which indicates whether a disease is going to invade the population. Although, given an specific model, obtaining this number is straighforward (cf. [10]), since it is generally computed as the spectral radius of a matrix, it is rather involved to measure its sensibility to the different parameters in the model.

In this chapter, we propose an alternative quantity, the critical virulence, which bounds the basic reproductive number (sometimes they even coincide), and, being easier to compute, it is more malleable. Especially when many different species carrying a common pathogen are involved in the transmission of the disease. Furthermore, as with the basic reproductive number, which measures the secondary infections arising by introducing an infected individual into a population of susceptible, the critical virulence also has a biological interpretation.

We think this approach is highly relevant to study the disease transmission risk of some highly prevalent vector-borne diseases that include multiple hosts, including humans, such as chagas, malaria, or leishmaniasis [6, 9, 13].

2 The Model

To model the transmission of a disease in an environment with the organisms of different type, we use the following assumptions and parameters:

- There are N different types of organisms in our environment; we denote by x_i the abundance of i-organisms (i.e., organisms of type i) susceptible of acquiring the disease, and by y_i the abundance of infected i-organisms, $i = 1, \ldots, N$.
- The probability of i-organisms getting infected by j-organisms, in a unit of time, will be denoted by β_{ij}.
- The natural mortality rate of the i-organisms will be denoted by μ_i.
- The mortality rate of the i-organisms infected by the disease will be denoted by $\mu_i + d_i$.
- The recruitment of the i-organisms will be denoted by Λ_i. We assume no newly recruited organism comes with the disease, i.e., there is no vertical transmission.

Our model is summarized by the following system of equations:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \Lambda_i - x_i \left(\mu_i + \sum_{j=1}^{N} \beta_{ij} y_j \right) \tag{1}$$

$$\frac{\mathrm{d}y_i}{\mathrm{d}t} = x_i \left(\sum_{j=1}^{N} \beta_{ij} y_j \right) - y_i(\mu_i + d_i) \tag{2}$$

where $i = 1, \ldots, N$.

For simplicity, we eliminate any vital dynamics assuming that the abundance of each organism $n_i = x_i + y_i$ is constant (i.e., $\Lambda_i = \mu_i n_i + d_i y_i$). This way, we may

disregard the susceptible state variables by putting $x_i = n_i - y_i$, so that Eqs. (1) and (2) collapse into:

$$\frac{\mathrm{d}y_i}{\mathrm{d}t} = (n_i - y_i)\left(\sum_{j=1}^{N} \beta_{ij} y_j\right) - y_i(\mu_i + d_i) \tag{3}$$

for $i = 1, \ldots, N$.

Note that here we do not explicitly distinguish vectors from hosts. In our model, vectors are the i-organisms with $\beta_{ii} = 0$ (i.e., organisms not capable of transmitting the disease to other organisms of the same type).

2.1 The NGM

The capacity a disease has for invading the environment will be measured using the spectral radius of the *NGM* G [7, 10]. To get G, we use the influx in (3) of new infected i-organisms:

$$F_i(y_1, \ldots, y_N) = (n_i - y_i)\left(\sum_{j=1}^{N} \beta_{ij} y_j\right)$$

and the outflux:

$$V_i(y_1, \ldots, y_N) = y_i(\mu_i + d_i).$$

The NGM is then given by

$$G = FV^{-1}$$

where F and V are the Jacobians of $\mathcal{F} = (F_1, \ldots, F_N)$ and $\mathcal{V} = (V_1, \ldots, V_N)$, respectively, evaluated at the disease-free equilibrium:

$$(y_1, \ldots, y_N) = (0, \ldots, 0).$$

Whence, if g_{ij} denotes the ij-entry in G,

$$g_{ij} = \frac{n_i \beta_{ij}}{\mu_j + d_j}. \tag{4}$$

Indeed, the ijth entry of F is

$$\frac{\partial F_i}{\partial y_j}(0, \ldots, 0) = n_i \beta_{ij}$$

and V is the diagonal matrix with diagonal entries $\mu_1 + d_1, \ldots, \mu_N + d_N$.

The NGM is telling us that if we consider our environment in a disease-free state and introduce y_1 1-organisms, y_2 2-organisms, …, y_N N-organism, all of these infected, then we can expect $\sum_j g_{ij}\, y_j$ newly infected i-organisms.

2.2 Basic Reproductive Number

The invading capacity of a disease is given by the basic reproductive number $\mathcal{R}_0$ defined as the spectral radius of G:

$$\mathcal{R}_0 = \rho(G).$$

Recall that $\rho(G)$ is defined as the greatest module of the eigenvalues of G. This number is a threshold value telling us whether the disease could eventually disappear by itself [10, Theorem 2]: if $\mathcal{R}_0 < 1$ it could, but if $\mathcal{R}_0 >, 1$ it will not. Intuitively, this number can be thought as the expected number of secondary cases arising from introducing a single infected organism into our disease-free environment.

The NGM is nonnegative ($g_{ij} \geq 0$ for every i, j), so we use the Perron–Frobenius theory to study $\mathcal{R}_0$; or more precisely, its *max version* [2]. To G, we associate a weighted-directed graph $\Gamma(G)$ with vertices $1, 2, \ldots, N$ and with an edge from i to j of weight g_{ij}, if and only if, $g_{ij} > 0$ [3, Sect. 1.1]. A *path* in $\Gamma(G)$ of *length* k is a sequence of vertices:

$$i_1 \to i_2 \to \cdots \to i_k \to i_{k+1}$$

such that there is an edge from i_j to i_{j+1} for $j = 1, \ldots, k$; and, its *path geometric mean* is the geometric mean of the weight of the edges:

$$\left(\prod_{j=1}^{k} g_{i_j i_{j+1}} \right)^{\frac{1}{k}}.$$

A *circuit* of length k is a path in $\Gamma(G)$ of length k with $i_{k+1} = i_1$ and not crossing any vertex more than once; and, the *circuit geometric mean* is the path geometric mean of a circuit. The greatest circuit geometric mean in $\Gamma(G)$ will be denoted by $\mu(G)$; it is commonly known as *limit eigenvalue* of G [12, 16]. A circuit is *critical* if its circuit geometric mean is equal to $\mu(G)$. Finally, the subgraph of $\Gamma(G)$ spanned by the vertices of a critical circuit is called *critical subgraph*.

Using the limit eigenvalue $\mu(G)$, we can get bounds on $\mathcal{R}_0 = \rho(G)$ [12, Inequality (6) and Theorem 2]:

$$\mu(G) \leq \rho(G) \leq \rho(S(G))\mu(G) \tag{5}$$

where $S(G)$ is the matrix whose ij-entry is signum(g_{ij}):

$$\text{signum}(g_{ij}) = \begin{cases} 0 & \text{if } g_{ij} = 0 \\ 1 & \text{if } g_{ij} > 0. \end{cases}$$

Note that $S(G)$ is a matrix with zeroes and ones, so it follows from the Perron–Frobenius theorem that its spectral radius $\rho(S(G))$ in (5) can vary from 0 to N, and may not be an integer.

3 Biological Interpretation

In the NGM G, the ij-term g_{ij} tells us that if into a disease-free environment we introduce a j-organism, it will result in g_{ij} infected i-organism. We think of our environment as the weighted-directed graph associated to G, $\Gamma(G)$, with vertices $1, 2, \ldots, N$ (i.e., each node represents an organism type) and with an edge from i to j of weight g_{ij} whenever $g_{ij} > 0$.

We include two examples to illustrate the results from the previous section.

Example 1 We take an environment with only two organisms: a vector and a host. We consider a slight modification of the vector–host model in [11]:

$$\begin{aligned}
\frac{\mathrm{d}S_v}{\mathrm{d}t} &= \mu N_v - \beta I_h S_v - \mu S_v \\
\frac{\mathrm{d}I_v}{\mathrm{d}t} &= \beta I_h S_v - \mu I_v \\
\frac{\mathrm{d}S_h}{\mathrm{d}t} &= mN_h - bI_v S_h - mS_h \\
\frac{\mathrm{d}I_h}{\mathrm{d}t} &= bI_v S_h - mI_h
\end{aligned}$$

where N_v, S_v, I_v (respectively N_h, S_h, I_h) is the abundance of vectors, susceptible vectors and infected vectors (respectively of hosts), μ (respectively m) their death rate, and, β and b are the probabilities that a vector will get the disease from a host and *vice versa*. Having no vital dynamics, assuming $N_v = N_h = 1$, we can summarize the model by

$$\begin{aligned}
\frac{\mathrm{d}I_v}{\mathrm{d}t} &= \beta I_h(1 - I_v) - \mu I_v \\
\frac{\mathrm{d}I_h}{\mathrm{d}t} &= bI_v(1 - I_h) - mi_h.
\end{aligned}$$

In such case the NGM is

$$G = \begin{pmatrix} 0 & \beta/m \\ b/\mu & 0 \end{pmatrix}.$$

The values $g_{12} = \beta/m$, which can be thought as the infectability of the vector and we denote by $\mathcal{R}_{0v}$, measures the number of infected hosts arising from the introduction of an infected vector in an otherwise disease-free system. It is obtained by combining the life expectancy of a vector m^{-1} and its probability of infecting a host β. Similarly, the value $g_{21} = \mathcal{R}_{0h} = b/\mu$ measures the infectability of the host. For G, its associated weighted-directed graph is displayed in Fig. 1. The limit eigenvalue is obtained by taking the circuit $1 \to 2 \to 1$:

$$\mu(G) = \sqrt{g_{12}g_{21}} = \sqrt{\frac{b\beta}{m\mu}} = \rho(G).$$

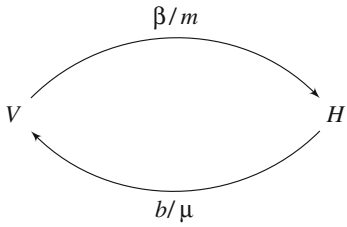


Fig. 1 Directed graph associated to the vector–host model

Note that in this case $\rho(S(G)) = 1$. So, regardless of the value of infectability of the vector $\mathcal{R}_{0v} = \beta/m$, the disease will not invade the environment whenever the host can contain the disease, i.e., whenever $\mathcal{R}_{0h} = b/\mu < m/\beta = \mathcal{R}_{0v}^{-1}$.

Based on this vector–host model our associated directed graph has two vertices, labeled V and H; and two directed edges, one from V to H, with weight β/m, and another from H to V, weight b/μ. See Fig. 1.

The limit eigenvalue $\mu(G)$ is the greatest circuit geometric mean in $\Gamma(G)$. Given a circuit, we will call its circuit geometric mean the *virulence*, and $\mu(G)$ the *critical virulence*. From (5), the critical virulence bounds the reproductive number $\mathcal{R}_0$ of the disease.

Example 2 As a generalization of the previous example, we may consider an environment with two host, H_1 and H_2, capable of transmitting the disease among themselves, and a vector V which can infect both organisms H_1 and H_2 but cannot infect other vectors and acquires the disease from infected hosts. This model is summarized in graph of Fig. 2.

The parameter β_i corresponds to the probability of the vector infecting host H_i, $i = 1, 2$, μ_i to the mortality rate of H_i, b_i is the probability that a vector will get infected by a host of type i, and m is the mortality of V. The circuits are:

$$H_1 \to H_1$$
$$H_2 \to H_2$$
$$H_1 \to V \to H_1$$
$$H_2 \to V \to H_2$$
$$H_1 \to V \to H_2 \to V \to H_1.$$

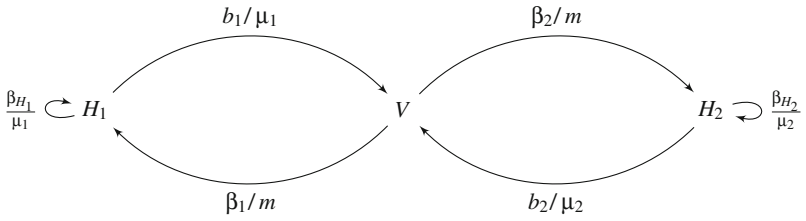


Fig. 2 Directed graph associated to the host–vector–host model

The one with the greatest virulence will determine the critical virulence $\mu(G)$ of the system. For this model:

$$S(G) = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix},$$

so that $\rho(S(G)) = 2$, and we have $\mu(G) \leq \mathcal{R}_0 \leq 2\mu(G)$.

The advantage of working with the critical virulence, rather than with the reproductive number, as a way of measuring the invading capacity of a disease in a biodiverse environment, lies on the fact that measuring the sensibility and the elasticity to the parameters of the model of the latter is considerably harder than that of the former. Furthermore, the critical virulence is telling us that the reproductive capacity of a disease is concentrated on the critical circuit.

4 Discussion

As the risk of human infection can be stated in terms of the abundance of competent species, the number of encounters between humans and infected subjects, and vector-feeding preferences; habitat destruction can either increase or decrease the risk of infection depending on what species remain and their relative contribution to parasite flow. Here, we presented a novel approach to identify critical cycles in the network as those that are dominant when describing the ultimate risk of human infection. In this chapter, we propose that the cycle with the maximum virulence for one particular node is called a critical cycle and is such because it explains the maximum amount of secondary infections of that type.

As very often the introduction of an infected organisms does not directly produces the infection of a particular type, but involves infection of multiple types in between, it is important to consider the chain of species involved in a specific parasite transmission network. The NGM provides a framework to establish the number of secondary cases after the introduction of any infected type in every other type in the network by averaging the contribution via different pathways. In this sense, removal of one node can be tested in terms of infection level in another but one would never know how this change came about. With the proposed critical cycle, we can tease apart the effect of one node into another by its paths and compare the virulence of each one. In this way, control programs could target more efficiently some privileged paths because of their ability to host the parasite.

This might be of importance in chagas disease. In chagas, *Didelphis marsupialis* has been proposed to be the main host of *Trypanosome cruzi* in great parts of South and Central America, it has been suggested that this mammal becomes infected via insect (vector) biting while searching for food or shelter in palm trees, the natural habitat of many vectors [13]. The mammal, because of its peridomicilary habits,

can support big insect populations close to human dwellings. Therefore, this host could be playing the role of moving the parasite from sylvatic populations to insect population close to humans. Other hosts include birds and reptiles that do not host the parasite but provide meals for insects. Habitat deforestation might increase the abundance of *D. marsupialis* while reducing other host types, what would be the effect of this land use change on human chagas disease risk? We believe that for diseases with complex parasite transmission networks such as Chagas the notion of critical cycle can provide answer to some of these questions; the challenge resides in measuring disease transmission parameters accurately.

Acknowledgement The work of J. M. Cordovez was partially supported by Vicerrectoría de Investigaciones de la Universidad de los Andes. The work of C. Sanabria was partially supported by Vicerrectoría de Investigaciones de la Universidad de los Andes grant PEP P13.160422.030 FAPA-Camilo Sanabria.

References

1. Anderson, R.M., May, R.M.: Infectious diseases of humans, dynamics and control. Oxford University Press (1991)
2. Bapat, R.B.: A max version of the Perron-Frobenius theorem. Linear Algebra Appl. **275–276**, 3–18 (15 May 1998). ISSN 0024-3795, 10.1016/S0024-3795(97)10057–X
3. Bapat, R.B., Raghavan, T.E.S.: Nonnegative matrices and applications. Encyclopedia of mathematics, vol. 64. Cambridge University Press, Cambridge (1997)
4. Bailey, N.: The mathematical theory of infectious diseases. Charles Griffin (1975)
5. Brauer, F., Castillo-Chavez, C.: Mathematical models in population biology and epidemiology. Springer (2000)
6. Chaves, L.F., Hernandez, M., Dobson, A.P., Pascual, M.: Sources and sinks: revisiting the criteria for identifying reservoirs for american cutaneous leishmaniasis. Trends Parasitol. **23**, 311–316 (2007)
7. Diekmann, O., Heesterbeek, J.A.P., Metz, J.A.J.: On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. J. Math. Biol. **28**(4), 365–382 (1990)
8. Dizney, L.J., Ruedas, L.A.: Increased host species diversity and decreased prevalence of sin nombre virus. Emerg. Infect. Dis. **15**, 1012–1018 (2009)
9. Dobson, A.: Population dynamics of pathogens with multiple host species. Am. Nat. **164**, S64–S78 (2004)
10. van den Driessche, P., Watmough, J.: Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. Math. Biosci. **180**(1–2), 29–48 (November–December 2002). ISSN 0025-5564, 10.1016/S0025-5564(02)00108–6
11. Feng, Z., Velasco-Hernández, J.X., Competitive exclusion in a vector-host model for the Dengue fever. J. Math. Biol. **35**, 523–544 (1997)
12. Friedland, S.: Limit eigenvalues of nonnegative matrices. Linear Algebra Appl. **74**, 173–178 (February 1986). ISSN 0024-3795, 10.1016/0024-3795(86)90120–5
13. Gottdenker, N.L., Chaves, L.F., Calzada, J.E., Saldaña, A., Carroll, C.R.: Host life history strategy, species diversity, and habitat influence trypanosoma cruzi vector infection in changing landscapes. PLoS Negl. Trop. Dis. **6**(11), e1884 (2012). 10.1371/journal.pntd.0001884
14. Holt, R., Pickering, J.: Infectious disease and species coexistence—a model of Lotka-Volterra form. Am. Nat. **126**, 196–211 (1985)

15. Holt, R., Dobson, A., Begon, M., Bowers, R., Schaube, E.: Parasite establishment in host communities. Ecol. Lett. **6**, 837–842 (2003)
16. Karlin, S., Ost, F., Some monotonicity properties of Schur powers of matrices and related inequalities. Linear Algebra Appl. **68**, 47–65 (July 1985). ISSN 0024-3795, 10.1016/0024-3795(85)90207–1
17. Keesing, F., Holt, R.D., Ostfeld, R.S.: Effects of species diversity on disease risk. Ecol. Lett. **9**, 485–498 (2006)
18. Keesing, F., Belden, L.K., Daszak, P., Dobson, A., Harvell, C.D.: Impacts of biodiversity on the emergence and transmission of infectious diseases. Nature **468**, 647–652 (December 2010). 10.1038/nature09575
19. Murray, J.D.: Mathematical Biology I and II. Springer (2004)
20. Rudolf, V.H.W., Antonovics, J.: Species coexistence and pathogens with frequency-dependent transmission. Am. Nat. **166**, 112–118 (2005)
21. Schmidt K, K., Ostfeld, R.: Biodiversity and the dilution effect in disease ecology. Ecology **82**, 609–619 (2001)
22. Suzan, G., Marce, E., Giermakowski, J.T., Armien, B., Pascale, J., et al. The effect of habitat fragmentation and species diversity loss on Hantavirus prevalence in Panama. Ann. NY Acad. Sci. **1149**, 80–83 (2008)
23. Telfer, S., Bown, K.J., Sekules, R., Begon, M., Hayden, T., et al. Disruption of a host-parasite system following the introduction of an exotic host species. Parasitology **130**, 661–668 (2005)

Simulation Model for AIDS Dynamics and Optimal Control Through Antiviral Treatment

Carlos Andrés Trujillo-Salazar and Hernán Darío Toro-Zapata

Abstract A mathematical model based on ordinary differential equations to study AIDS dynamics at a population level and giving importance to diagnosis of infected is proposed. Five populations are considered: susceptibles, healthy diagnosed, healthy undiagnosed HIV positives, sick diagnosed, and undiagnosed HIV positives. The number R_0 is analytically calculated and used in numerical results interpretation to determine the long-term population behavior and which parameters are the most influential on the dynamics. Subsequently, antiviral treatment is incorporated into the model as a control strategy and the Pontryagin maximum principle is used to find out an optimal control function. Finally, different simulations are performed and interpreted.

Keywords AIDS · Antiviral treatment · Diagnosis · Dynamic system · Optimal control

1 Introduction

AIDS study does not require additional motivation, the importance of research around this problem is sufficiently illustrated with numbers. In the *Report on the global AIDS epidemic 2013* from the United Nations (UN) [7], it is said that in 2012, 35.3 million people were living with HIV worldwide, and 1.6 million people have died from AIDS. Not surprisingly fighting this disease is a generalized slogan of most governments and ongoing research on several fronts; most of them include *prevention* and *treatment*.

In regard to prevention, the UN reveals that the annual number of new HIV infections in adults and adolescents decreased by 50 % or more in 26 countries between 2001 and 2012, a goal that was set for 2015. However, other countries are

C. A. Trujillo-Salazar (✉) · H. D. Toro-Zapata
Universidad del Quindío, Carrera 15 Calle 12 Norte Armenia, Quindío, Colombia
e-mail: catrujillo@uniquindio.edu.co

H. D. Toro-Zapata
e-mail: hdtoro@uniquindio.edu.co

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_18

failing to halve sexual HIV transmission, highlighting *the importance of intensifying prevention efforts* [7].

Antiviral therapy, is a reactive strategy against disease, although it improves HIV positives quality of life, does not prevent disease transmission. Indeed, rapid expansion of access to treatment has helped reduce the number of AIDS-related deaths but also contributes to the increased HIV prevalence. This increase has been documented in sex workers (SW) in populations of men who have sex with men (MSM), and confined populations, as is the case of inmates in prisons. Given the current disease status at a global scale, it is necessary to develop theoretical studies to provide scientific basis for decision making in the HIV and AIDS treatment [3, 8–10, 14, 16].

The UN also reports that the first-line antiviral therapy cost in some low- and middle-income countries has been reduced approximately to $140 per year per person, a significative number, taking into account that in the mid-1990s, the cost was about $ 10,000 per year per person. Accompanied by other policies, reduction of treatment cost has allowed 9.7 million people in low- and middle-income countries to have access to antiviral therapy in late 2012 [13]. This represents 61 % of those who were eligible under the guidelines of HIV treatment established by the World Health Organization (WHO) in 2010 [7].

Recent scientific evidence, based on clinical trials, has shown that early access to treatment can save lives. In 2013, the WHO revised its guidelines in the light of this new evidence and began to recommend treatment to be started long before and immediately in some cases. This means that 28.6 million people were eligible for treatment in 2013. Science has also shown that if HIV-positive pregnant women have access to antiviral drugs, risk of transmitting the virus to her child can be reduced to below 5 %. In 2012, about 62 % of these women had access to antiviral drugs and in many countries coverage levels exceeded 80 % [13].

According to the above, it is evident that the treatment is a very important aspect to be considered in AIDS research. This chapter presents an ordinary differential equations system based on AIDS dynamics, studied from a numerical perspective and subsequently control strategies based on treatment.

2 Basic Model

Three terms are defined from the medical point of view, as a simplification of the most precise definitions given in [15]. The defined terms are very important throughout the chapter:

- *HIV positive:* patient having antibodies to HIV.
- *Healthy HIV positive:* HIV positive with no symptoms of disease associated with infection. It is classified as diagnosed and undiagnosed.
- *Sick HIV positive:* HIV positive with symptoms of disease associated with infection. It is classified as diagnosed and undiagnosed.

Human population is divided into five well-differentiated categories: $x = x(t)$ denotes the average number of healthy people susceptible to get infected, $y_1 = y_1(t)$ denotes the average number of healthy undiagnosed HIV positives, $y_2 = y_2(t)$ denotes the average number of healthy diagnosed HIV positives, $z_1 = z_1(t)$ denotes the average number of sick undiagnosed HIV positives, and $z_2 = z_2(t)$ denotes the average number of sick diagnosed HIV positives.

Growth rate is assumed constant. It is considered only sexual transmission between healthy and infected HIV positives in a mixed population, i.e., no difference in age, gender, or sexual orientation is made. It is assumed that susceptible individuals acquire the virus through sexual contact with healthy and sick undiagnosed HIV positives, with transmission rate β_1, i.e., terms $\beta_1 x y_1$ and $\beta_1 x z_1$ are the average number of susceptibles that get infected. Infection rate is the same because there is no change in sexual behavior due to ignorance of their HIV status.

It is assumed that despite the diagnosis, it is still possible that susceptible people acquire HIV from healthy and sick diagnosed HIV positives, with transmission rates β_2 and β_3, respectively, i.e., terms $\beta_2 x y_2$ and $\beta_3 x z_2$ represent the average of susceptibles that get infected. This is a somewhat delicate consideration, because diagnosed HIV positives irresponsible sexual behavior is assumed. However, when simulating, an additional consideration made was β_1 greater than β_2 and β_3.

Healthy HIV positives evolve to sick HIV positives, preserving their diagnosed condition in which they are located. In this case, $\gamma_1 y_1$ and $\gamma_2 y_2$ are the average number of healthy HIV positives that evolve to sick HIV positives, diagnosed and undiagnosed, respectively. Sick HIV positives, diagnosed and undiagnosed, die by infection-related causes at rates ω_1 and ω_2, so the terms $\omega_1 z_1$ is the average of sick diagnosed HIV positives who dies and $\omega_2 z_2$ is the average of sick undiagnosed HIV positives who dies. Healthy HIV positives and sick HIV positives are diagnosed at rates δ_1 and δ_2; in this case, $\delta_1 y_1$ is the average of healthy HIV positives who gets diagnosed and $\delta_2 z_1$ is the average of sick HIV positives who gets diagnosed.

Taking into account definitions, variables, and assumptions above, a mathematical model based on ordinary differential equations arises, which describe, at least in theory, the interaction between the populations considered. The dot on each of the variables represents derivative, i.e., variation respect to time. The model is as follows:

$$\begin{cases} \dot{x} = \Lambda - \beta_1 x y_1 - \beta_2 x y_2 - \beta_1 x z_1 - \beta_3 x z_2 - \mu x \\ \dot{y}_1 = \beta_1 x y_1 + \beta_2 x y_2 + \beta_1 x z_1 + \beta_3 x z_2 - \delta_1 y_1 - \theta y_1 \\ \dot{y}_2 = \delta_1 y_1 - \phi y_2 \\ \dot{z}_1 = \gamma_1 y_1 - \delta_2 z_1 - \rho z_1 \\ \dot{z}_2 = \gamma_2 y_2 + \delta_2 z_1 - \eta z_2, \end{cases} \tag{1}$$

where $\theta = \gamma_1 + \mu$, $\phi = \gamma_2 + \mu$, $\rho = \omega_1 + \mu$, $\eta = \omega_2 + \mu$ and initial conditions:

$$x(0) = x_0, \quad y_1(0) = y_{10}, \quad y_2(0) = y_{20}, \quad z_1(0) = z_{10}, \quad z_2(0) = z_{20}. \tag{2}$$

Proposition 1 *System (1) is defined in the positively invariant region,*

$$\Omega = \left\{(x, y_1, y_2, z_1, z_2) \in \mathbb{R}^5 : 0 < x + y_1 + y_2 + z_1 + z_2 \leq \frac{\Lambda}{\mu}\right\}.$$

The region defined in Proposition 1 is important because it guarantees nonnegativity of the solutions and that they will remain in Ω as $t \to \infty$.

2.1 Disease-Free Equilibrium and R_0

System (1) has a *trivial equilibrium given by* $E_0 = \left(\frac{\Lambda}{\mu}, 0, 0, 0, 0\right)$, which represents the equilibrium state in the absence of infection, i.e., without disease, one would expect the susceptible population to reach the equilibrium value of Λ/μ, corresponding to the total population. It is known that the stability of the equilibrium points determines the future behavior of the infection, the study is done from the sign of the eigenvalues of the Jacobian matrix of the system, evaluated at such points. In the particular case of E_0, the characteristic equation is given by

$$-\frac{1}{\mu}(\lambda + \mu)(p_0\lambda^4 + p_1\lambda^3 + p_2\lambda^2 + p_3\lambda + p_4) = 0.$$

Obviously, one of the eigenvalues is $\lambda = -\mu$, but the other four eigenvalues are very uncomfortable to handle explicitly, since the coefficients $p_i, i = 0, \ldots, 4$ correspond to very large algebraic expressions. Therefore, stability analysis is omitted.

One way to study the disease behavior is to analyze the *basic reproduction number,* R_0, which represents the number of secondary cases that are produced by an infected individual in an entirely susceptible population. Using the method of *the next-generation matrix,* this number is obtained

Proposition 2 *The basic reproduction number* R_0 *for the system (1) is*

$$R_0 = \frac{\beta_1 x^*}{\delta_1 + \theta} + \frac{\beta_2 x^* \delta_1}{\phi(\delta_1 + \theta)} + \frac{\beta_1 x^* \gamma_1}{(\delta_2 + \rho)(\delta_1 + \theta)} + \frac{\beta_3 x^*(\gamma_2\delta_1\delta_2 + \gamma_2\delta_1\rho + \delta_2\gamma_1\phi)}{\phi\eta(\delta_1 + \theta)(\delta_2 + \rho)},$$

where $x^* = \frac{\Lambda}{\mu}$.

2.2 Basic Model Numerical Results

Numerical simulations were performed to visualize the long-term behavior of model (1); in order to obtain different settings, 21 simulations for each parameter were made. It was considered a *baseline* value (*ad hoc*) for each parameter, and randomly 20 other values were assigned using a uniform probability distribution. These values

Table 1 Baseline values considered in numerical simulations

Parameters	Description	Value	+100%
Λ	Constant growth rate	10	
β_1	Transmission rate by contact with an undiagnosed HIV+	0.0002	0.0004
β_2	Transmission rate by contact with a healthy diagnosed HIV+	0.00002	0.00004
β_3	Transmission rate by contact with an sick diagnosed HIV+	0.00001	0.00002
μ	Natural death rate	0.01	0.02
γ_1	Evolution from healthy undiagnosed HIV+ to sick undiagnosed HIV+	0.01	0.02
γ_2	Evolution from healthy diagnosed HIV+ to sick diagnosed HIV+	0.001	0.002
ω_1	Death rate of sick undiagnosed HIV+	0.01	0.02
ω_2	Death rate of sick diagnosed HIV+	0,001	0.002
δ_1	Diagnostic rate of healthy HIV+	0,5	1
δ_2	Diagnostic rate of sick HIV+	0,5	1

did not exceed 100 % of the baseline value, both by default or excess. In Table 1 are shown the baseline values considered and the maximum value for each one.

It is noteworthy that the simulations were made up to 1500 months, which is equivalent to 125 years, as shown in Fig. 1. This time scale is not unusual if taken as reference works like [1, 4, 6], where simulations at 100, 60, and 80 years, respectively, were made. Moreover, in [11], the simulation was performed up to 1000 years.

Another goal of simulations was to see which parameters were most influential in the dynamic. It was then determined that the transmission rate by contact with a healthy diagnosed HIV positive, denoted by β_2, was the parameter that generated major changes in the populations behavior, situation reflected in Fig. 1. In this case, the minimum value of β_2 was 0.00000444, associated to the curve drawn with black solid line, which is not seen in the graph, since the value of R_0 is 0.8768, less than 1, and makes no infected populations thrive. The maximum value of β_2 is 0.00003919 and has an associated graph drawn with green dots, in this case the value of R_0 is 3.9140.

3 Control Model

3.1 Background

In a previous work, one of the authors studied the model (1) incorporating optimal control strategies based on diagnosis solely [12]. In terms of structure, the model is the same, but an important assumption do change: the diagnosis rate of healthy and sick HIV positives corresponds to the control functions. That is, δ_1 and δ_2 were

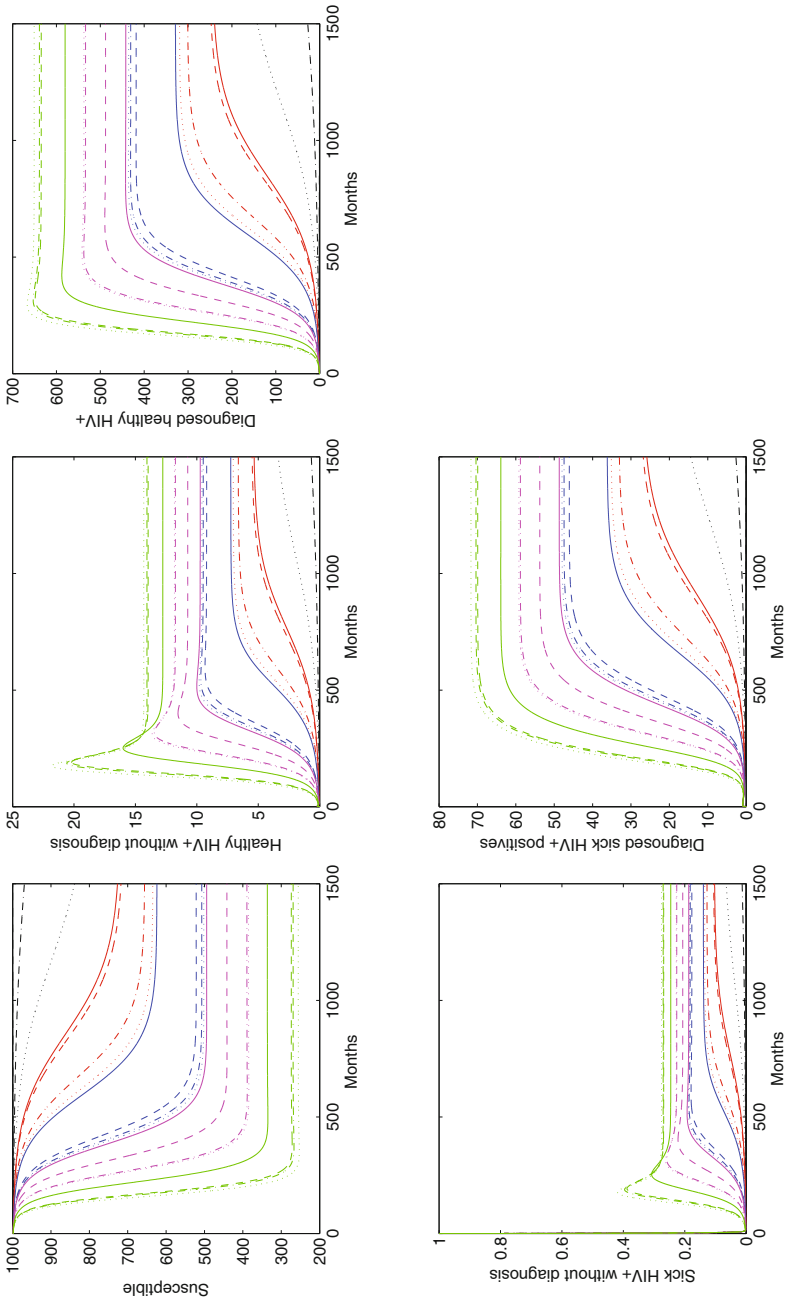


Fig. 1 Simulation of model (1) varying β_2 in the interval $[0, 4 \times 10^{-4}]$

denoted by $u_1 = u_1(t)$ and $u_2 = u_2(t)$, respectively. The model (3) describes the above consideration:

$$\begin{cases} \dot{x} = \Lambda - \beta_1 x y_1 - \beta_2 x y_2 - \beta_1 x z_1 - \beta_3 x z_2 - \mu x \\ \dot{y}_1 = \beta_1 x y_1 + \beta_2 x y_2 + \beta_1 x z_1 + \beta_3 x z_2 - u_1 y_1 - \theta y_1 \\ \dot{y}_2 = u_1 y_1 - \phi y_2 \\ \dot{z}_1 = \gamma_1 y_1 - u_2 z_1 - \rho z_1 \\ \dot{z}_2 = \gamma_2 y_2 + u_2 z_1 - \eta z_2. \end{cases} \tag{3}$$

Optimal control theory and the Pontryagin maximum principle were used to study the effect of diagnosis on HIV control. Model (3) analysis left very important lessons, including:

1. The diagnostic strategy alone was not sufficient to significantly reduce the disease.
2. Control based solely on HIV diagnosis is not effective in controlling the transmission of the disease and on the contrary, it is necessary to resort to other strategies such as prophylaxis or treatment.

3.2 Model with Antiviral Treatment

Primary goal of treatment is to reduce morbidity and mortality associated with HIV, which is achieved by focusing on inhibiting virus replication. During the past 20 years, the "Panel on Antiretroviral Guidelines for Adults and Adolescents" has made changes in recommendations about when to start therapy based on clinical trials, cohort data, and therapeutic options available at the time of each review. The standard procedure of the panel is to make recommendations when there is an agreement among two-thirds of the members; however, for the 2011 version of the guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents [9], it has not been possible to reach an agreement about when to start therapy. According to the panel, controlled trials provide evidence that applying treatment in patients with CD4 counts <350 cells/mm^3 brings benefits. The panel recommends treatment for patients with cell counts between 350 and 500 cells/mm^3. Finally, for patients with cell counts >500 cells/mm^3, the panel members are divided; out of this, 50 % favor starting therapy at early stages, while the other 50 % consider it optional [9].

Considering the last part of Sect. 3.1, and the system (1), it is proposed as a control problem which takes into account antiviral treatment that corresponds to the inclusion of a control function which varies with time. The goal is to find an optimal function in terms of reducing the impact of disease. Next assumptions are made:

- The control function $u = u(t)$, with $0 \leq u \leq 1$ models treatment application in diagnosed HIV positives. Thus, $u = 0$ means no treatment and $u = 1$ indicates full treatment is applied, i.e., the probability of infecting a healthy person is zero.

- Viral load decreases in the diagnosed population under antiretroviral treatment, resulting in a reduction of transmission rates.
- The reduction factors $(1 - f_1 u)$ and $(1 - f_2 u)$ denote the effectiveness of antiretroviral treatment in healthy and sick diagnosed HIV positives, where f_1 and f_2 are related to antiviral dose administration [2].

Under the above assumptions, the model (1) takes the following form:

$$\begin{cases} \dot{x} = \Lambda - \beta_1 x y_1 - \beta_2(1 - f_1 u) x y_2 - \beta_1 x z_1 - \beta_3(1 - f_2 u) x z_2 - \mu x \\ \dot{y}_1 = \beta_1 x y_1 + \beta_2(1 - f_1 u) x y_2 + \beta_1 x z_1 + \beta_3(1 - f_2 u) x z_2 - \delta_1 y_1 - \theta y_1 \\ \dot{y}_2 = \delta_1 y_1 - \phi y_2 \\ \dot{z}_1 = \gamma_1 y_1 - \delta_2 z_1 - \rho z_1 \\ \dot{z}_2 = \gamma_2 y_2 + \delta_2 z_1 - \eta z_2. \end{cases} \tag{4}$$

The idea now is to formulate an *optimal control problem* with model (4), that allows to determine an optimal function $\mathrm{u}^* = \mathrm{u}^*(\mathrm{t})$ to be the most effective treatment scheme to reduce disease transmission. For this, it is considered a cost functional to be minimized, denoted by J, which collects information of antiviral treatment and treated diagnosed populations. Such functional arises as follows:

$$J(u) = \int_0^\tau \left(A_1 y_2 + A_2 z_2 + \frac{A_3}{2} u^2 \right) \mathrm{d}t. \tag{5}$$

Functional (5) is subject to the initial value problem consisting of system (4) and initial conditions (2). The goal is to find an optimal function $u^* \in \Gamma$ such that $J(u^*) \leq J(u)$, for all $u \in \Gamma$ and where Γ is the *set of accessibility,* given by

$$\Gamma = \{u : u \in L^2([0, \tau]), 0 \leq u \leq 1\}.$$

To find this function, the *Pontryagin maximum principle is used,* whereby minimizing the functional J is equivalent to minimizing the Hamiltonian function H, given by

$$H(\mathbf{x}) = I + \mathbf{L} \cdot \mathbf{F}(\mathbf{x}),$$

where $\mathbf{x}$ is the vector of state variables, I is the integrand in J, $\mathbf{F}$ is the vector field in (4), and $\mathbf{L}$ is a vector of adjoint variables. Explicitly,

$$\begin{aligned} H(\cdot) = \; & A_1 y_2 + A_2 z_2 + \tfrac{A_3}{2} u^2 + L_1(\Lambda - \beta_1 x y_1 - \beta_2(1 - f_1 u) x y_2 - \beta_1 x z_1 \\ & - \beta_3(1 - f_2 u) x z_2 - \mu x) + L_2(\beta_1 x y_1 + \beta_2(1 - f_1 u) x y_2 + \beta_1 x z_1 \\ & + \beta_3(1 - f_2 u) x z_2 - \delta_1 y_1 - \theta y_1) + L_3(\delta_1 y_1 - \phi y_2) + L_4(\gamma_1 y_1 \\ & - \delta_2 z_1 - \rho z_1) + L_5(\gamma_2 y_2 + \delta_2 z_1 - \eta z_2) + M_1 u + M_2(1 - u) \end{aligned}$$

where $M_{1,2}$ are nonnegative penalty multipliers satisfying the following conditions:

$$M_1 u = 0 \text{ and } M_2(1 - u) = 0, \tag{6}$$

Used to guarantee that $0 \leq u \leq 1$. In addition, the variables L_i, for $i = 1, \ldots, 5$, are adjoint variables that satisfy the final-value problem consisting of the system:

$$\begin{cases} \dot{L}_1 = (L_1 - L_2)(\beta_1(y_1 + z_1) + \beta_2(1 - f_1 u)y_2 + \beta_3(1 - f_2 u)z_2) + \mu L_1 \\ \dot{L}_2 = (L_1 - L_2)\beta_1 x + (\delta_1 + \theta)L_2 - \delta_1 L_3 - \gamma_1 L_4 \\ \dot{L}_3 = (L_1 - L_2)(1 - f_1 u)\beta_2 x + \phi L_3 - \gamma_2 L_5 - A_1 \\ \dot{L}_4 = (L_1 - L_2)\beta_1 x + (\delta_2 + \rho)L_4 - \delta_2 L_5 \\ \dot{L}_5 = (L_1 - L_2)(1 - f_2 u)\beta_3 x + \eta L_5 - A_2, \end{cases} \tag{7}$$

and the final conditions $L_i(\tau) = 0$, for $i = 1, \ldots, 5$. To characterize the optimal control, it is solved the first-order condition $\frac{\partial H}{\partial u} = 0$ to have,

$$u = \frac{(L_2 - L_1)(\beta_2 f_1 y_2 + \beta_3 f_2 z_2) - M_1 + M_2}{A_3}.$$

Using the penalty conditions given in (6), it follows that an appropriate way to characterize u^* is

$$u^* = \max\left(0, \min\left(\frac{(L_2 - L_1)(\beta_2 f_1 y_2 + \beta_3 f_2 z_2)}{A_3}, 1\right)\right). \tag{8}$$

The described process is developed based on control theory applied to epidemics, described by several authors such as [5].

3.3 Control Model Numerical Results

The effect of optimal treatment schemes u^* on the dynamic is shown in the Figs. 2–4, which are the result of simulating the boundary value problem formed by (4), (7), and (8) with initial conditions (2) and final conditions $L_i(\tau) = 0$, for $i = 1, \ldots, 5$. Parameter baseline values are used according to Table 1, variations are performed on the weighting A_3, which is the denominator of the optimal control (8) and the cost-related parameter. Twenty random values were assigned to A_3, generated using a uniform probability distribution. In all cases, the smaller A_3 values correspond to the black line, whereas the green lines are associated with large values.

Preliminary simulations were made varying the values of A_3 between 0 and 10,000, while f_1 and f_2 effectiveness were set at 1. It was noted that to perceive dynamic changes such high weighting values for A_3 were not necessary. Thus, from the numerical results, the variation to A_3 was made between 0 and 4500, retaining effectiveness f_1 and f_2 in 1. It was obtained in Fig. 2, in which it is observed that HIV positives do not prosper (black solid line), while control levels remain between 50 and 70 %. However, when the value of A_3 decreases (green dotted line), so do the control and the susceptible population, while increases the HIV positives. It is

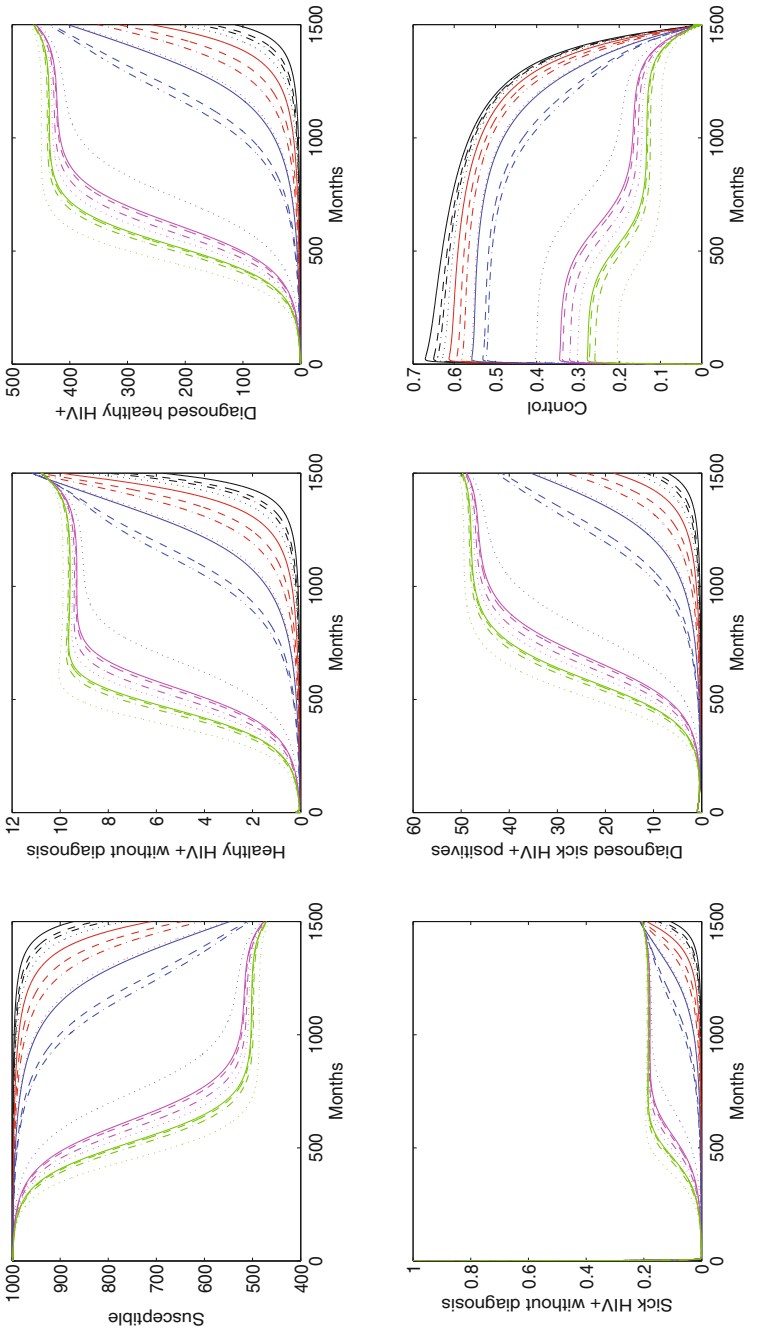


Fig. 2 Simulation of system (4), varying A_3 between 0 and 4500, with $f_1 = f_2 = 1$ as effectiveness values

important to say that at the end of the time scale, black curves give the feeling of apparent prosperity for the HIV positives, but this is because the treatment is no longer applied. It is also important to remember that final conditions for adjoint variables are zero, i.e., $L_i(\tau) = 0$.

In Fig. 2, a change of concavity in the graphs of the five populations was observed. From another perspective, a kind of jump was seen between some curves. This situation is then studied with another simulation, in which the variation now for A_3 is given between 2100 and 2400 to obtain Fig. 3, in which no news were found; it just happened that by decreasing the range of variation of the values of A_3, the curves were closer together. Of course, infected populations prosper because the application of the control did not exceed 50 % .

Finally, Fig. 4 shows that even the high values of A_3 reduces both the control and the susceptible, whereas infected populations thrive. However, an interesting situation with the solid black curve, which is obtained for the smallest random value of A_3 is presented; it is seen, despite a permanent application of control by nearly 1200 months, that the infected populations are on the rise. This is undoubtedly due to the value of one of the effectiveness was reduced, as it was considered $f_1 = 0.5$, while it remained $f_2 = 1$. Simulation to $f_1 = 1$ and $f_2 = 0.5$ was also performed, but the results were similar to those observed in Fig. 2.

4 Conclusions

After making different simulations of model (1), assigning to each parameter multiple values, it was concluded that β_2, the transmission rate by contact with a healthy diagnosed HIV positives, proved to be the parameter that caused major alterations in the population behavior. It is worthwhile to mention that this situation is not influenced by high numerical values; indeed, in Table 1 precisely β_2 is one of the parameters of lesser value and has not been giving special priority in the model; even in this order of ideas, β_1 was expected to be the one which would generate this behavior. Therefore, the influence of β_2 indicates that diagnosed HIV positives plays an important role on HIV/AIDS epidemics, and prevention strategies should be strengthened on this population.

Incorporation of dose parameters in optimal control models, as illustrated with f_1 and f_2 in the model (8) strengthen the numerical results. Compared with previous studies of the authors, they had just explored simulations as shown in Figs. 2 and 3. This results could help physicians and health services to improve their decision making on antiviral treatment schemes.

This study allows to understand how diagnosis and antiviral treatment considered together is an important issue on HIV transmission and evolution to AIDS. Researchers should incorporate this assumptions into their studies to provide more accurate interpretations on the phenomenon. Further studies should be focused on how individual antiviral schemes impact disease's behavior when generalized into populations with similar characteristics, and evaluate alternate (suboptimal) schemes as a way to improve patients' quality of life and reduce interventions cost.

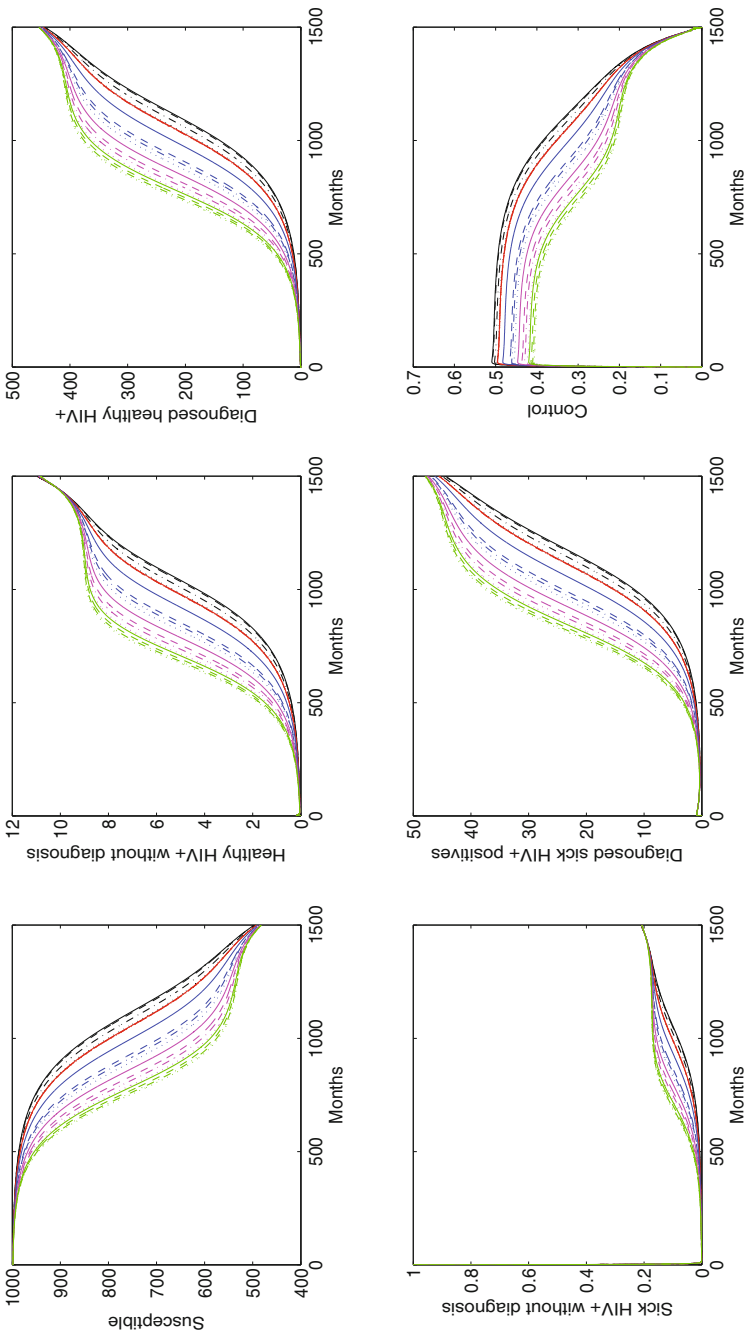


Fig. 3 Simulation of control model (4), with A_3 between 2100 and 2400, $f_1 = f_2 = 1$

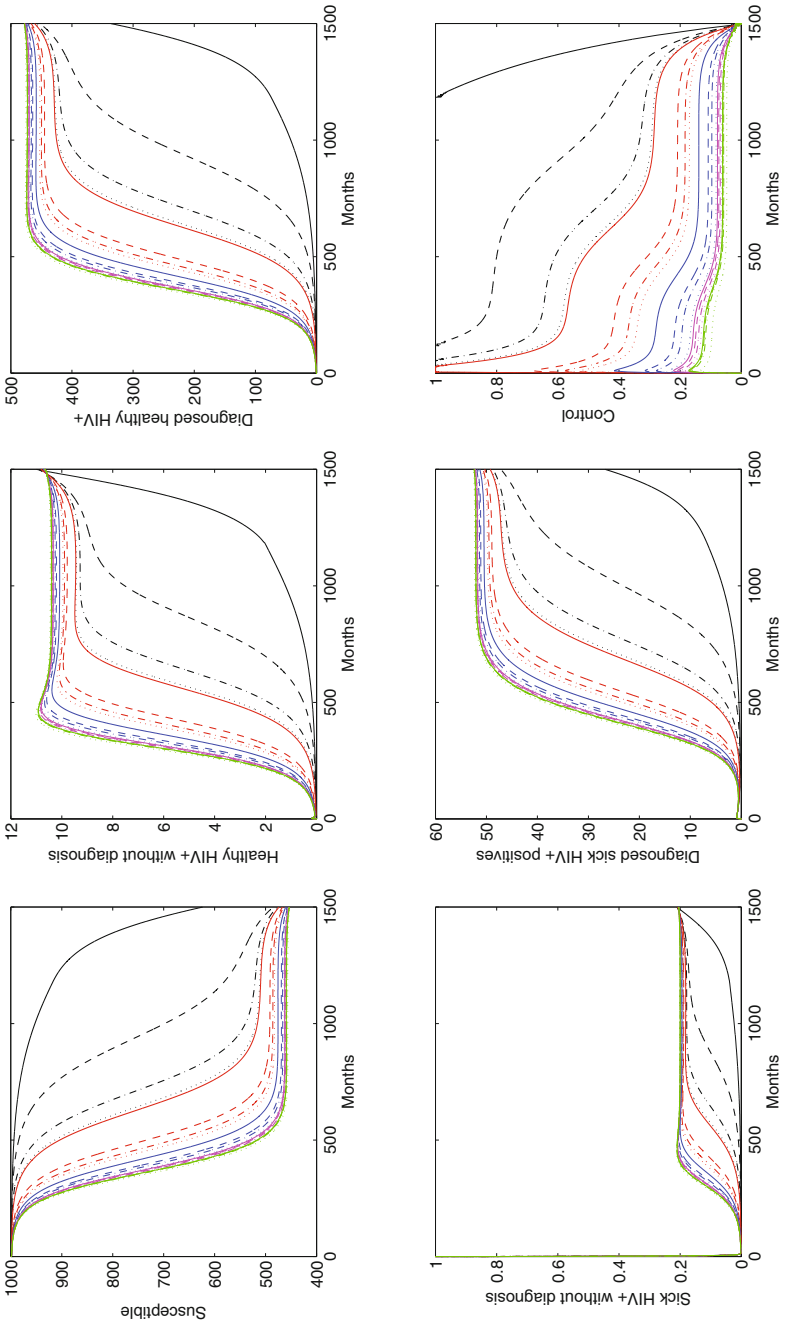


Fig. 4 Simulation of control model (4), with A_3 between 0 and 4500, $f_1 = 0.5$ and $f_2 = 1$

References

1. Al-Sheihk, S.: Stability analysis of an HIV/AIDS epidemics model with screening. Int. Math. Forum. **6**(66), 3251–3273 (2011)
2. Banks, H.: Modeling HIV immune response and validation with clinical data. J. Biol. Dyn. **2**(4), 357–385(2008)
3. Ding, A., Wu, H.: Relationships between antiviral treatment effects and biphasic viral decay rates in modeling HIV dynamics. Math. Biosci. **160**, 63–82 (1999)
4. Estrada, H., Mantilla, I.: Estudio de un modelo matemático para la propagación del SIDA. Rev. Acad. Colomb. Cien. **19**(72), 107–116 (1994)
5. Greenhalgh, D.: Some results on optimal control applied to epidemics. Math. Biosci. **88**, 125–158 (1986)
6. Levin, B., Bull, J., Stewart, F.: The intrinsic rate of increase of HIV/AIDS: epidemiological and evolutionary implications. Els. Math. Biosci. **132**, 69–96 (1996)
7. ONUSIDA. Informe sobre la epidemia mundial del SIDA (2013)
8. Orellana, J.M.: Optimal drug scheduling for HIV therapy efficiency improvement. Biomed. Signal Process. Control **6**, 379–386 (2004)
9. Panel on Antiretroviral Guidelines for Adults and Adolescents: Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents. Department of Health and Human Services, pp. 1–166 (2011)
10. Roshanfekr, M., Farahi, M.H., Rahbarian, R.: A different approach of optimal control on an HIV immunology model. Ain Shams Eng. J. http://dx.doi.org/10.1016/j.asej.2013.05.004 (2013)
11. Sharomi, O., Podder, C., Gumel, A.: Mathematical analysis of the transmission dynamics of HIV/TB coinfection in the presence of treatment. Math. Biosci. Eng. **5**(1), 145–174 2008
12. Toro-Zapata, H.D., Mesa-Mazo, M.J., Prieto-Medellín, D.A.: Modelo de simulación para la transmisión del VIH y estrategias de control basadas en diagnóstico. Revista de Salud Pública **16**(1), 139–152 (2014)
13. UNAIDS: AIDS by the numbers (2013)
14. UNAIDS, WHO: Global Report. UNAIDS report on the global AIDS epidemic (2013)
15. WHO: WHO case definitions of HIV for surveillance and revised clinical staging and immunological classification of HIV-related disease in adults and children. WHO Library Cataloguing-in-Publication Data. ISBN 978 92 4 159562 9 (2007)
16. Zarei, H., Kamyad, A.V., Effati, S.: Multiobjective optimal control of HIV dynamics. Math. Probl. Eng. Article ID 568315, 29 (2010). doi:10.1155/2010/568315

Orbital Relative Movement Applied the Formation Flight of Artificial Satellites Around the Earth

Jorge Soliz and Daniel Molano

Abstract In the past years, space missions have become a difficult task due to the ambition of the experiments. For this reason, the techniques needed to achieve the objectives are becoming more complex, as the needs of autonomy, accuracy, flexibility, etc., are very important. One solution found to accomplish these requirements is to send a formation of multiple satellites. This is a good solution because it lets us to make different measurements simultaneously, improving the accuracy and reliability, and also it us lets to build cheaper and smaller satellites, which not only improves the performance of each satellite but also the flexibility of the entire mission. There has been a considerable interest in distributing the functions of a single large satellite among several small cooperative units. Many potential applications of this enabling technology exist, one of which is to improve the performance of the Earth observation. A cluster of satellites will be able to synthesize a much larger aperture than can be achieved with a single platform, thus providing significant increases in image resolution through interferometry.

Keywords Formation flight · Dynamical systems · Celestial mechanics · Relative movement

1 Introduction

This work was inspired by the research of Koon et al. [4]. The research developed dynamical systems techniques appropriate to the near Earth case and found a family of candidate reference orbits whose nearby orbits may support formation flight.

Two important predecessors in the formation flying field are the European Space Agency (ESA) and the National Aeronautics and Space Administration (NASA).

J. Soliz (✉)
Universidad Sergio Arboleda, Bogotá, Colombia
e-mail: jorge.soliz@usa.edu.co

D. Molano
Universidad Sergio Arboleda, Universidad Nacional de Colombia, Bogotá, Colombia
e-mail: damolanom@unal.edu.co

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_19

Both, with their missions of multiple satellites are contributed to the development of this concept and the improvement of the necessary technology.

The first mission of this type was Deep Space 1 (DS1), launched in 1999 by NASA. The DS1 developed different important technologies for future missions of multiple satellites, like an autonomous agent architecture, on-board deduction and search, and goal-directed closed-loop commanding.

The orbital relative movement and the formation flight was studied in past years as it is seen in refs. [5–8]. We have used Routh reduction and Poincaré section techniques where a procedure was developed for locating orbits such that the cluster of satellites remains close for many years, with very little dispersing, even with no controls. Rather than using orbital elements, our analysis is done directly in physical space which makes the connection with physical requirements more direct. These orbits are called quasi-periodic orbits and are obtained due to Poincaré's map where one can find a fixed stable point for which passes the desired orbit. This methodology of finding dynamically favorable orbits, if coupled with control and optimal control, may provide an effective way to deal with the maintenance and reconfiguration of formation flight of near Earth satellites.

In the satellite motion, the first perturbation which will be considered is the so-called J_2 effect. This is due to the fact that the Earth is not an homogeneous perfect sphere, but its a geoid. This means that our planet has an irregular shape which is characteristic of the Earth alone. The most evident difference from the spherical shape is the flattening at the poles. Therefore, the Earth can be modeled not as a sphere, but as a spheroid (an ellipsoid of revolution). This is the cause of the J_2 perturbation, which results to be the most important perturbation at almost every altitude.

2 Movement Equations

The Routh reduction technique was used to rewrite the equations of motion of the full system in a simpler form. This procedure will enable us to study first the reduced dynamics in the meridian plane of the satellite before dealing with the dynamics in the longitudinal direction [4].

Recall that in spherical coordinates (ρ, θ, ϕ), distance from the origin to a given point (satellite), latitude, longitude, respectively. The potential energy including the J_2 effect is given by

$$U = -\frac{\mu}{\rho} + \frac{\mu R_e^2 J_2}{\rho^3}\left(\frac{3}{2}\cos^2\theta - \frac{1}{2}\right),$$

where μ is the gravitational constant of the Earth, R_e is the radius of the Earth ($\mu = GM_e = 3.986005 \times 10^{14}\ \mathrm{m}^3/\mathrm{s}^2$), ($R_e = 6378140$ m), and J_2 is the second zonal harmonic coefficient due to the oblateness of the Earth ($J_2 = 0.00108263$).

The potential energy equation with J_2 in $U(r,z)$,

$$U(r,z) = -\frac{\mu}{(r^2+z^2)^{1/2}} + \frac{\mu R_e^2 J_2}{(r^2+z^2)^{3/2}}\left(\frac{3}{2}\frac{z^2}{r^2+z^2} - \frac{1}{2}\right),$$

following [1],

$$R = \frac{1}{2}(\dot{\rho}^2 + \rho^2\dot{\theta}^2) - \frac{H_z^2}{2\rho^2\sin^2\theta} - U(\rho,\theta).$$

In the rectangular coordinates (r,z), Routhian function becomes

$$R = \frac{1}{2}(\dot{r}^2+\dot{z}^2) - \frac{H_z^2}{2r^2} - \left[-\frac{\mu}{(r^2+z^2)^{1/2}} + \frac{\mu R_e^2 J_2}{(r^2+z^2)^{3/2}}\left(\frac{3}{2}\frac{z^2}{r^2+z^2} - \frac{1}{2}\right)\right],$$

where $\rho^2 = r^2+z^2$ and $\cos\theta = z/\rho$. The reduced equation are then given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial R}{\partial \dot{r}}\right) = \frac{\partial R}{\partial r}, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial R}{\partial \dot{z}}\right) = \frac{\partial R}{\partial z}.$$

Equivalently, these equations are given by [4]:

$$\ddot{r} = H_z^2\frac{1}{r^3} - \mu\frac{r}{(r^2+z^2)^{3/2}} - \frac{3\mu R_e^2 J_2}{2}\frac{r}{(r^2+z^2)^{5/2}} + \frac{15\mu R_e^2 J_2}{2}\frac{rz^2}{(r^2+z^2)^{7/2}},$$

$$\ddot{z} = -\mu\frac{z}{(r^2+z^2)^{3/2}} - \frac{3\mu R_e^2 J_2}{2}\frac{z}{(r^2+z^2)^{5/2}} + \frac{3\mu R_e^2 J_2}{2}\frac{(3z^2-2r^2)z}{(r^2+z^2)^{7/2}},$$

where

$$\ddot{r} = f(r,z), \quad \ddot{z} = g(r,z), \quad \ddot{\phi} = \frac{H_z}{r^2},$$

and the energy is given by

$$E = \frac{1}{2}(\dot{r}^2+\dot{z}^2) + \frac{H_z^2}{2r^2} + U(r,z).$$

2.1 The Poincaré Map

The most basic tool for studying the stability and bifurcations of periodic orbits is the Poincaré map or first return map. The idea of the Poincaré map is quite simple: If Γ is a periodic orbit of the system:

$$\dot{\mathbf{x}} = f(\mathbf{x}),$$

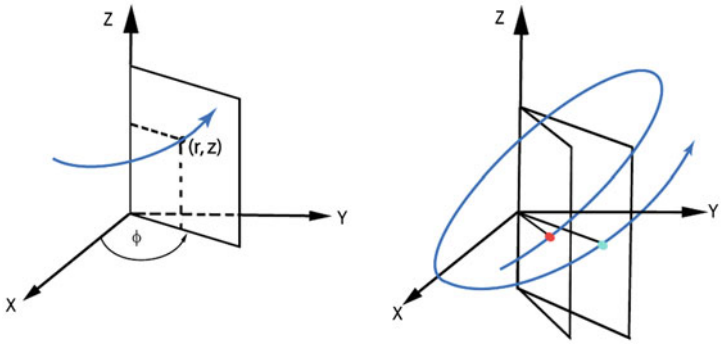


Fig. 1 The Poincaré map [4]

through the point $\mathbf{x_0}$ and Σ is a hyperplane perpendicular to Γ at $\mathbf{x_0}$, then for any $\mathbf{x} \in \Sigma$ sufficiently near $\mathbf{x_0}$, the solution of the last equation through $\mathbf{x}$ at $t = 0$, $\phi_t(\mathbf{x})$, will cross Σ again at a point $\mathbf{P}(\mathbf{x})$ near $\mathbf{x_0}$. The mapping $\mathbf{x} \rightarrow \mathbf{P}(\mathbf{x})$ is called the Poincaré map (Fig. 1).

The Poincaré map can also be defined when Σ is a smooth surface, through a point $\mathbf{x_0} \in \Gamma$, which is not tangent to Γ at $\mathbf{x_0}$. In this case, the surface Σ is said to intersect the curve Γ transversally at $\mathbf{x_0}$.

After performing Routh reduction, we can use the method of Poincaré section to find the initial conditions for orbits that are dynamically favorable to the formation flight.

According to [4] the energy E is conserved (in the meridian variables (r, z)), the constant energy surface for the reduced system is three-dimensional and the hyperplane $z = 0$ can be used as the transversal plane to obtain the two-dimensional Poincare section, (Fig. 2). Notice that the plane $z = 0$ is the plane of the Earth's equator.

By studying this Poincaré section (Fig. 2) and looking for the stable fixed point, we can find the pseudo-circular orbit (which corresponds to the fixed point in the middle of Fig. 2) whose nearby orbits can be used for formation flight.

3 Simulation

Due to the reduced equations, we can fix values and obtain $(r, \dot{r})$ of the Poincaré section that gives the initial conditions $(r, z, \phi, \dot{r}, \dot{z}, \dot{\phi})$ for an orbit of the full system. This is because $z = 0$ and $\dot{z} = \dot{z}(r, z, \phi, \dot{r}, \dot{z}, \dot{\phi})$ (where $\dot{z} > 0$) and $\dot{\phi} = H_z/r^2$ can be computed from the fixed energy E and the fixed z-component of the angular momentum H_z once $(r, \dot{r})$ are known. Also, since ϕ is ignorable, it can be chosen arbitrary. For convenience sake, we can set $\phi = 0$ at $t = 0$. Hence, $(r, \dot{r})$ (or more fully $(r, 0, \phi, \dot{r}, E, H_z)$) provides all the initial conditions for an orbit of the full system.

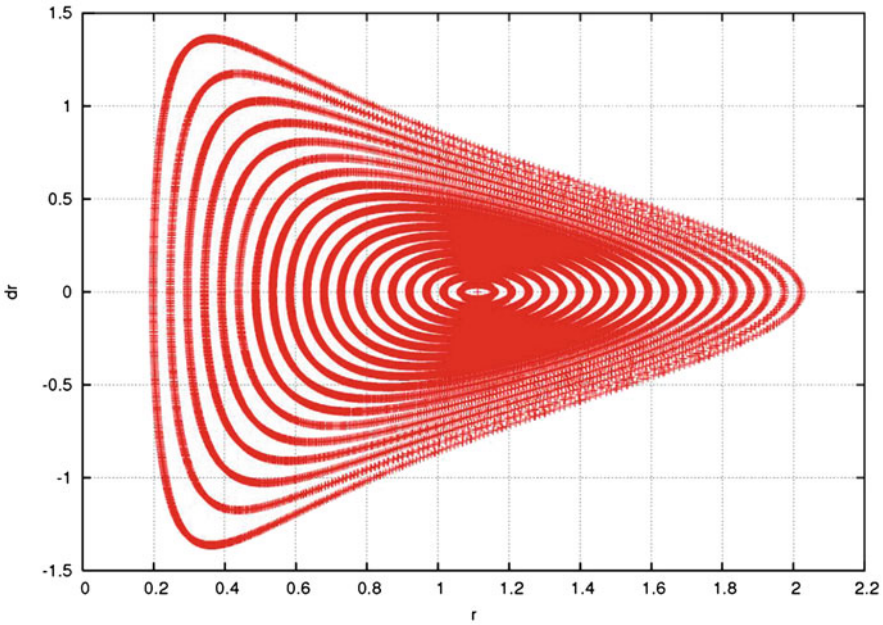


Fig. 2 Poincaré section of $(r, \dot{r})$ at $z = 0$, $E = -0.45$, $H_z^2 = 0.3$

Table 1 Initial conditions for the cluster of satellites

Sat	$r\ (DU)$	$\dot{r}\ (DU/TU)$	ϕ	$t\ (TU)$
1	1.11133496883	0.0	1×10^{-5}	1×10^{-5}
2	1.11134196883	0.0	0.0	1×10^{-5}
3	1.11133496883	0.0	-1×10^{-5}	1×10^{-5}

3.1 Triangular Cluster Near the Pseudo-Circular Orbit

By using the stable fixed point and the points nearby as well as making slight changes in the longitudinal angle ϕ (and possibly in the time t), we can construct different kinds of cluster which will remain together after many years (corresponding to thousands of revolutions around the Earth).

For example, if we fix $E = -0.45$, $H_z^2 = 0.3$ (for example), the fixed point for the Poincaré section at $z = 0$ will be $(r_f, 0)$, where $r_f = 1.11133496883$ fixed point this is about 710 km above the Earth.

The following initial conditions give a triangular cluster (isosceles triangle, with sides 80, 80, and 140 m approximately; Tables 1 and 2).

Recall that the length units have been chosen to make the radius of the Earth 6.4×10^6 m equal to $1DU$ and $1TU$ is equivalent to 806.810 s.

The evolution of these three satellites in a triangular cluster were plotted in a frame whose origin is at their instantaneous barycenter, with the yz-plane orthogonal to the

Table 2 Initial conditions in rectangular coordinates

Sat	$x\,(DU)$	$y\,(DU)$	$z\,(DU)$	$\dot{x}\,(DU/TU)$	$\dot{y}\,(DU/TU)$	$\dot{z}\,(DU/TU)$
1	1.111334	0.111133×10^{-4}	0.810878×10^{-05}	$-0.108502 \times 10^{-04}$	0.492851	0.810878
2	1.111341	0.0	0.810873×10^{-05}	$-0.592164 \times 10^{-05}$	0.492847	0.810873
3	1.111334	$-0.111133 \times 10^{-04}$	0.810878×10^{-05}	$-0.993194 \times 10^{-06}$	0.492851	0.810878

Table 3 Initial conditions in rectangular coordinates

Sat	$x\,(DU)$	$y\,(DU)$	$z\,(DU)$	$\dot{x}\,(DU/TU)$	$\dot{y}\,(DU/TU)$	$\dot{z}\,(DU/TU)$
1	1.1113349	0	0	0	0.492851	0.746675

Table 4 Maximum and minimum value of the orbital elements mains for the fixed point

r	Altitude (km)	Eccentricity	Inclination (degree)
1.11133496883	703.632–717.350	0.00031–0.000929	58.678–58.712

line of sight, the x-axis pointing toward the center of the Earth, and the y-axis and the z-axis pointing toward the (instantaneous) west and north, respectively.

Figure 3 (left side) shows the trajectories of these three satellites projected onto the yz-plane for 100 revolutions around the Earth (about a week). Figure 3 (right side) shows the trajectories of the same cluster of satellites cluster in the yz-plane for 5000 revolutions around the Earth (about a year), notice how small the dispersion is during a year (see Fig. 3).

3.2 Orbital Elements Range

By having the initial condition $(x, y, z, \dot{x}, \dot{y}, \dot{z})$ of the stable fixed point (see Table 3), the orbital elements are calculated for a longtime (5000 revolutions around the Earth).

As one can see, the mean orbital elements do not show large variations (Table 4).

4 Method of Monitoring the Formation of Satellites

In this section, we studied the behavior of the cluster of satellites. First, we made an area analysis of the triangle that forms the three satellites. Second, we studied the behavior of the relative distances between the satellites.

4.1 Area

The triangle's area formed by the cluster of satellites will be calculated for knowing how the area changes along time. In Fig. 4, we can see the change of the triangular area in four revolutions. The area changes between 16 and 3445 m^2, approximately.

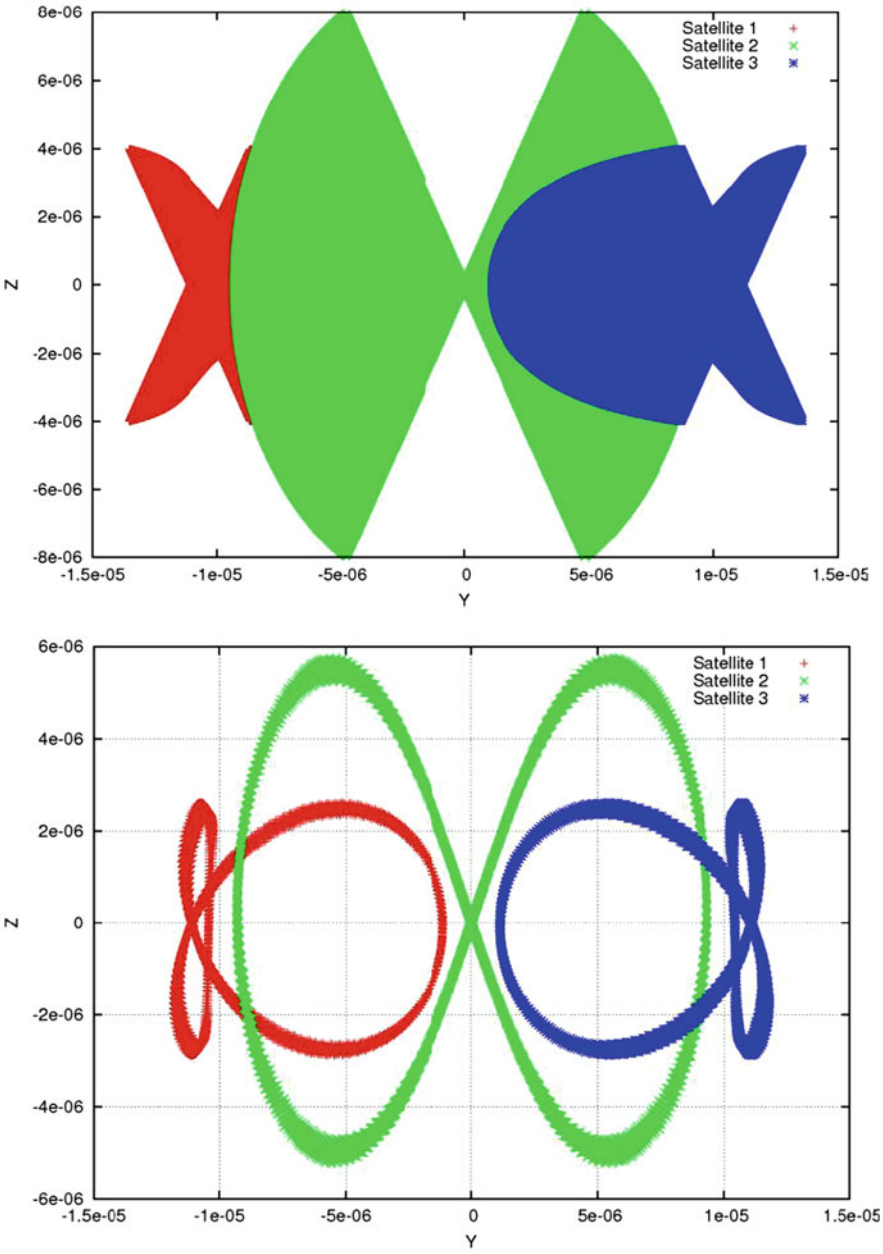


Fig. 3 The trajectories of three satellites in the yz-plane for 100 (*left*) and 5000 revolutions around the Earth (*right*)

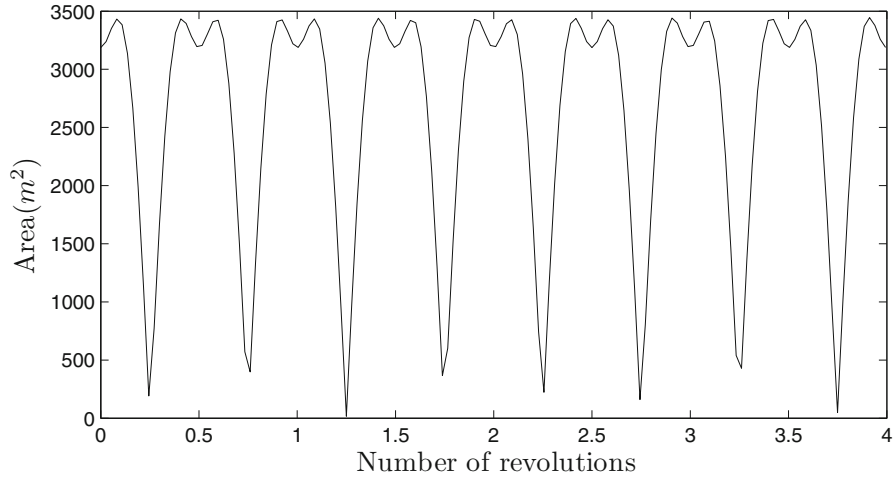


Fig. 4 Area formed by the cluster in four revolutions

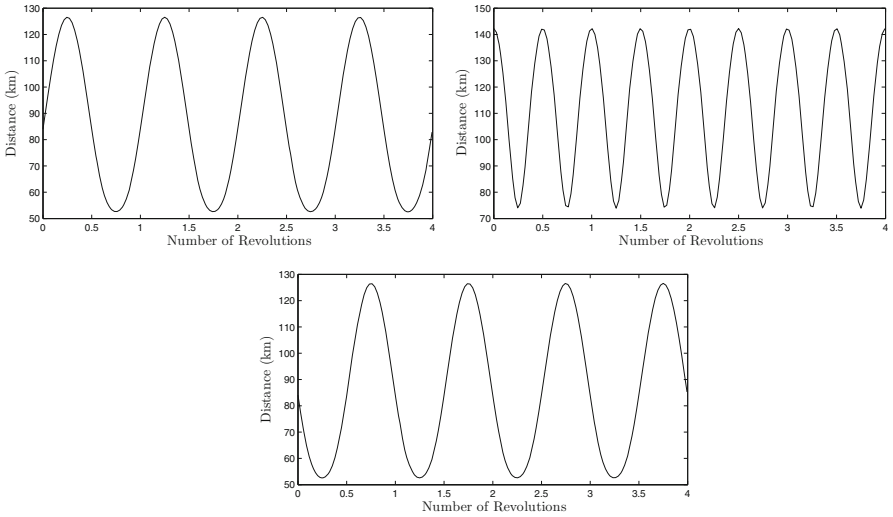


Fig. 5 Relative distance between the satellites 1 and 2, 1 and 3, 2 and 3 respectively, for four revolutions

4.2 Relative Distances

In this section, we realized the analysis of the relative distances between satellites and their performance along the time. The relative distance for four revolutions is plotted in the following figure (Fig. 5).

Briefly, the maximum and minimum variation are given in Table 5:

Table 5 Values of the maximum and minimum distance between two satellites

Satellites	Max. distance (m)	Min. distance (m)
1 and 2	126.5	52.6
2 and 3	126.5	52.6
1 and 3	142.2	73.9

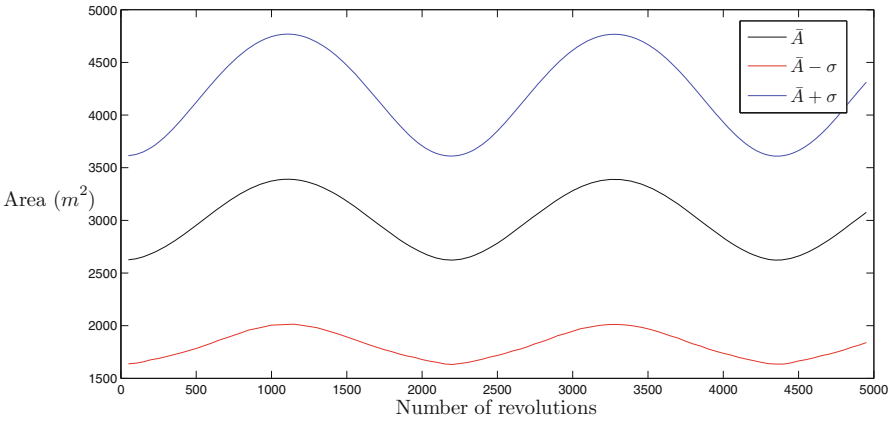


Fig. 6 Area formed by the cluster

4.3 Statistical Analysis

By making an statistical analysis for understanding the behavior of the cluster's area in the next revolutions of Fig. 4, we will get the mean area "$\bar{A}$" (black line) and the standard deviations "σ" (red line for maximun value and blue line for minimun value). These were calculated each 50 revolutions (total time of simulation, 5000 revolutions). The result gives a qualitative behavior of the areas as we can see in Fig. 6.

We made the same qualitative analysis but in this case for the relative distances, see Fig. 7. Here, we have no dispersion between relative distances and this is a proof that the method works so well. However, in some applications we need that the distance between satellites be constant. For this reason, in our next chapter, we will analyze several control methods like [3] and [2].

5 Conclusions

By mean Routh reduction and Poincaré section, a procedure was developed for locating orbits such that the cluster of satellites remains close for many years, with very little dispersion [4]. The fixed points in the Poincaré section has provided periodic orbits around invariant tori. The satellites trajectories have a little dispersion on the tori very close to periodic orbit. Figures 6 and 7 have demonstrated the little dispersion already mentioned for many years.

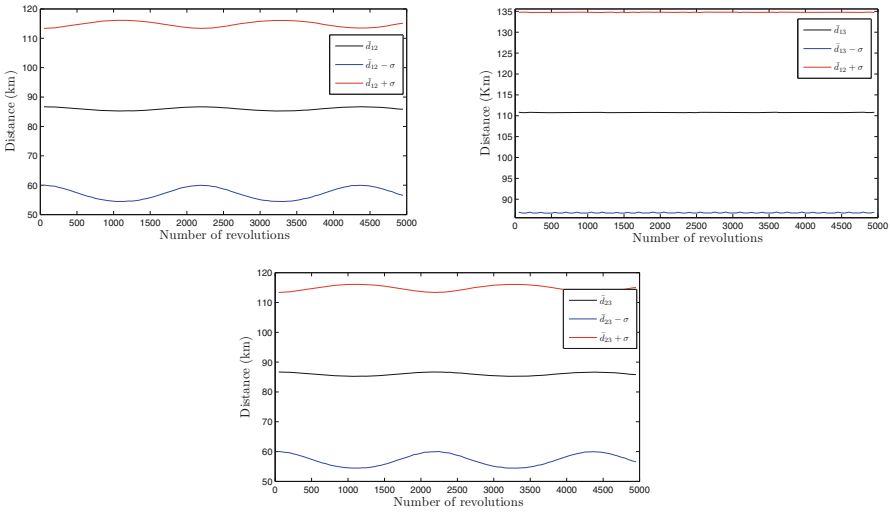


Fig. 7 Distance of the satellites 1 and 2, 1 and 3, 2 and 3 respectively for 5000 revolutions

References

1. Broucke, R.A.: Numerical integration of the periodics orbits in the main problem of artificial satellite theory. Celest. Mech. Dyn. Astron. **58**, 99–123 (1994)
2. Cai, W.-w., Yang, l.-p., Zhu, Y.-w., Zhang, Y.-w.: Optimal satellite formation reconfiguration actuated by inter-satellite electromagnetic forces. Acta Astronaut. **89**, 154–165 (2013)
3. Junge, O., Blobaum, O.: Optimal reconfiguration of formation flying satellites. Proceedings of the 44th IEEE Conference on Decision and Control, and the European Control Conference (2005)
4. Koon, W.S., Marsden, J.E., Masdemont, J.J., Murray, R.M.: J_2 dynamics and formation flight. AIAA-2001-4090 (2001)
5. Kong, E.M.C., Kwon, D.W., Schweighart, S.A., Elias, L.M.: Electromagnetic formation flight for multisatellite arrays. J. Spacecr Rockets **41**(4), 659–666 (2004)
6. Sabatini, M., Izzo, D., Palmerini, G.B.: Invariant relative satellite motion. ACT Workshop on Innovative Concepts, ESA-ESTEC 28–29 January 2008
7. Vadali, S. R., Sengupta, P., Yan, H., Alfriend, K. T.: On the fundamental frequencies of relative motion and the control of satellite formations. Proceedings the AAS/AIAA Astrodynamics Specialist Conference, paper AAS 07-427 (2007)
8. Wei, C., Park, S.-Y., Park, C.: Linearized dynamics model for relative motion under a J_2-perturbed elliptical reference orbit. Int. J. Non-Linear Mech. **55**, 55–69 (2013)

Some Mathematical Aspects in the Expanding Universe

Daniel Molano and Leonardo Castañeda

Abstract Recent astronomical observations of supernovae (SNIa) and barionic acoustic oscilations (BAO) indicate that the Universe is in an accelerated expansion period. Interpreted within the framework of general relativity (GR), the acceleration is explained by a positive cosmological constant or exotic matter models known in the literature as dark energy. However, there is an alternative approach to explain the acceleration without exotic matter models. Modifications of GR such as scalar–tensor gravity and high-order derivative gravity theories, naturally offer the explanation for the accelerated phase coming from the geometrical side. One of this higher-order theories is $f(R)$ modified gravity. In this work, we use some mathematical results concerning to the Taylor expansions of tensor fields under the action of one-parameter families of diffeomorphism in the context of $f(R)$ theories in the expanding universe. We mean gauge invariant in the sense of the second-kind gauge following the work exposed in Nakamura (Adv. Astr. 2010(576273):2010). We obtain the general gauge invariant at first-order equations in $f(R)$ gravity. As an example, we write these first-order equations in $f(R)$ gravity for a perturbed Friedmann–Lemaître–Robertson–Walker (FLRW) space-time. The gauge invariant scalar perturbations equations for perturbed FLRW are obtained explicitly in $f(R)$ gravity.

Keywords General relativity · Cosmological perturbation theory · Modified gravity · Exact solutions · Field equations · Manifold · Diffeomorfism · Pullback

D. Molano (✉)
Universidad Nacional de Colombia, Universidad Sergio Arboleda,
Observatorio Astronómico Nacional, Colombia
e-mail: damolanom@unal.edu.co

L. Castañeda
Universidad Nacional de Colombia, Observatorio Astronómico Nacional, Colombia
e-mail: lcastanedac@unal.edu.co

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_20

1 Introduction

Recent high-precision observations [1, 6, 15, 17, 18] have provided strong evidence that the Universe is in a phase of accelerated expansion. In the cosmological standard model with theory of gravity as general relativity (GR), this accelerating phase can be explained with a component of negative pressure called dark energy. However, we do not have direct evidence of dark energy until now. It has motivated alternative models [4]. The models are modifications to the Einstein–Hilbert action. An important subset of these models are called generalized gravity theories, and they are based on nonlinear Lagrangians written in the generic form $f(R)$, where f is a general differentiable function of the Ricci scalar R.

On the other hand, the cosmological perturbation theory (CPT) in GR and extended theories is a very active field of research. One of the main goals in CPT is to clarify the relation between scenarios of the early universe from probes designed with the cosmological data sets. One of the most important field to reach this objective is the cosmic microwave background (CMB) and its anisotropy temperature power spectrum [8]. The CPT is the tool to investigate the phenomenology in the CMB power spectrum and the link with the galactic correlation function, in order to understand the structure formation in the universe. CPT is a very sophisticated theory with deep consequences in applied mathematics [10, 13, 21].

2 Field Equations in the Metric Formalism

The action in the metric formalism for $f(R)$ gravity is:

$$S = \frac{1}{2k^2}\int \mathrm{d}^4x\sqrt{-g}\,f(R) + \int \mathrm{d}^4x\mathcal{L}_M(g_{\mu\nu}, \Psi_M), \tag{1}$$

where $k^2 = 8\pi G$, g is the determinant of the metric $g_{\mu\nu}$, and $\mathcal{L}_M$ is the Lagrangian for matter fields which depends on $g_{\mu\nu}$ and the matter fields Ψ_M. The Ricci scalar R is defined by the contraction of the Ricci tensor, i.e., $R = g^{\mu\nu}R_{\mu\nu}$, and the Ricci tensor is defined by $R_{\mu\nu} = R_{\mu\alpha\nu}{}^{\alpha}$ where the Riemann tensor is

$$R_{\mu\nu\rho}{}^{\sigma} = \Gamma^{\sigma}{}_{\mu\rho,\nu} - \Gamma^{\sigma}{}_{\nu\rho,\mu} + \Gamma^{\alpha}{}_{\mu\rho}\Gamma^{\sigma}{}_{\alpha\nu} - \Gamma^{\alpha}{}_{\nu\rho}\Gamma^{\sigma}{}_{\alpha\mu}. \tag{2}$$

In the case of the torsion free, the connections $\Gamma^{\alpha}{}_{\beta\lambda}$ are the usual Christoffel symbols defined in terms of the metric tensor $g_{\mu\nu}$, as:

$$\Gamma^{\alpha}{}_{\beta\gamma} = \frac{1}{2}g^{\alpha\lambda}(g_{\gamma\lambda,\beta} + g_{\lambda\beta,\gamma} - g_{\beta\gamma,\lambda}). \tag{3}$$

The field equations are derived by varying the action (1) with respect to $g^{\mu\nu}$, or equivalently $\delta S/\delta g^{\mu\nu} = 0$,

$$\Sigma_{\mu\nu} \equiv f'(R)R_{\mu\nu} - \frac{1}{2}f(R)g_{\mu\nu} - \nabla_\mu\nabla_\nu f'(R) + g_{\mu\nu}\Box f'(R) = \kappa^2 T_{\mu\nu}, \quad (4)$$

where $f'(R) \equiv \partial f/\partial R$ and $\Box \equiv \nabla^\mu\nabla_\mu$. $T_{\mu\nu}$ is the energy–momentum tensor of the matter fields defined by the variational derivative of $\mathcal{L}_M$ with respect to $g^{\mu\nu}$. Einstein gravity corresponds to $f(R) = R - 2\Lambda$ [7]. We can add a boundary term in the action (1), for details see [9] and for physical interpretations see [14].

3 Perturbation Theory

There are a few exact solutions of physical interest in GR theory (and at the most physical theories). Einstein equations are a complicated coupled system of nonlinear partial differential equations, and we do not have a general method to solve it. In our case, we want to know how "small" inhomogeneities in the Robertson–Walker Universe evolve in time. We have an exact solution (FLRW, see Sect. 5) and we want to quantify how "small" deviations in the metric field are properly described in $f(R)$ theories of gravity.

3.1 Taylor Expansion of Tensors on a Manifold

Considering the Einstein's equation:

$$\mathcal{E}(g, \tau) = 0, \quad (5)$$

where g is the space-time metric and τ is the matter distribution. Suppose that an exact solution, g_0, is known. From this, we build a one-parameter family g_λ of exact solutions,

$$\mathcal{E}(g_\lambda, \tau_\lambda) = 0. \quad (6)$$

We regard that g_λ and τ_λ depend smoothly on the parameter λ. The parameter λ is a measure of the amount by which a specific $(\mathcal{M}, g_\lambda, \tau_\lambda)$ differs from the idealized background solution. In some cases, λ is a formal parameter and for convenience one can set $\lambda = 1$ at the end[1] for the physical space-time. In other cases, λ is a parameter that depends of the physical problem [3, 21]. Sometimes, depending on the physical problem, it is more convenient to introduce two or more parameters [20]. In our case, we choose one parameter and it is purely formal.

In this case, we introduce a $(4 + 1)$-dimensional manifold $\mathcal{N}$, foliated by submanifolds diffemorphic to $\mathcal{M}$ so that $\mathcal{N} = \mathbb{R} \times \mathcal{M}$ [3]. We now want to define the perturbation in any tensor T, therefore, we must find a way to compare T_λ with T_0.

[1] like perturbation theory in quantum mechanics.

Being $T_0(p)$ and $T_\lambda(q)$ where $p \in \mathcal{M} \times \{0\}$ and $q \in \mathcal{M} \times \{\lambda\}$. T_0 and T_λ "live" in different points on $\mathcal{N}$ and for this reason we cannot compare them directly. For this, we define a diffeomorphism:

$$\begin{aligned} \mathcal{X} : \mathcal{N} &\longrightarrow \mathcal{N} \\ p &\longmapsto \mathcal{X}_p = q. \end{aligned}$$

The diffeomorphism $\mathcal{X}$ generates a vector field X^a on $\mathcal{M}$. We assume that X^a is transversal to each $\mathcal{M}_\lambda$. Now considering the pullback $\mathcal{X}^*$ of T_λ on T_0, the tensor can be Taylor expanded as [19]:

$$\begin{aligned} \mathcal{X}_\lambda^* T_\lambda &= T_0 + \lambda \mathcal{L}_X T + \frac{\lambda^2}{2} \mathcal{L}_X^2 T + \cdots \\ &= \overset{(0)}{T} + \lambda \overset{(1)}{T} + \frac{\lambda^2}{2} \overset{(2)}{T} + \cdots \end{aligned} \tag{7}$$

We denote the pulled-back $\mathcal{X}_\lambda^* T_\lambda$ on $\mathcal{M}_0$ by $\bar{T}$ and $\overset{(0)}{T} \equiv T_0$, $\overset{(1)}{T} \equiv \mathcal{L}_X T, \overset{(2)}{T} \equiv \mathcal{L}_X^2 T$. Thus, we have a representation of T_λ in p and by this way we can compare it with T_0, see Fig. 1. This defines a function between $\mathcal{M}_\lambda$ and $\mathcal{M}_0$ and such a correspondence is called *gauge choice* (of second kind).

3.2 Gauge Degree of Freedom in GR

There are two kinds of gauge transformations. The first kind of gauge transformation is intrinsic of the manifold definition, i.e., a change of coordinate transformation. The second kind of gauge transformation is a change between correspondences $\mathcal{M}_\lambda$ and $\mathcal{M}_0$. Let $\mathcal{X}$ and $\mathcal{Y}$ be two gauge choices with the above restrictions [13]. The gauge transformations $\mathcal{X}_\lambda \to \mathcal{Y}_\lambda$ are given by the diffeomorphism $\Psi_\lambda \equiv \mathcal{X}_\lambda^{-1} \circ \mathcal{Y}_\lambda$. Now,

$$\begin{aligned} {}_{\mathcal{Y}}\mathbf{T} &= \mathcal{Y}_\lambda^* \mathbf{T}|_{M_0} = \left(\mathcal{Y}_\lambda^* \left(\mathcal{X}_\lambda \circ \mathcal{X}_\lambda^{-1}\right)^* \mathbf{T}\right) \mathbf{T}|_{M_0} \\ &= \left(\mathcal{X}_\lambda^{-1} \circ \mathcal{Y}_\lambda\right)^* \left(\mathcal{X}_\lambda^* \mathbf{T}\right)|_{M_0} = \Psi_\lambda^* {}_{\mathcal{X}}\mathbf{T}, . \end{aligned} \tag{8}$$

where ${}_{\mathcal{X}}\mathbf{T} \equiv \mathcal{X}_\lambda^* T_\lambda$ and ${}_{\mathcal{Y}}\mathbf{T} \equiv \mathcal{Y}_\lambda^* T_\lambda$. And finally using the Taylor expansion (7) we have

$${}_{\mathcal{Y}}\mathbf{T}_\lambda = {}_{\mathcal{X}}\mathbf{T}_\lambda + \lambda \mathcal{L}_{\xi_{(1)}} {}_{\mathcal{X}}\mathbf{T}_\lambda + O(\lambda^2). \tag{9}$$

In Fig. 2, we can see a gauge transformation between two different gauge choices $\mathcal{X}$ and $\mathcal{Y}$.

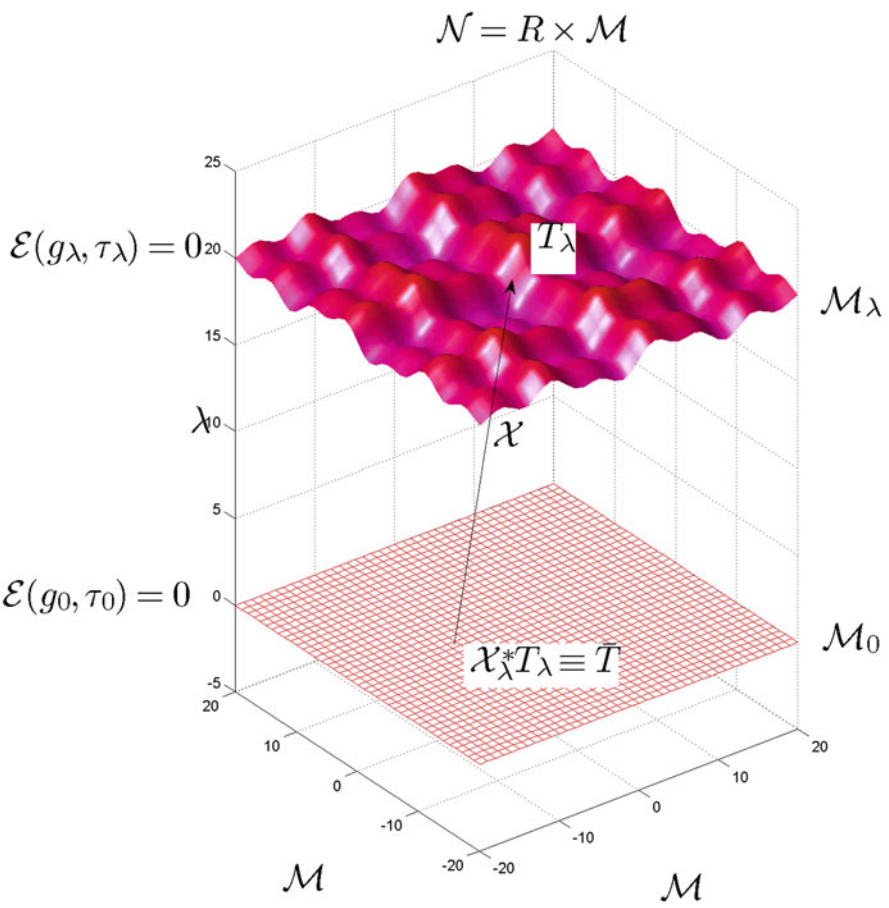


Fig. 1 The (4+1)-dimensional manifold $\mathcal{N}$. $\mathcal{M}_0$ is called *background space-time*

3.3 Gauge Invariant Variables

We assume that $(\mathcal{M}_\lambda, g_\lambda)$ is our physical space-time. All experiments and observations are done in $\mathcal{M}_\lambda$. The space-time background $\mathcal{M}_0$ is fictitious. Thus, our observations do not depend of the gauge choice, i.e., the gauge choice is not a physical degree of freedom, but we can notice from Eq. (7) that the pulled-back tensor field depends on the gauge choice. In this section, we are going to decompose the tensorial quantities in a gauge invariant component and a gauge variant component.

Consider the metric perturbation,

$$\mathcal{X}^*_\lambda \overline{g}_{\alpha\beta} \equiv \overline{g}_{\alpha\beta} = g_{\alpha\beta} + \lambda h_{\alpha\beta} + \frac{\lambda^2}{2} l_{\alpha\beta} + O(\lambda^3), \tag{10}$$

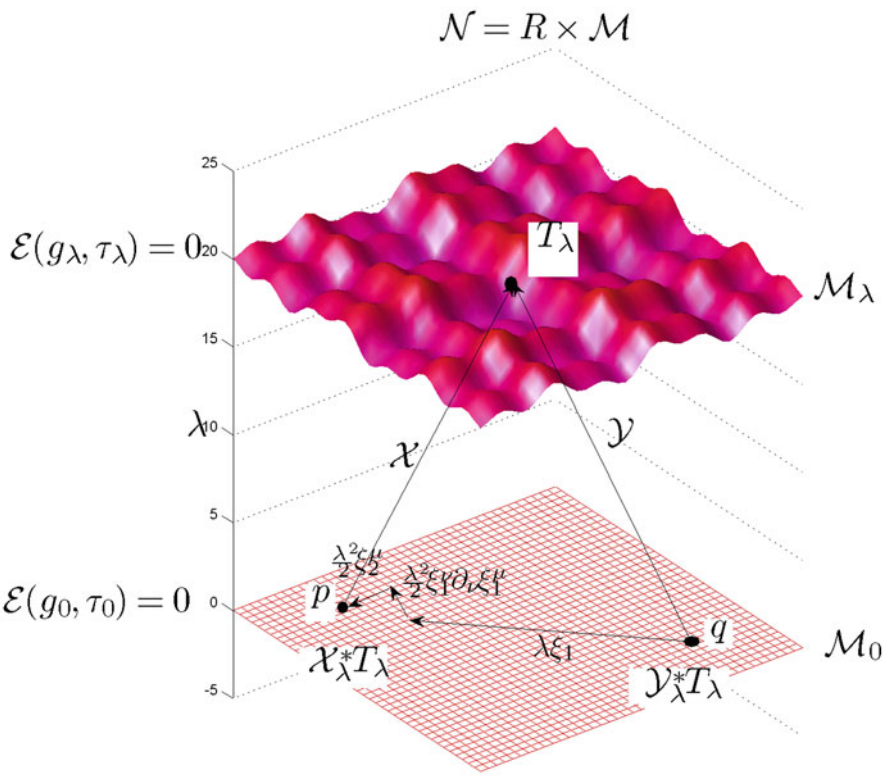


Fig. 2 The second kind of gauge is a point identification between the physical space-time $\mathcal{M}_\lambda$ on the extended manifold $\mathcal{N}$. We assume the existence of a point identification map between $\mathcal{M}_\lambda$ and $\mathcal{M}_0$. However, this point identification is not unique by virtue of the general covariance in the theory [3, 13]

Our starting point to construct gauge invariant variables is the assumption that we already know the procedure for finding gauge invariant variables for the linear metric perturbations. Then a linear metric perturbation h_{ab} is decomposed as [13]:

$$h_{ab} := \mathcal{H}_{ab} + \mathcal{L}_X g_{ab}, \tag{11}$$

where $\mathcal{H}_{ab}$ and $\mathcal{L}_X g_{ab}$ are the gauge invariant and variant parts, respectively.

4 Perturbations of the Field Equations in $f(R)$ Gravity

Consider the equation:

$$\mathcal{E}'(g,\tau)=0,$$

similar to Eq. (5), but in this case $\mathcal{E}'$ are the $f(R)$-modified gravity field equations. Again suppose an exact solution, g_0, is known. We build a one-parameter family g_λ of exact solutions,

$$\mathcal{E}'(g_\lambda,\tau_\lambda)=0 \tag{12}$$

and we want to know how to change g_λ.

We consider a manifold $\mathcal{N}$ like in Sect. 3 but in this case each geometrical tensor fields must satisfy the modified equation for each λ. Now, being f a scalar function like in Sect. 2. We regard that this scalar function admits a Taylor expansion like (7), thus

$$\mathcal{X}^*_\lambda f_\lambda \equiv \bar{f}=\overset{(0)}{f}+\lambda\overset{(1)}{f}+\frac{\lambda^2}{2}\overset{(2)}{f}+\cdots, \tag{13}$$

$$\mathcal{X}^*_\lambda f'_\lambda \equiv \bar{f}'=\overset{(0)}{f'}+\lambda\overset{(1)}{f'}+\frac{\lambda^2}{2}\overset{(2)}{f'}+\cdots. \tag{14}$$

$$\tag{15}$$

And for the Ricci tensor and the Ricci scalar:

$$\mathcal{X}^*_\lambda (R_{\mu\nu})_\lambda \equiv \bar{R}_{\mu\nu}=\overset{(0)}{R}_{\mu\nu}+\lambda\overset{(1)}{R}_{\mu\nu}+\frac{\lambda^2}{2}\overset{(2)}{R}_{\mu\nu}+\cdots \tag{16}$$

$$\mathcal{X}^*_\lambda R_\lambda \equiv \bar{R}=\overset{(0)}{R}_{\mu\nu}+\lambda\overset{(1)}{R}_{\mu\nu}+\frac{\lambda^2}{2}\overset{(2)}{R}_{\mu\nu}+\cdots \tag{17}$$

$$\tag{18}$$

The covariant derivative operator [13, 23]:

$$(\mathcal{X}^*_\lambda(\nabla_\alpha)_\lambda)\omega_\beta \equiv \overline{\nabla}_\alpha\omega_\beta=\nabla_\alpha\omega_\beta-C^\gamma_{\alpha\beta}\omega_\gamma, \tag{19}$$

where

$$C^\gamma_{\alpha\beta}=\frac{1}{2}\overline{g}^{\gamma\delta}(\nabla_\alpha\overline{g}_{\beta\delta}+\nabla_\beta\overline{g}_{\alpha\delta}-\nabla_\delta\overline{g}_{\alpha\beta}). \tag{20}$$

thus

$$C_{\mu\nu}{}^\delta=\overset{(0)}{C}_{\mu\nu}{}^\delta+\lambda\overset{(1)}{C}_{\mu\nu}{}^\delta+\frac{\lambda^2}{2}\overset{(2)}{C}_{\mu\nu}{}^\delta+\cdots, \tag{21}$$

and for the $f(R)$ field equation we have

$$\bar{\Sigma}_{\mu\nu}=\overset{(0)}{\Sigma}_{\mu\nu}+\lambda\overset{(1)}{\Sigma}_{\mu\nu}+\frac{\lambda^2}{2}\overset{(2)}{\Sigma}_{\mu\nu}+\cdots. \tag{22}$$

One can see that replacing Eqs. (13)–(21) in $f(R)$ field Eqs. (4) and comparing with (22) we have,

$$\begin{aligned}\overset{(1)}{\Sigma}_{\mu\nu} &= \overset{(1)}{G}_{\mu\nu} + \overset{(1)}{f'}\, R_{\mu\nu} + (f'-1)\overset{(1)}{R}_{\mu\nu} - \frac{1}{2}[(\overset{(1)}{f} - \overset{(1)}{R})g_{\mu\nu} \\ &+ (f-R)h_{\mu\nu}] - \nabla_\mu\nabla_\nu \overset{(1)}{f'} + \overset{(1)}{C}{}^{\alpha}_{\mu\nu}\nabla_\alpha f' + g_{\mu\nu}\Box \overset{(1)}{f'} \\ &- g_{\mu\nu}g^{\alpha\beta}\overset{(1)}{C}{}^{\delta}_{\alpha\beta}\nabla_\delta f' - g_{\mu\nu}h^{\alpha\beta}\nabla_\alpha\nabla_\beta f' + h_{\mu\nu}\Box f'.\end{aligned}$$

Now, if we use (11), we can decompose the first-order perturbation of the Ricci tensor and the Ricci scalar by [13]

$$\overset{(1)}{R}_{ab} = \overset{(1)}{\mathcal{R}}_{ab} + \mathcal{L}_X \overset{(0)}{R}_{ab}, \qquad \overset{(1)}{R} = \overset{(1)}{\mathcal{R}} + \mathcal{L}_X \overset{(0)}{R}, \tag{23}$$

and we assume that we can decompose f, f' $\overset{(1)}{G}_{ab}$, and $\overset{(1)}{\Sigma}_{ab}$ in a part gauge invariant and gauge variant, i.e.,

$$\begin{aligned}\overset{(1)}{f} &= \overset{(1)}{\mathcal{F}} + \mathcal{L}_X \overset{(0)}{f}, \qquad \overset{(1)}{f'} = \overset{(1)}{\mathcal{F}'} + \mathcal{L}_X \overset{(0)}{f'}, \\ \overset{(1)}{G}_{ab} &= \overset{(1)}{\mathcal{G}}_{ab} + \mathcal{L}_X \overset{(0)}{G}_{ab}, \qquad \overset{(1)}{\Sigma}_{ab} = \overset{(1)}{\mathcal{S}}_{ab} + \mathcal{L}_X \overset{(0)}{\Sigma}_{ab}\end{aligned}$$

and after some calculations we get [12]:

$$\begin{aligned}\overset{(1)}{\Sigma}_{\mu\nu} &= \overset{(1)}{\mathcal{G}}_{\mu\nu} + \overset{(1)}{\mathcal{F}'}\, R_{\mu\nu} + (f'-1)\overset{(1)}{\mathcal{R}}_{\mu\nu} - \frac{1}{2}[(\overset{(1)}{\mathcal{F}} - \overset{(1)}{\mathcal{R}})g_{\mu\nu} \\ &+ (f-R)\mathcal{H}_{\mu\nu}] - \nabla_\mu\nabla_\nu \overset{(1)}{\mathcal{F}'} + H^{\alpha}_{\mu\nu}[\mathcal{H}]\nabla_\alpha f' \\ &+ g_{\mu\nu}\Box \overset{(1)}{\mathcal{F}'} - g_{\mu\nu}g^{\alpha\beta}H^{\delta}_{\alpha\beta}[\mathcal{H}]\nabla_\delta f' \\ &- g_{\mu\nu}\mathcal{H}^{\alpha\beta}\nabla_\alpha\nabla_\beta f' + \mathcal{H}_{\mu\nu}\Box f' + \mathcal{L}_X \overset{(0)}{\Sigma}_{\mu\nu}.\end{aligned} \tag{24}$$

where $H^{\alpha}_{\mu\nu}[g] = \overset{(0)}{C}{}^{\alpha}_{\mu\nu}$. This is a general result. We can consider any background space-time with the condition that we can perform the gauge invariant and gauge variant decomposition. This is one of the most important results of this chapter and it will be used for cosmology in the next section.

5 Cosmological Background Space-Time and Field Equations

The background space-time $\mathcal{M}_0$ considered in CPT is a homogeneous and isotropic universe. The space-time metric of this universe is given by

$$\overset{(0)}{g}_{ab} = a^2(\eta)(-(d\eta)_a(d\eta)_b + \gamma_{ij}(dx^i)_a(dx^j)_b), \tag{25}$$

with η the conformal time.

6 Equations for the First-Order Cosmological Perturbations

6.1 Gauge Invariant Metric Perturbation

If we consider a 3+1 decomposition, for the linear-order metric perturbation, we have

$$h_{ab} = h_{\eta\eta}(d\eta)_a(d\eta)_b + 2h_{\eta i}(d\eta)_{(a}(dx^i)_{b)} + h_{ij}(dx^i)_a(dx^j)_b. \tag{26}$$

Also, considering the decomposition in scalar–vector–tensor of the linear-order metric perturbation h_{ab}, thus we decompose $h_{\eta i}$ and h_{ij} [22]:

$$\begin{aligned}
h_{\eta i} &= D_i h_{(VL)} + h_{(V)i}, \quad D^i h_{(V)i} = 0 \\
h_{ij} &= a^2(h_{(L)}\gamma_{ij} + h_{(T)ij}), \quad h^i_{(T)_i} \equiv \gamma^{ij}h_{(T)ij} = 0 \\
h_{(T)ij} &= (D_i D_j - \frac{1}{3}\gamma_{ij}\Delta)h_{(TL)} + 2D_{(i}h_{(TV)j)} + h_{(TT)ij}, \\
D^i h_{(TV)i} &= 0, \quad D^i h_{(TT)ij} = 0.
\end{aligned}$$

Now, subtracting gauge variant part $\mathcal{L}_X g_{ab}$ from h_{ab}, we have the gauge variant part $\mathcal{H}_{ab}$ in Eq. (11):

$$\begin{aligned}
\mathcal{H}_{ab} = a^2\Big[&-2\overset{(1)}{\Phi}(d\eta)_a(d\eta)_b + 2\overset{(1)}{\nu}_i(d\eta)_{(a}(dx^i)_{b)} \\
&+ (-2\overset{(1)}{\Psi}\gamma_{ij} + \overset{(1)}{\chi}_{ij})(dx^i)_a(dx^j)_b\Big],
\end{aligned} \tag{27}$$

with the following properties $D^i\overset{(1)}{\nu}_i := \gamma^{ij}D_i\overset{(1)}{\nu}_j = \overset{(1)}{\chi}{}^i{}_i := \gamma^{ij}\overset{(1)}{\chi}_{ij} = D^i\overset{(1)}{\chi}_{ij} = 0$ [11]. The quantities (27) are defined by

$$\begin{aligned}
\mathcal{H}_{\eta\eta} &\equiv -2a^2\overset{(1)}{\Phi} = h_{\eta\eta} - 2(\partial_\eta - H)\bar{X}_\eta, \\
\mathcal{H}_{\eta i} &\equiv a^2\overset{(1)}{\nu}_i = h_{(V)i} - a^2\partial_\eta h_{(TV)i},
\end{aligned}$$

$$\mathcal{H}_{ij} \equiv -2a^2 \overset{(1)}{\Psi} \gamma_{ij} + a^2 \overset{(1)}{\chi}_{ij} = a^2(h_{(L)} - \frac{1}{3}\Delta h_{(TL)}) + 2H\bar{X}_\eta + a^2 h_{(TT)ij},$$

with $H = \partial_\eta a/a$, and

$$\bar{X}_\eta \equiv h_{(VL)} - \frac{1}{2}(\partial_\eta - 2H)(a^2 h_{(TL)}) = h_{(VL)} - \frac{1}{2}a^2 \partial_\eta h_{(TL)}.$$

6.2 First-Order $f(R)$ Field Equations

Now, we derive the linear-order $f(R)$ gravity Eqs. (4), where we consider the invariant part of the perturbed metric tensor (27) and the result (24). Also, we will consider the decomposition of the first-order perturbed field equations for $f(R)$ in the gauge invariant part and gauge variant part:

$$\overset{(1)}{\Sigma}_{ab} = \overset{(1)}{\mathcal{S}}_{ab} + \mathcal{L}_X \overset{(0)}{\Sigma}_{ab}. \tag{28}$$

Finally, taking the scalar part equations. Thus, the $\eta - \eta$ component:

$$-a^2 \overset{(1)}{\mathcal{S}}{}^{\eta}_{\eta} = f'[(6\dot{H} + 3H\partial_\eta + \Delta)\overset{(1)}{\Phi} + (3H\partial_\eta + 3\partial_\eta^2)\overset{(1)}{\Psi}] + \frac{a^2}{2}\overset{(1)}{\mathcal{F}} - (3\partial_\eta H + \Delta - 3H\partial_\eta)\overset{(1)}{\mathcal{F}'} - (6H\overset{(1)}{\Phi} + 3\partial_\eta \overset{(1)}{\Psi})\partial_\eta f', \tag{29}$$

here $\Delta \equiv D^i D_i$. The $i - \eta$ component:

$$a^2 \overset{(1)}{\mathcal{S}}{}^{\eta}_{i} = -f'(2HD_i\overset{(1)}{\Phi} + 2\partial_\eta D_i \overset{(1)}{\Psi}) + (\partial_\eta D_i - HD_i)\overset{(1)}{\mathcal{F}'} - (D_i \overset{(1)}{\Phi})\partial_\eta f', \tag{30}$$

the $i - i$ component:

$$\begin{aligned}\frac{a^2}{3}\overset{(1)}{\mathcal{S}}{}_i^i &= f'[(-4H^2 - 2\partial_\eta H - H\partial_\eta - \frac{1}{3}\Delta)\overset{(1)}{\Phi} \\ &+ (4K - 5H\partial_\eta - \partial_\eta^2 + \frac{4}{3}\Delta)\overset{(1)}{\Psi}] \\ &- \frac{a^2}{2}\overset{(1)}{\mathcal{F}} + (2H^2 + \partial_\eta H + 2K - \partial_\eta^2 + \frac{2}{3}\Delta - H\partial_\eta)\overset{(1)}{\mathcal{F}'} \\ &+ [(\partial_\eta + 2H)\overset{(1)}{\Phi} + 2(\partial_\eta - H)\overset{(1)}{\Psi} + 2\overset{(1)}{\Phi}\partial_\eta]\partial_\eta f',\end{aligned} \tag{31}$$

and the $i - j, i \neq j$ component:

$$a^2\overset{(1)}{\mathcal{S}}{}_i^j = D_i D^j[f'(\overset{(1)}{\Psi} - \overset{(1)}{\Phi}) - \overset{(1)}{\mathcal{F}'}]. \tag{32}$$

7 Conclusions and Future Work

In this work, we use the Taylor expansion on a manifold to obtain the first-order perturbed field equations in $f(R)$ gravity in the general case. We regard the first-order function $f(R)$ can be decomposed in a gauge invariant part and gauge variant part exist. We use such decomposition to obtain the first-order $f(R)$ gravity in the general case. This is one of the most important results of this chapter (24). Then we apply it in the cosmological scenario (29)–(32). Our results agree with refs. [2, 5, 16] where the particular case of Newtonian gauge is chosen (29), (30), and (32). We stress that the Nakamura formalism [13] is in principle naturally extended to $f(R)$ gravity. We found the scalar perturbation equations in the cosmological case and in our next chapter we will show the scalar, vector, and tensor perturbations equations in $f(R)$ gravity for an universe filled of a perfect fluid.

References

1. Astier, P. et al.: The Supernova Legacy Survey: measurement of Ω_M, Ω_Λ and w from the first year data set. Astron. Astrophys. **447**, 31 (2006)
2. Bean, R., Bernat, D., Pogosian, L., Silvestri, A., Trodden, M.: Dynamics of Linear Perturbations in f(R) Gravity. Phys. Rev. D **75**, 064020 (2007). astro-ph/0611321v2
3. Bruni, M., Materrese, S., Mollerach, S., Sonego, S.: Perturbations of spacetime: gauge transformations and gauge invariance at second order and beyond. Class. Quantum Grav. **14**, 2585–2606 (1997)
4. Clifton, T., Ferreira, P.G., Padilla, A., Skordis, C.. Modified gravity and cosmology. Phys. Rep. **513**(1–3), 1–189 (2012)
5. de la Cruz-Dombriz, A., Dobado, A., Maroto, A.L.: On the evolution of density perturbations in f(R) theories of gravity. Phys. Rev. D **77**, 123515 (2008). arXiv:0802.2999v2
6. Daly, R.A., Djorgovski, S.G.: A Model-Independent Determination of the Expansion and Acceleration Rates of the Universe as a Function of Redshift and Constraints on Dark Energy. Astrophys. J. **597**, 9 (2003)
7. De Felice, A., Tsujikawa, S.: $f(R)$ theories. Living Rev. Relativ. **13**, 3 (2010)
8. Durrer, R.: The cosmic microwave background. Cambridge University Press Date Published: September 2008 isbn: 9780521847049. http://www.cambridge.org/9780521847049.
9. Guarnizo, A., Castañeda, L., Tejeiro, J.M.: Boundary term in metric $f(R)$ gravity: field equations in the metric formalism. Gen. Rel. Grav. **42**, 2713–2728 (2010)
10. Hortua, J., Castañeda, L., Tejeiro, J.M.: Evolution of magnetic fields through cosmological perturbation theory. Phys. Rev. D **87**, 103531 (2013)
11. Kodama, H., Sasaki, M.: Cosmological perturbation theory. Prog. Theor. Phys. Suppl. **78**, 141–142 (1984)
12. Molano, D.: Teoría de perturbaciones cosmológicas en teorías de gravedad modificada $f(R)$. M.Sc. thesis, Universidad Nacional de Colombia, Observatorio Astronómico Nacional (to be submitted)
13. Nakamura, K.: Second-order gauge-invariant cosmological perturbation heory: current status. Adv. Astr. **2010**, ID 576273 1–26 (2010)
14. Padmanabhan, T.: Equipartition energy, Noether energy and boundary term in gravitational action. Gen. Rel. Grav. **44**, 2681 (2012)
15. Perlmutter, S. et al. [Supernova Cosmology Project Collaboration]: Measurements of Omega and Lambda from 42 high redshift supernovae. Astrophys. J. **517**, 565 (1999)
16. Pogosian, L., Silvestri, A.: The pattern of growth in viable f(R) cosmologies. Phys. Rev. D **77**, 023503 (2008). arXiv:0709.0296v3
17. Riess, A.G. et al. [Supernova Search Team Collaboration]: Observational evidence from supernovae for an accelerating universe and a cosmological constant. Astron. J. **116**, 1009 (1998)
18. Riess, A.G., et al.: Type Ia Supernova Discoveries at $z > 1$ from the Hubble Space Telescope: Evidence for Past Deceleration and Constraints on Dark Energy Evolution. Astrophys. J. **607**, 665 (2004)
19. Schouten, J.A.: Ricci Calculus. Springer, Berlin (1954)
20. Sopuerta, C., Bruni, M., Gualtieri, L.: Non-linear N-parameter spacetime perturbations: Gauge transformations. Phys. Rev. D **70**, 064002 (2004)
21. Stewart, J.M., Walker, M.: Perturbations of space-times in general relativity. Proc. R. Soc. Lond. A **341**, 49–74 (1974)
22. Stewart, J.M.: Perturbations of Friedmann-Robertson-Walker cosmological models. Class. Quantum Grav **7**. 1169–1180 (1990)
23. Wald, R.M.: General Relativity. The University of Chicago Press, Chicago (1984)

Liouvillian Propagators and Degenerate Parametric Amplification with Time-Dependent Pump Amplitude and Phase

Primitivo B. Acosta-Humánez and Erwin Suazo

Abstract This chapter is complementary to previous work of the authors, see (Acosta-Humánez et al., http://arxiv.org/abs/1311.2479, 2013; J. Phys. A Math. Theor. 46(45):455203–455219, 2013). We present in detail missed computations using differential Galois theory dealing with the construction of one-dimensional propagators for the degenerate parametric amplification with time-dependent pump amplitude and phase $\varphi = 0$ and $\varphi = \pi/2$. Also presented is a generalization of Liouvillian propagators for the n-dimensional case, which concerns to the study of explicit solutions for the Cauchy problem of the Schrödinger equation in $\mathbb{R}^d$:

$$\mathrm{i}\frac{\partial\psi}{\partial t} = -\frac{1}{2}\Delta\psi + \sum_{j=1}^{d}\frac{b_j(t)}{2}x_j^2\psi - f_j(t)x_j\psi + \mathrm{i}g_j(t)\frac{\partial\psi}{\partial x_j} - \mathrm{i}\frac{c_j(t)}{2}\left(2x_j\frac{\partial\psi}{\partial x_j} + \psi\right)$$

using differential Galois theory.

Keywords Cauchy initial value problem · Degenerate harmonic oscillator · Differential Galois theory · Mehler's formula · Linear Schrödinger equation · Liouvillian propagator

1 Introduction

In this work, we assemble some results on Liouvillian propagators [1] and generalized Huygens–Fresnel integral in several dimensions with the purpose of unifying the

P. B. Acosta-Humánez (✉)
Department of Mathematics, Universidad del Atlántico and Intelectual. Co, Barranquilla, Colombia
e-mail: primi@intelectual.co

E. Suazo
Department of Mathematical Sciences, University of Puerto Rico, Mayaguez, Puerto Rico, USA

School of Mathematics and Statistics, Arizona State University, AZ, USA
e-mail: erwin.suazo@asu.edu

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_21

Table 1 Some quadratic Hamiltonians having Liouvillian propagators (we assume E, k constants)

Hamiltonian $H(t)$	Fundamental solution (propagator)
Free particle $H_0(t)\psi = -\frac{1}{2}\frac{\partial^2\psi}{\partial x^2}$	$G_0(x,y,t) = \frac{1}{\sqrt{2\pi i t}} e^{i\lvert x-y\rvert^2/2t}$
Constant electric field $H_1(t)\psi = -\frac{1}{2}\frac{\partial^2\psi}{\partial x^2} + E\cdot x\psi$	$G_1(x,y,t) = \frac{1}{\sqrt{2\pi i \sin t}} \exp\left(\frac{i(x-y)^2}{2t}\right)$ $\times \exp\left(\frac{iE(x+y)}{2}t - \frac{iE^2}{24}t^3\right)$
Isotropic oscillator $H_2(t)\psi = -\frac{1}{2}\frac{\partial^2\psi}{\partial x^2} + \frac{1}{2}x^2\psi$	$G_2(x,y,t) = \frac{1}{\sqrt{2\pi i \sin t}}$ $\times \exp\left(i\frac{1}{4\sin(t)}\left(\left(x^2+y^2\right)\cos t - 2xy\right)\right)$
Repulsive harmonic potential $H_3(t)\psi = -\frac{1}{2}\frac{\partial^2\psi}{\partial x^2} - \frac{1}{2}x^2\psi$	$G_3(x,y,t) = \frac{1}{\sqrt{2\pi i \sinh t}}$ $\times \exp\left(i\frac{1}{4\sinh t}\left(\left(x^2+y^2\right)\cosh t - 2xy\right)\right)$
Anisotropic oscillator $H_4(t)\psi = -\frac{1}{2}\frac{\partial^2\psi}{\partial x^2} + \frac{1}{2}\omega^2x^2\psi$	$G_4(x,y,t) = \frac{\omega}{\sqrt{2\pi i \sin \omega t}}$ $\times \exp\left(i\frac{\omega}{4\sin(\omega t)}\left(\left(x^2+y^2\right)\cos \omega t - 2xy\right)\right)$
Parametric oscillator $H_6(t)\psi = -\cos^2 t\frac{\partial^2\psi}{\partial x^2} + \sin^2 t x^2\psi$ $-i\frac{\sin 2t}{2}\left(2x\frac{\partial}{\partial x} - 1\right)\psi$	$G_6(x,y,t) = \frac{1}{\sqrt{2\pi i(\cos t \sinh t + \sin t \cosh t)}}$ $\times \exp\left(\frac{\left(x^2-y^2\right)\sin t \sinh t + 2xy - \left(x^2+y^2\right)\cos t \cosh t}{2i(\cos t \sinh t + \sin t \cosh t)}\right)$
Damped harmonic oscillator $H_7(t)\psi = -\frac{\omega_0}{2}\frac{\partial^2\psi}{\partial x^2} + \frac{\omega_0}{2}x^2\psi$ $+i\frac{\lambda}{2}\left(2x\frac{\partial}{\partial x} + 1\right)\psi$	$G_7(x,y,t) = \sqrt{\frac{\omega}{2\pi i \omega_0 \sin \omega t}}$ $\times \exp\left(\frac{i\omega}{2\,\omega_0 \sin \omega t}\left(\left(x^2+y^2\right)\cos \omega t - 2xy\right)\right)$ $\times \exp\left(\frac{i\lambda}{2\,\omega_0}\left(x^2-y^2\right)\right), \omega = \sqrt{\omega_0^2 - \lambda^2} > 0$
Analog of heat equation with Linear drift $H_8(t)\psi = -\frac{\partial^2\psi}{\partial x^2} - ikx\frac{\partial\psi}{\partial x}, \quad k > 0$	$G_8(x,y,t) = \frac{\sqrt{k}e^{kt/2}}{\sqrt{2\pi i \sinh(kt)}} \exp\left(\frac{ike^{kt}\left[e^{-kt}x - e^{kt}y\right]^2}{4\sinh(kt)}\right)$

results of the authors with the idea of creating a criterium for Galoisian integrability in several dimensions for partial differential equations. In this note, we first show the power of this combination dealing with the construction of one-dimensional propagators for the degenerate parametric amplification with time-dependent pump amplitude and phase $\varphi = 0$ and $\varphi = \pi/2$, we present in detail missed computations in [3] using differential Galois theory, see Sect. 3. We also present the bases, see Proposition 1, to establish the integrability of Liouvillian propagators in several dimensions extending the results presented in [1].

2 Quadratic Hamiltonians in Quantum Mechanics

In [1], Liouvillian propagators (fundamental explicit solutions) of the linear Schrödinger equation (LSE) were studied:

$$i\partial_t\psi = H\psi, \quad H = a(t)p^2 + b(t)x^2 + d(t)(px + xp), \quad p = -i\partial_x. \tag{1}$$

Liouvillian propagators are obtained through Liouvillian functions giving a Galoisian formulation for this kind of integrability: LSE (1) is integrable in Galoisian sense (Galois integrable), when it has a Liouvillian propagator. We construct models of propagators of the form:

$$G(x,y,t)=\frac{1}{\sqrt{2\pi \mathrm{i}\mu(t)}}\,\mathrm{e}^{\mathrm{i}\left(\alpha(t)x^2+\beta(t)xy+\gamma(t)y^2+\delta(t)x+\varepsilon(t)y+\kappa(t)\right)}, \tag{2}$$

for time-dependent Schrödinger equations inspired by solvable Riccati equations:

$$\frac{\mathrm{d}\alpha}{\mathrm{d}t}+b(t)+2c(t)\alpha+4a(t)\alpha^2=0. \tag{3}$$

In this work, we will consider a Hamiltonian more general [14]:

$$H(t)=a(t)p^2+\frac{d(t)}{2}(p\cdot x+x\cdot p)+\frac{b(t)}{2}x^2-g(t)p-f(t)x+\zeta(t). \tag{4}$$

and in several dimensions too (see 36). The expert would recognize (4) as a quantum mechanical self-adjoint Hamiltonian, which is a quadratic polynomial in x and $p=-\mathrm{i}\partial/\partial x$ with time-dependent coefficients. As pointed out before [14], one can assume $\zeta(t)=0$ since it causes a trivial phase factor in the propagator. We have also assumed in (4) the coefficient of the Laplacian constant, $a(t)$, because we can disregard it after substitution. Table 1 reviews the most popular one-dimensional quadratic Hamiltonians with Liouvillian propagators that can be found by the method presented in this note. However, important examples where $a(t)$ is not constant are covered by our approach. In this section we review briefly two of them and we leave degenerate parametric harmonic oscillators for Sects. 3 and 4 to be solved in detail using differential Galois theory.

Caldirola–Kanai Hamiltonian. The Caldirola–Kanai Hamiltonian [4, 16] was introduced more than 60 years ago:

$$H_{\mathrm{CK}}(t)\psi=-\frac{\omega_0\mathrm{e}^{-\lambda t}}{2}\frac{\partial^2\psi}{\partial x^2}+\frac{\omega_0\mathrm{e}^{\lambda t}}{2}x^2\psi. \qquad \text{(Caldirola–Kanai Hamiltonian)}$$

In [7], the authors did not know about this case and called it "third model"; it was one of the models of the damped harmonic oscillator with an explicit propagator considered in that publication. The fundamental solution for the Caldirola–Kanai's Hamiltonian is given by

$$G_{CK}(x,y,t)=\sqrt{\frac{\omega\mathrm{e}^{\lambda t}}{2\pi \mathrm{i}\omega_0\sin\omega t}}\,\mathrm{e}^{\mathrm{i}(\alpha(t)x^2+\beta(t)xy+\gamma(t)y^2)}, \tag{5}$$

where

$$\alpha(t)=\frac{\omega\cos\omega t-\lambda\sin\omega t}{2\,\omega_0\sin\omega t}\mathrm{e}^{2\lambda t}, \qquad \beta(t)=-\frac{\omega}{\omega_0\sin\omega t}\mathrm{e}^{\lambda t}, \tag{6}$$

$$\gamma(t)=\frac{\omega\cos\omega t+\lambda\sin\omega t}{2\,\omega_0\sin\omega t}. \tag{7}$$

Oscillator with $a(t)$ NonConstant Another example with $a(t)$ nonconstant that can be solved using the same ideas presented in this work and has been studied before is

$$H_{10}(t)\psi = -a(t)\frac{\partial^2\psi}{\partial x^2} + \frac{b(t)}{2}x^2\psi, \tag{8}$$

$$a(t) = \frac{\left(\Omega^2\cos(\Omega t) - \gamma\sin(\Omega t)\tanh(\gamma t)\right)}{\cosh(\gamma t)(\cos(\gamma t)\cosh(\gamma t) - 2\gamma)}, \qquad b(t) = -\frac{\omega^2}{4a(t)}, \tag{9}$$

$$\Omega = \sqrt{\omega^2 - \gamma^2}, \tag{10}$$

and its fundamental solution is given by

$$G_{10}(x, y, t) = \sqrt{\frac{m_0\Omega\cosh\gamma t}{2\pi\mathrm{i}\sin(\Omega t)}}\mathrm{e}^{\mathrm{i}(\alpha(t)x^2+\beta(t)xy+\gamma(t)y^2)}, \tag{11}$$

with

$$\alpha(t) = \frac{\cosh(\gamma t)\,(m_0\Omega\cosh(\gamma t)\cos(\Omega t) - \gamma)}{2\sin(\Omega t)}, \qquad \beta(t) = -\frac{m_0\Omega\cosh\gamma t}{2\pi i\sin(\Omega t)}, \tag{12}$$

$$\gamma(t) = \frac{m_0\Omega\cos(\Omega t)}{2\sin(\Omega t)}. \tag{13}$$

Suslov et al. in [8] studied the quantum integrals of motion for these last two examples between other models.

3 Differential Galois Theory and a Generalized Ince's Equation

Following [1], we define Liouvillian propagators in terms of differential Galois theory. An effective algorithm to solve second-order linear differential equations with rational coefficients using differential Galois theory is Kovacic algorithm. When the coefficients are not rational functions, this problem can be solved using *Hamiltonian algebrization* procedure developed by the first author and used in [1, 3]. We say that $\tau = \tau(t)$ is a *Hamiltonian change of variable* if $(\tau, \partial_t\tau)$ is a solution curve of the Hamiltonian:

$$H = \frac{p^2}{2} + V(\tau), \quad \partial_t\tau = \partial_p H = p, \quad \partial_t p = -\partial_\tau H = -\partial_\tau V(\tau), \quad V(\tau) \in \mathbb{C}(\tau).$$

We choose to write

$$\alpha = 2H - 2V(\tau) = (\partial_t\tau)^2, \quad \partial_t\tau = \sqrt{\alpha}.$$

Thus, we can transform differential equations:

$$\partial_t^2\mu + p\partial_t\mu + q\mu = 0 \rightsquigarrow \widehat{\partial}_\tau^2\widehat{\mu} + \widehat{p}\widehat{\partial}_\tau\widehat{\mu} + \widehat{q}\widehat{\mu} = 0,$$

where $\widehat{\partial}_\tau = \sqrt{\alpha}\partial_\tau$, $\widehat{\mu} \circ \tau = \mu$, $\widehat{p} \circ \tau = p$, $\widehat{q} \circ \tau = q$. Moreover, the differential equation $\widehat{\partial}_\tau^2\widehat{\mu} + \widehat{p}\widehat{\partial}_\tau\widehat{\mu} + \widehat{q}\widehat{\mu} = 0$ can be explicitly written as:

$$\partial_\tau^2\widehat{\mu} + \left(\frac{1}{2}\partial_\tau(\ln\alpha) + \frac{\widehat{p}}{\sqrt{\alpha}}\right)\partial_\tau\widehat{\mu} + \left(\frac{\widehat{q}}{\alpha}\right)\widehat{\mu} = 0. \tag{14}$$

In case that $\sqrt{\alpha}$, $\widehat{p}$, and $\widehat{q}$ are rational functions in τ, Eq. 14 is the algebraic form of the first one, i.e., the equation $\partial_t^2\mu + p\partial_t\mu + q\mu = 0$ has been algebrized through a Hamiltonian change of variable. This procedure is called *Hamiltonian algebrization*.

It was studied in [1] the solution of the Ince's equation:

$$\partial_t^2\mu + \frac{2\lambda\,\omega\,\sin(2\,\omega\,t)}{\omega + \lambda\,\cos(2\,\omega\,t)}\partial_t\mu + \frac{\omega^3 - 3\,\omega\,\lambda^2 - \left(\omega^2\lambda + \lambda^3\right)\cos(2\,\omega\,t)}{\omega + \lambda\,\cos(2\,\omega\,t)}\mu = 0 \tag{15}$$

by means of a Galoisian approach to LSE, which is summarized in Theorem 1 of the next section.

Recently, in [3], a variation of degenerate parametric oscillator and Ince's equation was studied by a Galoisian approach to study explicit solutions, statistic, means and variances related with squeezed photons. The Ince's equation considered there is given by

$$\partial_t^2\mu + \frac{2\lambda\omega\sin\left(2\,\omega t + \frac{\pi}{2}\right)}{\omega + \lambda\cos\left(2\,\omega t + \frac{\pi}{2}\right)}\partial_t\mu + \frac{\omega^3 - 3\omega\lambda^2 - \left(\omega^2\lambda + \lambda^3\right)\cos\left(2\,\omega t + \frac{\pi}{2}\right)}{\omega + \lambda\cos\left(2\,\omega t + \frac{\pi}{2}\right)}\mu = 0, \tag{16}$$

The procedure to arrive to the solution of the characteristic equation (16) is missing in [3] and as motivation to the readers, we present here such computations using Hamiltonian algebrization and Kovacic algorithm as in [1].

Owing to the trigonometrical relations:

$$\sin\left(2\,\omega t + \frac{\pi}{2}\right) = \cos(2\,\omega t) \quad \text{and} \quad \cos\left(2\,\omega t + \frac{\pi}{2}\right) = -\sin(2\,\omega t),$$

Equation 16 corresponds to

$$\partial_t^2\mu + \frac{2\lambda\,\omega\,\cos(2\,\omega\,t)}{\omega - \lambda\,\sin(2\,\omega\,t)}\partial_t\mu + \frac{\omega^3 - 3\,\omega\,\lambda^2 + \left(\omega^2\lambda + \lambda^3\right)\sin(2\,\omega\,t)}{\omega - \lambda\,\sin(2\,\omega\,t)}\mu = 0 \tag{17}$$

For the Ince's equation (15), using the Hamiltonian algebrization procedure and the Kovacic algorithm and by properties of double angle, we can write Eq. 15 in terms of $\tan(\omega t)$, we can consider the differential field to $K = \mathbb{C}(\tan\omega t)$. After the Hamiltonian change of variable $\tau = \tan\omega t$, we obtain $\alpha = \omega^2(1+\tau^2)^2$, and by the

Hamiltonian algebrization procedure, we get as an algebraic form of Eqs. 16 and 17 to be

$$\partial_\tau^2\widehat{\mu}+\varphi_1(\tau)\partial_\tau\widehat{\mu}+\varphi_0(\tau)\widehat{\mu}=0,\quad \varphi_1(\tau)=\frac{2\,\omega\tau^3-6\lambda\tau^2+2\,\omega\tau+2\lambda}{\left(1+\tau^2\right)\left(\omega\tau^2-2\lambda\tau+\omega\right)},$$
$$\varphi_0(\tau)=\frac{\left(\omega^3-3\omega\lambda^2\right)\tau^2+\left(2\lambda^3+2\omega^2\lambda\right)\tau+\omega^3-3\omega\lambda^2}{\omega^2\left(1+\tau^2\right)^2\left(\omega\tau^2-2\lambda\tau+\omega\right)}.$$

We can eliminate one parameter through the change $\lambda=\kappa\omega$; thus, our algebraic form becomes

$$\begin{aligned}&\partial_\tau^2\widehat{\mu}+\varphi_1(\tau)\partial_\tau\widehat{\mu}+\varphi_0(\tau)\widehat{\mu}=0,\quad \varphi_1(\tau)=\frac{2\tau^3-6\kappa\tau^2+2\tau+2\kappa}{\left(1+\tau^2\right)\left(\tau^2-2\kappa\tau+1\right)},\\&\varphi_0(\tau)=\frac{\left(1-3\kappa^2\right)\tau^2+\left(2\kappa^3+2\kappa\right)\tau+1-3\kappa^2}{\left(1+\tau^2\right)^2\left(\tau^2-2\kappa\tau+1\right)}.\end{aligned}\tag{18}$$

The general solution for Eq. 18 is given by

$$\widehat{\mu}(\tau)=C_1\frac{\tau\mathrm{e}^{-\kappa\arctan\tau}}{\sqrt{1+\tau^2}}+C_2\frac{\mathrm{e}^{\kappa\arctan\tau}}{\sqrt{1+\tau^2}};$$

Recalling that $\tau=\tan\omega t$ and $\kappa=\lambda/\omega$, we get the general solution of the characteristic equation:

$$\mu(t)=C_1\mathrm{e}^{-\lambda t}\sin\omega t+C_2\mathrm{e}^{\lambda t}\cos\omega t.$$

Alternatively, after the Hamiltonian algebrization process over Eq. 17, we use the command `kovacicsols` over Eq. 18 to obtain

$$\left[\frac{\left(\frac{\mathrm{i}+\tau}{-\tau+\mathrm{i}}\right)^{\frac{1}{2}\mathrm{i}\kappa}}{\sqrt{1+\tau^2}},\frac{\left(\frac{-\tau+\mathrm{i}}{\mathrm{i}+\tau}\right)^{\frac{1}{2}\mathrm{i}\kappa}\tau}{\sqrt{1+\tau^2}}\right];$$

therefore, we can write the general solution as:

$$\hat{\mu}=C_1\frac{\left(\frac{\mathrm{i}+\tau}{-\tau+\mathrm{i}}\right)^{\frac{1}{2}\mathrm{i}\kappa}}{\sqrt{1+\tau^2}}+C_2\frac{\left(\frac{-\tau+\mathrm{i}}{\mathrm{i}+\tau}\right)^{\frac{1}{2}\mathrm{i}\kappa}\tau}{\sqrt{1+\tau^2}}.\tag{19}$$

Now, using the relations above, we get that Eq. (19) becomes

$$\hat{\mu}(\tau)=C_1\mathrm{e}^{\kappa\arctan\tau}\cos\left(\arctan\tau\right)+C_2\mathrm{e}^{-\kappa\arctan\tau}\sin\left(\arctan\tau\right).$$

Now, recalling $\tau=\tan\omega t$ and $\kappa=\lambda/\omega$, we obtain

$$\mu(t)=C_1\mathrm{e}^{\lambda t}\cos\omega t+C_2\mathrm{e}^{-\lambda t}\sin\omega t.$$

we can summarize the results of this section as:

Lemma 1 *The fundamental solution of Eq. (16) is given by*

$$\mu(t)=C_1\mathrm{e}^{\lambda t}\cos\omega t+C_2\mathrm{e}^{-\lambda t}\sin\omega t.\tag{20}$$

4 Degenerate Parametric Amplification with Phase $\varphi = 0$ and $\varphi = \pi/2$

In Schrödinger picture, the time evolution of degenerate parametric amplifier is governed by the time-dependent Schrödinger equation for the state vector $\mid \psi(t)\rangle$ [3]:

$$\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}t} \mid \psi(t)\rangle = \hat{H}(t) \mid \psi(t)\rangle. \tag{21}$$

The degenerate parametric amplification with time-dependent amplitude and phase has a corresponding Hamiltonian [see 3, 18] in terms of annihilation and creation operators, $\hat{a} = \sqrt{1/2\,\omega}\left(\omega\hat{q} + \mathrm{i}\hat{p}\right)$, $\hat{a}^{+} = \sqrt{1/2\,\omega}(\omega\hat{q} - \mathrm{i}\hat{p})$ with $\hat{a}\hat{a}^{+} - \hat{a}^{+}\hat{a} = 1$, of the form:

$$H(t) = \frac{\omega}{2}\left(\hat{a}\hat{a}^{+} + \hat{a}^{+}\hat{a}\right) - \frac{\lambda(t)}{2}\left(e^{i(2\omega t+\varphi(t))}\hat{a}^{2} + e^{-i(2\omega t+\varphi(t))}(\hat{a}^{+})^{2}\right). \tag{22}$$

ω is a given mode, $\lambda(t)$ describes the strength of the interaction between the quantized signal of frequency ω and the classical pump of frequency $2\,\omega$, and the pump phase $\varphi(t)$ are functions of time. The Hamiltonian (22) can be written as:

$$\begin{aligned}\hat{H}(t) = {} & \frac{1}{2}\left(1 + \frac{\lambda(t)}{\omega}\cos\left(2\,\omega t + \varphi(t)\right)\right)\hat{p}^{2} \\ & - \frac{\omega^{2}}{2}\left(1 - \frac{\lambda(t)}{\omega}\cos\left(2\,\omega t + \varphi(t)\right)\right)\hat{q}^{2} \\ & + \frac{\lambda(t)}{2}\sin\left(2\,\omega t + \varphi(t)\right)\left(\hat{p}\hat{q} + \hat{q}\,\hat{p}\right)\end{aligned} \tag{23}$$

and the corresponding characteristic equation (classical equation of motion [5–7]) takes the form

$$\begin{aligned}&\partial_{t}^{2}\mu + \frac{\lambda\,\sin\left(2\,\omega\,t + \varphi\right)\left(2\,\omega + \varphi'\right) - \lambda'\cos\left(2\,\omega\,t + \varphi\right)}{\omega + \lambda\cos\left(2\,\omega\,t + \varphi\right)}\partial_{t}\mu \\ &+ \frac{\omega(\omega^{2} - 3\lambda^{2}) - \lambda\varphi' - \lambda(\omega^{2} + \lambda^{2} + \omega\varphi')\cos\left(2\,\omega t + \varphi\right) - \lambda'\omega\sin\left(2\,\omega\,t + \varphi\right)}{\omega + \lambda\cos\left(2\,\omega\,t + \varphi\right)}\mu = 0.\end{aligned} \tag{24}$$

It was part of the first goal of this note to solve the equation above for the case $\varphi = \pi/2$ using differential Galois theory, see Sect. 3. For the case $\varphi = 0$, see Theorem 1 below, we also find the propagator of a Schrödinger equation associated.

Theorem 1 [1]. *The fundamental solution of Ince's characteristic equation (15) is given by*

$$\mu(t) = C_{1}(\sin\omega t + \cos\omega t)e^{\lambda t} + C_{2}(\sin\omega t - \cos\omega t)e^{-\lambda t}. \tag{25}$$

Furthermore, the propagator of its associated Schrödinger equation (21) with Hamiltonian:

$$H(t) = \frac{1}{2m}\left(1 + \frac{\lambda}{\omega}\cos(2\,\omega t)\right)p^2 + \frac{m\omega^2}{2}\left(1 - \frac{\lambda}{\omega}\cos(2\,\omega t)\right)x^2 + \frac{\lambda}{2}\sin(2\,\omega t)(px + xp), \quad p = -\mathrm{i}\partial_x,$$

is Galois integrable and its propagator is given by

$$G(x, y, t) = \frac{1}{\sqrt{2\pi\,\mathrm{i}(\cos\omega t\sinh\lambda t + \sin\omega t\cosh\lambda t)}} \times \exp\left[\frac{(\omega x^2 - y^2)\sin\omega t\sinh\lambda t + 2xy - (\omega x^2 + y^2)\cos\omega t\cosh\lambda t}{2i(\cos\omega t\sinh\lambda t + \sin\omega t\cosh\lambda t)}\right] \tag{26}$$

which is a Liouvillian propagator.

The main tools applied to prove the previous theorem were Kovacic algorithm and Hamiltonian algebrization procedure, see [2]. Details and proofs can be found in [1]. We have also used the following Lemma.

Lemma 1 [5, 6]. *If the second-order differential equation:*

$$\mu'' - \left(\frac{a'}{a} - 2c\right)\mu' + 4ab\mu = 0 \tag{27}$$

has two solutions μ_0 and μ_1 satisfying

$$\mu_0(0) = 0, \qquad \mu_0'(0) = 2a(0) \neq 0 \tag{28}$$

$$\mu_1(0) \neq 0, \qquad \mu_1'(0) = 0, \tag{29}$$

then the explicit propagator for the LSE:

$$\mathrm{i}\frac{\partial\psi}{\partial t} = -a(t)\frac{\partial^2\psi}{\partial x^2} + b(t)x^2\psi - \mathrm{i}c(t)x\frac{\partial\psi}{\partial x} - \mathrm{i}d(t)\psi, \tag{30}$$

is given by (2) with

$$\alpha(t) = \frac{1}{4a(t)}\frac{\mu_0'(t)}{\mu_0(t)} - \frac{c(t)}{2a(t)}, \tag{31}$$

$$\beta(t) = -\frac{1}{\mu_0(t)}\exp\left(-\int_0^t c(\tau)\,\mathrm{d}\tau\right), \tag{32}$$

$$\gamma(t) = \frac{\mu_1(t)}{2\mu_1(0)\mu_0(t)} + \frac{c(0)}{2a(0)}. \tag{33}$$

Furthermore, α is the solution of (3), and (27) can be obtained from (3) by the substituting (31).

We finish this section presenting the propagator for the Schrödinger equation associated to Ince's equation (17)

Theorem 2 *The fundamental solution of Ince's characteristic equation* (17) *given by* (20) *together with Lemma 1 allow us to find the propagator for the Schrödinger equation with Hamiltonian:*

$$H(t) = \frac{1}{2}\left(1 - \frac{\lambda}{\omega}\sin(2\,\omega t)\right)p^2 + \frac{\omega^2}{2}\left(1 + \frac{\lambda}{\omega}\sin(2\,\omega t)\right)x^2 + \frac{\lambda}{2}\cos(2\,\omega t)(px + xp), \quad p = -\mathrm{i}\partial_x$$

is Galois integrable and its propagator is given by

$$G(x, y, t) = \frac{\sqrt{\omega}}{\sqrt{2\pi\,\mathrm{i}\mathrm{e}^{-\lambda t}\sin\omega t}} \times \exp\left[\frac{\omega\cos\omega t}{2}x^2 - \frac{\omega}{\mathrm{e}^{-\lambda t}\sin\omega t}xy + \frac{\omega\cos\omega t}{2\mathrm{e}^{-2\lambda t}}x^2\right], \tag{34}$$

which is a Liouvillian propagator.

5 Generalized Huygens–Fresnel Integrals

For the second goal, we follow the same approach and theoretical framework of [1, and references therein], we consider the LSE in $\mathbb{R}^d$ with time-dependent quadratic Hamiltonian having the following form:

$$\mathrm{i}\frac{\partial\psi}{\partial t} = \sum_{j=1}^{d} H_j(t)\,\psi, \qquad \psi(x,0) = \varphi \tag{35}$$

with[1]

$$\sum_{j=1}^{d} H_j(t)\psi = \sum_{j=1}^{d} -\frac{1}{2}\frac{\partial^2\psi}{\partial x_j^2} + \frac{b_j(t)}{2}x_j^2\psi - f_j(t)x_j\psi + \mathrm{i}g_j(t)\frac{\partial\psi}{\partial x_j} - \mathrm{i}\frac{c_j(t)}{2}\left(2x_j\frac{\partial\psi}{\partial x_j} + \psi\right). \tag{36}$$

[1] Where $b_j, f_j, g_j, c_j \in C^1$ (b_j, f_j, g_j could be piecewise continuous functions) and $\varphi \in S(\mathbb{R}^n)$ ($S(\mathbb{R}^n)$ is the Schwartz space) to simplify the discussion.

We derive an explicit formula for the time evolution operator of (35) and (36) in the form:

$$\psi(x,t) = U_H(t,0)\varphi(x) = \int_{\mathbb{R}^n} G_H(x,y,t)\,\varphi(y)\,\mathrm{d}y. \tag{37}$$

In Lemma 2, for the convenience of the reader, we extend the explicit formula found in [6], see also Lemma 1, for the one-dimensional case to several variables. The study of the best formulation has been of great interest by his general applications on mathematical physics [9–13, 15, 17, and references therein]. Our approach gives the time evolution operator explicitly in terms of the original coefficients. As a consequence, uniqueness of the solution for (35) and (36) and its continuous dependence on the initial data and smoothness of the solution (well-posedness) is obtained. In Lemma 3, we also obtain estimates (48) and (49) that will be used for the study of the well-posedness of the nonlinear case.

In this section, we also present the analysis of the dynamics of the propagators (Corollary 1) and convenient inequalities to establish the well-posedness of the Cauchy initial value problem. A list of popular propagators that can be solved explicitly by lemma is listed in Table 1, they can be combined as a tensor product in order to have propagators of several dimensions.

Formula (47) is a generalization of Mehler's formula and it is a consequence of the following result:

Lemma 2 *(Fundamental solution) 1. The Cauchy initial value problem (35) and (36) has the following fundamental solution:*

$$G_H(x,y,t) = \prod_{j=1}^{n} \frac{1}{2\pi i\mu_j(t)} \mathrm{e}^{\mathrm{i}\left(\sum \alpha_j(t)x_j^2+\beta_j(t)x_jy_j+\gamma_j(t)y_j^2+\delta_j(t)x_j+\varepsilon_j(t)y_j+\kappa_j(t)\right)}. \tag{38}$$

μ_j satisfies

$$\mu_j'' + 4\sigma_j(t)\,\mu_j = 0, \tag{39}$$

with $\sigma_j(t) = b_j(t)/2 - c_j^2(t)/4 - c_j'(t)/4$, which must be solved subject to $\mu_j(0) = 0$, $\mu_j'(0) = 1$. Furthermore, $\alpha_j(t)$, $\beta_j(t)$, $\gamma_j(t)$, $\delta_j(t)$, $\varepsilon_j(t)$, $\kappa_j(t)$ are differentiable in time t only and are given explicitly by

$$\alpha_j(t) = \frac{1}{2}\frac{\mu_j'(t)}{\mu_j(t)} - \frac{c_j(t)}{2}, \tag{40}$$

$$\beta_j(t) = -\frac{1}{\mu_j(t)}, \tag{41}$$

$$\gamma_j(t) = \frac{1}{2\mu_j(t)\,\mu_j'(t)} - 2\int_0^t \frac{\sigma_j(\tau)}{\left(\mu_j'(\tau)\right)^2}\,\mathrm{d}\tau + \frac{c_j(0)}{2} \tag{42}$$

$$\delta_j(t) = \frac{1}{\mu_j(t)} \int_0^t \left(f_j(\tau) - c_j(\tau) g_j(\tau)\right) \mu_j(\tau) + g_j(\tau) \mu_j'(\tau)\, \mathrm{d}\tau, \tag{43}$$

$$\varepsilon_j(t) = -\frac{\delta_j(t)}{\mu_j'(t)} + 4 \int_0^t \frac{\mu_j(\tau)\, \delta_j(\tau)\, \sigma_j(\tau)}{\left(\mu_j'(\tau)\right)^2} \mathrm{d}\tau + \int_0^t \frac{1}{\mu'(\tau)} \left(f_j(\tau) - c_j(\tau) g_j(\tau)\right) \mathrm{d}\tau, \tag{44}$$

$$\kappa_j(t) = \frac{\mu_j(t)}{2\mu_j'(t)} \delta_j^2(t) - 2 \int_0^t \frac{\sigma_j(\tau)}{\left(\mu_j'(\tau)\right)^2} \left(\mu_j(\tau)\, \delta_j(\tau)\right)^2 \mathrm{d}\tau - \int_0^t \frac{\mu_j(\tau)\, \delta_j(\tau)}{\mu_j'(\tau)} \left(f_j(\tau) - c_j(\tau) g_j(\tau)\right) \mathrm{d}\tau \tag{45}$$

with

$$\delta_j(0) = g_j(0), \qquad \varepsilon_j(0) = -\delta_j(0), \qquad \kappa_j(0) = 0. \tag{46}$$

Thus, the fundamental solution (propagator) is explicitly given by (38) in terms of the characteristic function (39) with (40)–(45). The fundamental solution G_H includes several well-known examples see Table 1.

Lemma 3

1. *Let $\varphi \in S(\mathbb{R}^n)$ (Schwartz space), then the Cauchy initial value problem for (35) and (36) has the following unitary evolution operator:*

$$U_H(t)\varphi \equiv \prod_{j=1}^{n} \frac{1}{2\pi \mathrm{i} \mu_j(t)} \int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}\left(\sum_{j=1}^{n} \alpha_j(t) x_j^2 + \beta_j(t) x_j y_j + \gamma_j(t) y_j^2 + \delta_j(t) x_j + \varepsilon_j(t) y_j + \kappa_j(t)\right)} \varphi(y) \mathrm{d}y. \tag{47}$$

2. *If $\varphi \in S(\mathbb{R}^n)$, then $U_H(t)\varphi \in S(\mathbb{R}^n)$.*
   If ψ satisfies (35) and (36) and it is smooth, then:
3. *The following estimates hold:*

$$\|U_H(t)\varphi\|_{L^2(\mathbb{R}^n)} = \|\varphi\|_{L^2(\mathbb{R}^n)}, \tag{48}$$

$$\|U_H(t,s)\varphi\|_{L^\infty(\mathbb{R}^n)} \leq \prod_{j=1}^{d} \frac{1}{\sqrt{4\pi \mathrm{i} \mu_j(t) \mu_j(s) \left(\gamma_j(s) - \gamma_j(t)\right)}} \|\varphi\|_{L^1(\mathbb{R}^n)}. \tag{49}$$

The following properties for $U_H(t,s)$ are fundamental to describe the dynamics of the evolution operator.

Corollary 1 *The evolution operator associated to (35) and (36) satisfies the following properties:*

1. $U_H(t,s) = U_H(t)U_H^{-1}(s)$.
2. $U_H(t,t) = Id$.
3. *The map* $(t,s) \to U_H(t,s)$ *is strongly continuous.*
4. $U_H(t,\tau)U_H(\tau,s) = U_H(t,s)$.

We finish our presentation with results on Galoisian integrabilty.

Theorem 3 (Galoisian approach to LSE, [1]) *LSE (1) is Galois integrable if and only if the differential Galois group of the characteristic equation associated to (1) is virtually*[2] *solvable.*

Using this theorem and previous lemmas in this section, we arrive to our main result.

Proposition 1 *The equation:*

$$\mathrm{i}\frac{\partial\psi}{\partial t} = -\frac{1}{2}\Delta\psi + \sum_{j=1}^{d}\frac{b_j(t)}{2}x_j^2\psi - f_j(t)x_j\psi + \mathrm{i}g_j(t)\frac{\partial\psi}{\partial x_j} - \mathrm{i}\frac{c_j(t)}{2}\left(2x_j\frac{\partial\psi}{\partial x_j} + \psi\right)$$

has a Liouvillian propagator, if and only if, $DGal(L_j \mid K)$ *(associated to* $\mu_j'' + \sigma_j\mu_j = 0$*) is virtually solvable for all* $j \in \{1,...,n\}$.

Proof As we have n decoupled characteristic equations, we obtain a differential Galois group for each one. Therefore, the differential Galois group for the system of characteristic equations has a faithful representation as a subgroup of block matrices. Thus, the differential Galois for the system is virtually solvable if and only if the differential Galois group for each characteristic equation is virtually solvable. By the previous theorem, we have a Liouvillian propagator built through the solutions of the characteristic equations and therefore we obtain a Liouvillian propagator constructed with the system of characteristic equations, i.e., n-dimensional Liouvillian propagator.

Acknowledgement The first author was partially supported by the MICIIN/FEDER grant number MTM2009–06973, the Generalitat de Catalunya grant number 2009SGR859, and DIDI - Universidad del Norte (Raimundo Abello). The second author was partially supported by a grant from the Simons Foundation (#316295 to Erwin Suazo), Arizona State University, and University of Puerto Rico, Mayaguez. The authors thank to Greisy Morillo by their hospitality during the final process of this chapter.

[2] The connected identity component of its differential Galois group is a solvable group.

References

1. Acosta-Humánez, P.B., Suazo, E.: Liouvillian propagators, Riccati equations and differential Galois theory. J. Phys. A Math. Theor. **46**(45), 455203–455219 (2013)
2. Acosta-Humánez, P.B., Morales-Ruiz, J.J., Weil, J.-A.: Galoisian approach to integrability of Schrödinger equation. Rep. Math. Phys. **67**(3), 305–374 (2011)
3. Acosta-Humánez, P.B., Mahalov, A., Kryuchov, S., Suazo, E., Suslov, S.K.: Degenerate parametric amplification of squeezed photons: Explicit solutions, statistics, means and variances. Preprint (2013). http://arxiv.org/abs/1311.2479
4. Caldirola, P.: Forze non conservative nella meccanica quantistica. Nuovo Cim. **18**, 393–400 (1941)
5. Cordero-Soto,, R., Suslov, S.K.: The degenerate parametric oscillator and Ince's equation. J. Phys. A: Math. Theor. **44**(1), 015101(9 pages) (2011)
6. Cordero-Soto, R., Lopez, R.M., Suazo, E., Suslov, S.K.: Propagator of a charged particle with a spin in uniform magnetic and perpendicular electric fields. Lett. Math. Phys. **84**(2–3), 159–178 (2008)
7. Cordero-Soto, R., Suazo, E., Suslov, S.K.: Models of damped oscillators in quantum mechanics. J. Phys. Math. **1**, Article ID S090603, 16 pages (2009)
8. Cordero-Soto, R., Suazo, E., Suslov, S.K.: Quantum integrals of motion for variable quadratic Hamiltonians. Ann. Phys. **325**(9), 1884–1912 (2010)
9. Dodonov, V.V., Malkin, I.A., Man'ko, V.I.: Integrals of the motion, Green functions, and coherent states of dynamical systems. Intern. J. Theor. Phys. **14**, 37–54 (1975)
10. Feynman, R.P., Hibbs, A.R.: Quantum Mechanics and Path Integrals, vol. 4. McGraw-Hill, New York, 1965
11. Fujiwara, D.: A construction of the fundamental solution for the Schrödinger equation. J. Anal. Math. **35**, 41–96 (1979)
12. Fujiwara, D.: Remarks on the convergence of the Feynman path integrals. Duke Math. J. **47**(3), 559–600 (1980)
13. Fujiwara, D.: On a nature of convergence of some path integrals, I. Duke Math. J. **47**, 559–600 (1980)
14. Hagedorn, G.A., Loss, M., Slawny, J.: Non-stochasticity of time-dependent quadratic Hamiltonians and spectra of canonical transformations, J. Phys. A Math. Gen. **19**, 1986, 521–531
15. Hörmander, L.: Symplectic classification of quadratic forms, and general Mehler formulas. Math. Z. **219**(3), 413–449 (1995)
16. Kanai, E.: On the quatization of dissipative systems. Prog. Theor. Phys. **3**, 440–442 (1941)
17. Killip, R., Visan, M., Zhang, X.: Energy-critical NLS with quadratic potential. Comm. PDE. **34**, 1531–1565 (2009)
18. Raiford, M.T.: Degenerate parametric amplification with time-dependent pump amplitude and phase. Phys. Rev. A **9**(5), 2060–2069 (1974)

Construction of Shear Wave Models by Applying Multi-Objective Optimization to Multiple Geophysical Data Sets

Lennox Thompson, Aaron A. Velasco and Vladik Kreinovich

Abstract For this work, our main purpose is to obtain a better understanding of the Earth's tectonic processes in the Texas region, which requires us to analyze the Earth structure. We expand on a constrained optimization approach for a joint inversion least-squares (LSQ) algorithm to characterize a Earth's structure of Texas with the use of multiple geophysical data sets. We employed a joint inversion scheme using multiple geophysical data sets for the sole purpose of obtaining a three-dimensional velocity structure of Texas in order to identify an ancient rift system within Texas. In particular, we use data from the USArray, which is part of the EarthScope experiment, a 15-year program to place a dense network of permanent and portable seismographs across the continental USA. Utilizing the USArray data has provided us with the ability to image the crust and upper mantle structure of Texas. We prove through numerical and experimental testing that our multiobjective optimization problem (MOP) scheme performs inversion in a more robust, and flexible matter than traditional inversion approaches.

Keywords Teleseismic · Receiver functions · Seismographs · Body waves · Multi-objective optimization · Primal-dual interior point method

1 Introduction

For this chapter, we propose to combine multiple geophysical data sets for the purpose of assisting us in better determining physical properties of the Earth structure. By simultaneously inverting multiple data sets, we obtain a better estimate of the

L. Thompson (✉) · A. A. Velasco · V. Kreinovich
Department of Geological Sciences, University of Texas at El Paso (UTEP),
500 W. University Ave, El Paso, TX 79968, USA
e-mail: lethompson@miners.utep.edu

A. A. Velasco
e-mail: aavelasco@utep.edu

V. Kreinovich
e-mail: vladik@utep.edu

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_22

true Earth structure. In general, there are two reasons why the estimated Earth structure model differs from the true Earth structure. The first reason is the inherent nonuniqueness of the inverse problem that causes several (usually infinitely many) models to satisfy the data. The second reason is that real geophysical data are always affected by noise, which introduces error associated with the estimation of the Earth structure model after inversion. By jointly inverting multiple geophysical data sets, we reduce the inherent nonuniqueness typical for the geophysical data sets (e.g., receiver functions, surface wave dispersion, teleseismic delay travel times, and gravity) individually [6, 35]. For this research, we use receiver functions, surface wave dispersion measurements, and *P* wave travel times to characterize the crust and upper mantle structure of the Texas region.

In general, geophysical data sets such as receiver functions are suited to constrain the depth of discontinuities and are sensitive to relative changes in *S* wave velocities in different layers. Surface wave measurements, on the other hand, constrain the absolute shear velocities between discontinuities whereas receiver functions are unable to do that [14, 21, 29, 30, 32]. Seismic first-arrival travel times and gravity data are complementary to each other because one can recover the causative slowness and density distributions of the Earth structure [18]. The complementary information provided by the following data sets reduces the inherent ambiguity or nonuniquess of performing inversion (e.g., [4, 6, 9, 19, 22, 23]). By jointly inverting seismic data along with gravity data, we will be able to overcome the difficulties of nonuniqueness and be able to facilitate the construction of the true Earth model.

When we process a single data set (e.g., surface wave dispersion), we use the least-squares (LSQ) method to find the best-fit model. For multiple data sets (e.g., surface wave dispersion and receiver functions), if we knew the variance (uncertainty of data) of the different measurements of the multiple data sets, we could still use the LSQ approach to find the model space. In practice, we only have an approximate knowledge of the variances. So, instead of producing a single model, we want to generate several models corresponding to different possible variances. Once several models corresponding to different possible variances are computed, we can then proceed to select the most geophysical meaningful model from the Pareto Front [15]. The reason we use an optimization technique is to find the best possible solution for our nonlinear geophysics inverse problem. For example, in geophysics, most inverse problems require finding some minimization and that is why we will use an optimization technique called multi-objective optimization problem (MOP). The MOP technique generates several possible models. This is what sets it apart from other various joint inversion techniques. We will be able to select the final solution from a population of alternative solutions from the model space. Such methods are described in [15, 27, 28].

There are two types of seismic waves that travel through the Earth: body waves and surface waves. Both types of waves give us different sensitivities and information about the Earth structure, since they are sampling the interior and surface of the medium with different velocities and directions. The information collected from the body waves travels deeper into the Earth and translates into teleseismic *P* wave receiver functions. In order to obtain information about the Earth surface, surface

waves are analyzed, in our case, by means of surface waves dispersion. On one hand we have receiver functions, which resolve discontinuities (impedance contrasts) in seismic velocities, and provide good measurement of crustal thickness, without providing a good average of shear wave velocity. On the other hand, we have surface (Love and Rayleigh) waves whose energy is concentrated near the Earth's surface, and provide good average of absolute shear wave velocity, without good shear wave velocity contrasts in layered structures [5, 14, 21, 25, 29, 32]. Therefore, these two data sets can be considered as complimentary and consistent, as long as we sample the same medium. Hence, we expect a mutually consistent estimate of the Earth's structure. Since both data sets are sensitive to shear wave velocity structure [14], we can assume a forward operator F depending nonlinearly on our model parameter $x \in \mathsf{R}^n$ that represents the different shear velocities of a half space with n horizontal layers (a standard way of modeling Earth's structure). In the next subsections, we explain in more detail the nonlinear relationship with respect to shear wave velocities of this operator and the techniques used to compute each synthetic data set.

1.1 Receiver Functions

A receiver function is simply a time series representation of the Earth's response relative to an incoming *P* wave propagating near a recording station. Positive or negative spike amplitudes represent positive or negative seismic velocity contrasts. A receiver function technique can model the structure of the Earth by using seismograms from three component (vertical, north, and east) seismic stations from teleseismic earthquakes. The receiver function technique takes advantage of the fact that part of the energy of seismic *P* waves is converted into *S* waves at discontinuities along the ray path [2, 7], and has been utilized in many studies (e.g., [3, 11, 37, 38]). For data collection and processing, we use the standing order for data (SOD) [3, 26] to request three component seismograms for *P* wave arrivals and for events with a minimum magnitude 5.5, depth in the range of 1–600 km, and an epicentral distance ranging from 30° to 95° (e.g., [3]).

Receiver functions were first applied in the late 1970s at solitary stations to obtain local one-dimensional (1D) structural estimates [16]. Since then, there was an increase in the number of stations deployed for seismic experiments. It is now possible to generate detailed two- (2D) or three-dimensional (3D) images of structures, such as the moho and the upper mantle transition zone discontinuities near 410 and 670 km depths using receiver functions (e.g., [36]).

Receiver functions are derived using deconvolution, a mathematical method used to filter a signal and isolate the superimposed harmonic waves. Specially, receiver functions are calculated by deconvolving the vertical component of a seismogram from the radial component, resulting in the identification of converted phases where there is an impedance contrast (crustal–mantle boundary; Fig. 1) [29].

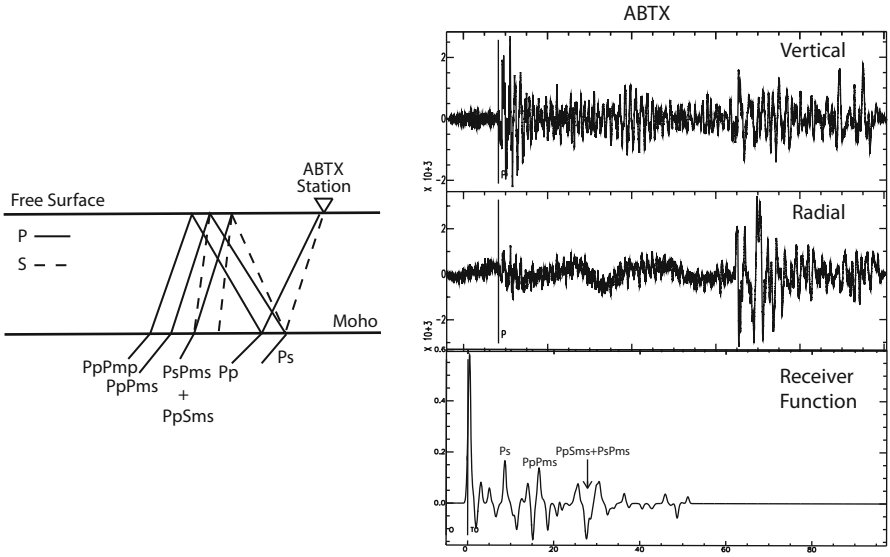


Fig. 1 *Left* Illustration of a simplified ray diagram, which identifies the Ps, converted phases, which comprise the receiver function for a single layer. *Right* Vertical and radial seismograms and the corresponding receiver function resulting from the deconvolution of the vertical component from the radial component

1.2 Receiver Function Stacking

We used the receiver function stacking technique introduced by Zhu and Kanamori [39], which estimates the crustal thickness and a *Vp/Vs* ratio based on the radial receiver function. This technique is the standard approach used by EarthScope Automated Receiver Survey (EARS). Assuming that no lateral velocity heterogeneities exist, the time separation between the P_s converted wave and the direct P wave obtained from receiver functions (t_{Ps}) can then be used to estimate crustal thickness (H), given the average crustal velocities V and a *Vp/Vs* ratio (κ), and the constant ray parameter p of the incident wave (e.g., [8]). The trade-off between the thickness and the crustal velocities presents an ambiguity that can be reduced by using the later multiple phases $tp_p p_s$ and $tp_s p_s + p_p s_s$, which provide additional constraints to both *Vp/Vs* and the crustal thickness (e.g., [8, 39]). Using and stacking multiple events helps to increase the signal-to-noise ratio (SNR), which may be caused by background noise, scattering from crustal heterogeneities, and P-to-S multiple conversions from other velocity discontinuities [20]. The H-κ domain stacking weights each phase and plots the stacked phases as a gridded image $s(H, \kappa)$,which reaches a maximum when all three phases (tp_s, $tp_p p_s$, $tp_s p_s + p_p s_s$) are stacked coherently with the correct H and κ [39]. The main advantage of this grid-search-based technique is that (1) large amounts of receiver functions can be processed without the need of picking P_s arrival times, and (2) the stacking results in an enhancement of

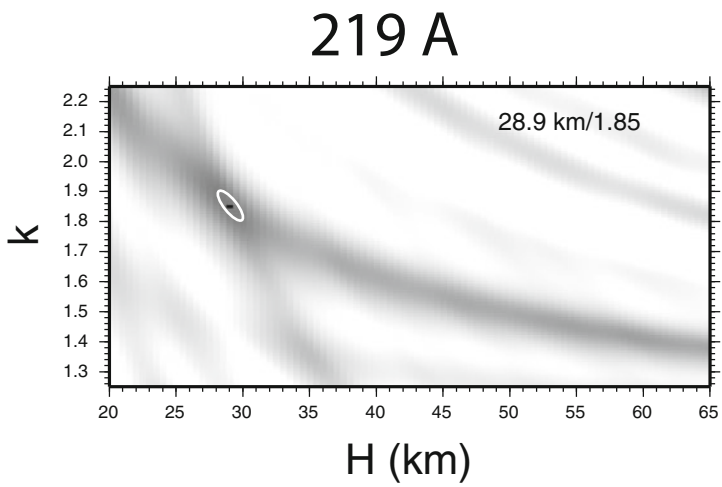


Fig. 2 This is a receiver function stack of station 219A, *Vp*/*Vs* vs. *H* (km). The *black dot* with the *white circle* around it represents the preferred value. Note the multiple *shaded regions* might result in a poor choice of crustal thickness

the signal/noise ratio and a suppression of lateral variations in the vicinity of the recording station [20]. We will use this technique to derive an average crustal model including H and *Vp/Vs* (κ). An example of this technique is shown in Fig. 2 for one of the EarthScope USArray stations, 219A. The dark dot with the white circle around the dot represents the possible solution in H and *Vp/Vs* space (Fig. 2).

1.3 Surface Wave Dispersion

Surface waves in general differ from body waves in many respects—they travel slower, lower frequencies, largest amplitudes, and their velocities are in fact dependent on frequency [29]. The surface wave velocities vary with respect to depth being sampled by each period of the surface wave. The sampling by each period of the surface wave is known as dispersion [31]. Valuable information can be inferred by measuring surface wave dispersion because it will allow you to be able to better understand the Earth's crustal and mantle velocity structure [17, 25, 31]. In particular, Love and Rayleigh wave group dispersion observations generally account for average velocity structure as a function of depth [14, 21]. The dispersion curves for surface waves are extracted from station records of three component seismograms for different frequencies and distances, by using reduction algorithms that rely on spectral analysis techniques. The important fact here is that, based on Rayleigh's principle, surface wave velocities are more sensitive to S wave velocity, although they are also theoretically sensitive to P wave velocity and density. The Rayleigh's principle states that the phase velocity perturbation, denoted by $\frac{\delta c}{c}$, can be viewed as

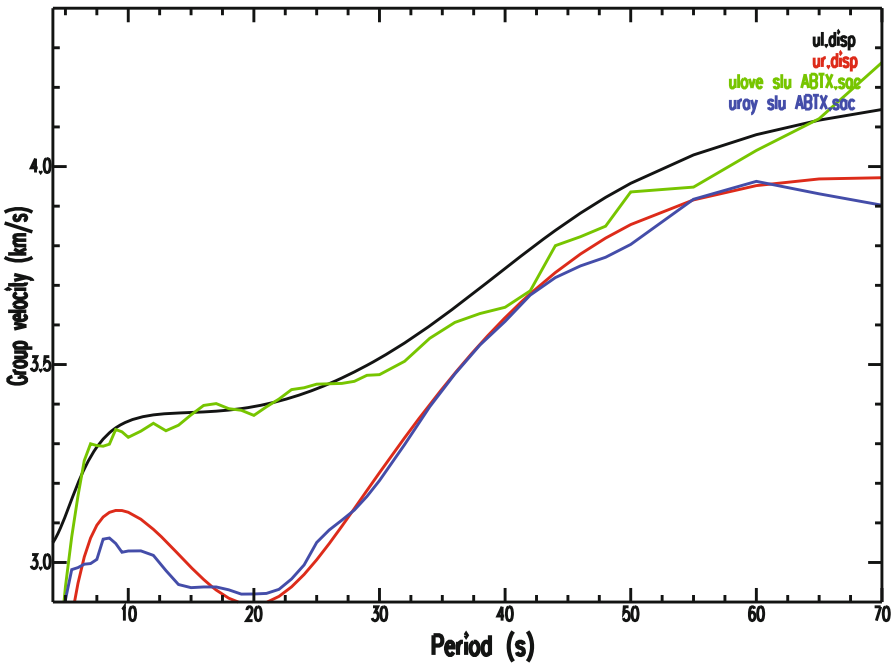


Fig. 3 Surface wave dispersion curves (Love and Rayleigh) for station ABTX using real data

a function of $(K_\alpha, K_\beta, K_\rho)$, the sensitivity coefficients for P wave velocity, S wave velocity, and density, respectively, i.e., (Fig. 3)

$$\frac{\delta c(T)}{c(T)} = \int\left(K_\alpha \frac{\delta\alpha(z)}{\alpha(z)} + K_\beta \frac{\delta\beta(z)}{\beta(z)} + K_\rho \frac{\delta\rho(z)}{\rho(z)}\right), \tag{1}$$

where T is the period and z is the depth. By investigating sensitivity function variation in depth, the relative contribution of each property to dispersion can be shown. This subject is beyond the scope of our work, thus we just mention here that such analysis allows geophysicists to show that the relative contribution of P wave velocity, and density to dispersion is smaller than the one for S wave velocity [14]. That is, surface wave dispersion is much more sensitive with respect to S wave velocity, and therefore we have established the dependence of this data set on shear wave velocity.

1.4 Delay Travel Times

The travel time T between a source and receiver along a ray L is given in integral form for a velocity field as:

$$T = \int_L \frac{\mathrm{d}s}{v(s)} \tag{2}$$

where s is the position vector in 2D or 3D media. Travel times are considered a nonlinear inverse problem given the relationship between the measured data (travel times) and the unknown model parameters (the velocity field). However, by transforming variables to use slowness, the reciprocal of velocity, instead of velocity as the unknown, a seemingly linear inversion problem is created:

$$\int_L \Delta u(s)\mathrm{d}s = \Delta T = T_{\text{obs}} - T_{\text{pred}} \tag{3}$$

However, the ray is also dependent on the velocity (or slowness) model, thus making the inverse problem nonlinear regardless of what form of model variable or parameterization is used. If the medium is subdivided into blocks, the path length l_j in the jth block and can be discretized to

$$\Delta T = \sum_j l_j \Delta u_j \tag{4}$$

The model can be parameterized any number of ways using velocity or slowness, and cells, nodes, or splines, since the problems' nonlinearity must be dealt with regardless of the parameterization. Most often a linearized gradient approach is applied in which a starting model is used and both the model and rays are updated over a series of iterations with the hope that there will be convergence to an acceptable model (the final model). The model is almost always discretized using cells, nodes, or other interpolating functions; in the latter two cases, the discrete model parameters are the coefficients of the interpolating functions. For the formulation of travel times for a tomography problem, the model is parameterized using constant slowness cells, in which case the equation for the ith data becomes

$$\Delta T_i = \sum_j l_{ij} \Delta u_j \tag{5}$$

where l_{ij} is the length of the ith ray in the jth model cell and Δu_j is the slowness in the jth cell. In this case, the path length of each ray in a block, l_{ij} is the partial derivative, $\partial T_i / \partial u_j$ of the travel time with respect to the slowness of that block (Fig. 4) [32].

1.5 Gravity Anomalies

In geophysics, gravity anomalies are generally defined as the difference between observed gravity field and the field of a reference model. Depending on the reference gravity model, two different types of anomaly variations are considered: gravity anomalies and gravity disturbances. The geodetic gravity anomaly is defined as the difference between gravity on the geoid and normal gravity on the reference ellipsoid [13]. On the other hand, the gravity disturbance is defined as the difference of the

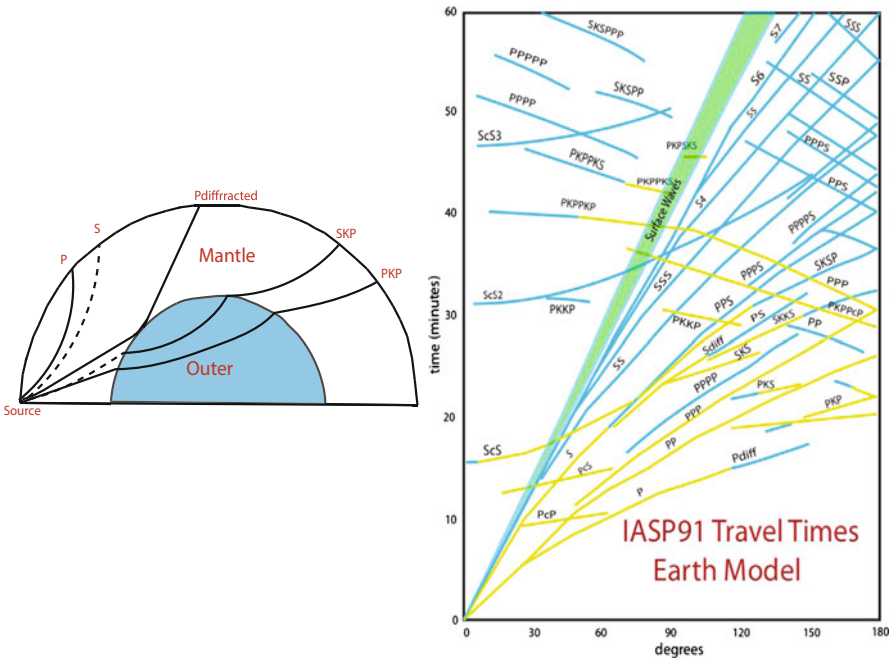


Fig. 4 *Left* Half-space showing different P and S waves created within the Earth. *Right* Different seismic phases within the Earth. Travel times are from the Array Network Facility (ANF) seismic catalog [1]

fields at the same point on the reference ellipisoid. It has been demonstrated that the gravity disturbances are more appropriate for geophysical purposes (e.g., [12]). In any case, its necessary to take into account the difference in the interpretation (Fig. 5).

The observed gravity anomalies reflect the effect of density variations relative to the homogeneous reference model. Interpretation of the gravity anomalies implies an estimation of the density heterogeneities. The density model should reproduce the observed gravity field, taking into account that the observations may be affected by measurement errors. Density heterogeneity of the Earth, associated with thermal and compositional variations or with deflected boundaries separating layers of different density, is one of the main factors that control dynamic processes and deformations at both shallow and deep levels. Therefore, interpretation of the gravity anomalies or gravity modeling is one of the principal methods, which help to understand the nature and origin of the tectonic processes and the Earth's dynamics.

2 Forward Problem

If we know the layered shear velocity distribution $x = (x_1, \ldots, x_n)$ at n different horizontal layers, then we can evaluate the measured quantities $y = (y_1, \ldots, y_m)$ (e.g., the travel times) by applying an appropriate nonlinear operator $F(x)$ that uses

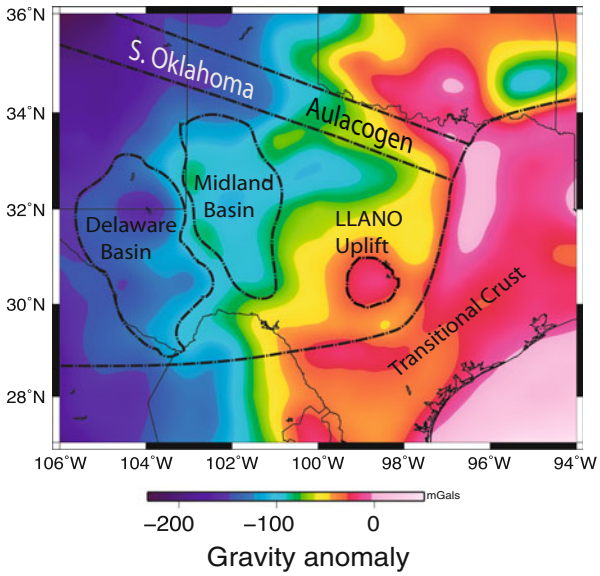


Fig. 5 Bouguer gravity anomaly map of the Texas region and surrounding area. High-amplitude gravity anomaly observed in Texas

the velocities x to predict the Earth's response $y = F(x)$,

$$F(x) = (F_1(x), \ldots, F_m(x)) \in \mathsf{R}^m, x = (x_1, \ldots, x_n) \in \mathsf{R}^n \ (m \gg n) \tag{6}$$

The operator F relates the data space and the model space. In other words, if we know the velocity model x, then we can predict the Earth's response based on the velocity model.

3 Inverse Problem

Given an observed data vector $y \in \mathsf{R}^m$, we want to find the unknown model x such that $F(x)$ approximates y as much as possible. For each specific type T of observations, this means that we are minimizing:

$$\min_x \left\| F^T(x) - y^T \right\|^2 = \min_x \left(F_i^T(x) - y_i^T \right)^2 \tag{7}$$

to match measurements of different types, researchers traditionally use weighted nonlinear LSQ method. For example, to simultaneously match the teleseismic receiver functions (RF), surface wave dispersion velocities (SW), travel times (TT), and gravity (GR), we minimize $\min_x J$, where

$$J = w_{\mathrm{RF}}^2 \left\| F^{\mathrm{RF}}(x) - y^{\mathrm{RF}} \right\|^2 + w_{\mathrm{SW}}^2 \left\| F^{\mathrm{SW}}(x) - y^{\mathrm{SW}} \right\|^2 + w_{\mathrm{TT}}^2 \left\| F^{\mathrm{TT}}(x) - y^{\mathrm{TT}} \right\|^2 +$$

$$w_{\mathrm{GR}}^2 \left\| F^{\mathrm{GR}}(x) - y^{\mathrm{GR}} \right\|^2. \tag{8}$$

This minimization problem can be reformulated as:

$$\min_x \|F(x) - y\|^2, \tag{9}$$

where

$$F(x) = W \begin{pmatrix} F^{\mathrm{SW}}(x) \\ F^{\mathrm{RF}}(x) \\ F^{\mathrm{TT}}(x) \\ F^{\mathrm{GR}}(x) \end{pmatrix} \in \mathsf{R}^m,$$

$$y = W \begin{pmatrix} y^{\mathrm{SW}} \\ y^{\mathrm{RF}} \\ y^{\mathrm{TT}} \\ y^{\mathrm{GR}} \end{pmatrix} \in \mathsf{R}^n$$

and

$$W = \mathrm{diag}(w_i), w_i = \sqrt{\frac{\eta_1}{\sigma_i^2 p}}, i = 1, \ldots, p, w_i = \sqrt{\frac{\eta_2}{\sigma_i^2 q}}, i = p+1, \ldots, p+q,$$

$$w_i = \sqrt{\frac{\eta_3}{\sigma_i^2 r}} i = p+q+1, \ldots, p+q+r, \tag{10}$$

$$w_i = \sqrt{\frac{1-\eta_1-\eta_2-\eta_3}{\sigma_i^2 s}}, i = p+q+r+1, \ldots, m = p+q+r+s,$$

with W a weighted diagonal matrix used to equalize the contribution of each data set with respect to physical units and number of data points, $\eta_i \in [0, 1]$ are influence parameters that measures the reliability of each data set used for the inversion, σ_i^2 is the approximate standard deviation of each point, and p, q, r, and s are the number of RF, SW, TT, and GR observations [31].

4 Need for Multiobjective Optimization

In practice, we do not know the exact values of the influence parameters. For different values of the influence parameters, we get, in general, different velocity distributions x; some of these velocity models are geophysically meaningful, some are not (e.g.,

some models x predict higher velocities in the crust and lower velocities in the mantle contrary to geophysics).

Traditionally, researchers avoid nonphysical nonsmooth velocity models by adding a regularization term $\lambda||Lx||^2$ to the minimized function [33]. The problem with this term is that it is not clear how to select λ, and different values of λ lead to different solutions; see, e.g., [10] and [34].

In this work, instead of using regularization, we explicitly formulate constraints that need to be satisfied, for example, the desired smoothness can be described as a bound on $|x_i - x_j| \le \Delta$ on the difference between velocities x_i and x_j at nearby locations. Then, we find the model x for which $J(x)$ is the smallest under these constraints. Additionally, we include bounds $a \le x \le b$ on the velocities at different depths. In geophysical applications, it is crucial to keep the physical parameters within appropriate bounds.

So, instead of selecting a single combination of influence parameters (and thus, of weights), we propose to use multiobjective optimization; namely, we generate all possible models corresponding to different combinations of weights, and then we use one of the MOP criteria to select the most promising model [15, 27, 28].

In this case, we want to mimimize the four criteria $f_1(x) = ||F^{\mathrm{RF}}(x) - y^{\mathrm{RF}}||^2$, $f_2(x) = ||F^{\mathrm{SW}}(x) - y^{\mathrm{SW}}||^2$, $f_3(x) = ||F^{\mathrm{TT}}(x) - y^{\mathrm{TT}}||^2$, $f_4(x) = ||F^{\mathrm{GR}}(x) - y^{\mathrm{GR}}||^2$. First, we find the Pareto optimal set P^*, i.e., the set of all feasible solutions x for which there is no other feasible solution x' which is better with respect to all criteria $f_1(x') < f_1(x), \ldots, f_k(x') < f_k(x)$ (Fig. 6).

Definition (Pareto Optimal Set) For a given multiobjective problem $F(x) = (f_1(x), \ldots, f_k(x))$, *the Pareto optimal set* $P^* \triangleq$, is defined as:

$$P \triangleq \{x \in \Omega | \neg \exists x' \in \Omega (F(x') \le F(x))\} \tag{11}$$

It is known that elements of the Pareto set can be obtained by solving the one-objective (scalar) optimization problem:

$$\min_{x \in X} f(x) = \sum_{i=1}^{k} w_i f_i(x), \tag{12}$$

where $w = (w_1, \ldots, w_k) \ge 0$ is the vector of weighting coefficients assigned prior to the solution of the problem. So, in our computations, we try all possible combinations of weights, and we find all solutions x corresponding to different combinations. For each criterion f_i, we then find the smallest value $f_i^{\min}$ and the largest value $f_i^{\max}$. The smallest values form an "ideal point" $f^{\min} = (f_1^{\min}, \ldots, f_4^{\min})$. We then select a solution x which is the closest to this ideal point. Specifically, we normalize each differences $f_i(x) - f_i^{\min}(x)$ to the interval (0,1) by dividing it by $f_i^{\max}(x) - f_i^{\min}(x)$, and then we minimize the corresponding normalized distance. In other terms, we select a solution x for which the distance

$$d^2(f^{\min}, f(x)) = \sum_{i=1}^{k} \left(\frac{f_i(x) - f_i^{\min}(x)}{f_i^{\max}(x) - f_i^{\min}(x)} \right)^2 \tag{13}$$

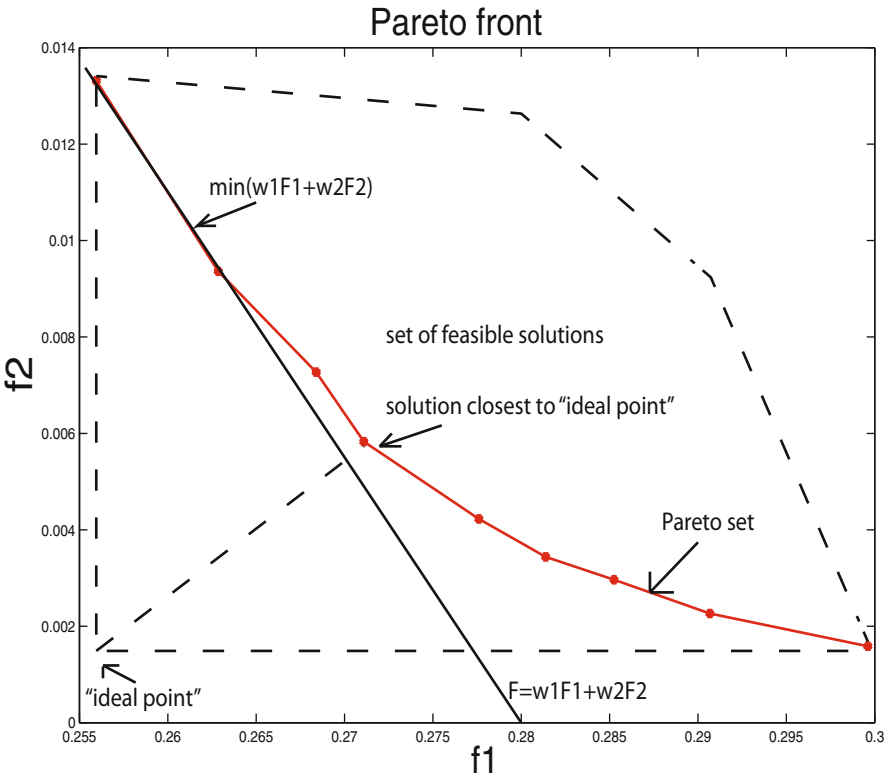


Fig. 6 Illustration of the solution set or Pareto front, which is, defined as the weights times the perspective objective functions

5 Numerical Algorithm

First, we use a first-order Taylor approximation of the operator F around some suitable model $\overline{x}_k$:

$$F(x) \cong F(\overline{x}_k) + F'(\overline{x}_k)\Delta x = F(\overline{x}_k) + F'(\overline{x}_k)(x - \overline{x}_k), \tag{14}$$

where $F'(\overline{x}_k)$ is the matrix formed by the partial derivatives of F. Therefore, we rewrite the problem (9) as:

$$\min_x \frac{1}{2} \left\| F'(\overline{x}_k)x + r(\overline{x}_k) \right\|^2$$

$$\text{s.t. } g(\overline{x}_k) \geq 0 \tag{15}$$

$$g(\overline{x}_k) = \begin{pmatrix} \overline{x}_k - a \\ b - \overline{x}_k \end{pmatrix}$$

where $r(\overline{x}_k) = F(\overline{x}_k) - y - F'(\overline{x}_k)\overline{x}_k$, and $g(\overline{x}_k)$ is a vector of constraints, including constraints $x_i - a_i \geq 0$ and $b_i - x_i \geq 0$ that describe the bounds $a_i \leq x_i \leq b_i$ on velocities x_i at different layers.

6 Primal-Dual Interior-Point Method

To implement the primal-dual interior-point (PDIP) method [24, 31], we first rewrite our problem in a standard form as follows:

$$\begin{gathered} \min_x \frac{1}{2} \left\| F'(\overline{x}_k)x + r(\overline{x}_k) \right\|^2 \\ \text{s.t. } g(\overline{x}_k) - s = 0 \\ s \geq 0 \end{gathered} \tag{16}$$

where $s \in \mathsf{R}^{2n}$ is a slack variable. Then we define the Lagrange function associated to problem (16) as:

$$l(\overline{x}_k, z, s, w) = \frac{1}{2}||F'(\overline{x}_k)x + r(\overline{x}_k)||^2 - (g(\overline{x}_k) - s)^T z - s^T w \tag{17}$$

with the Lagrangian multipliers $z, w \in \mathsf{R}^{2n}$, $(z, w) \geq 0$. For a given perturbation parameter $\mu > 0$, the perturbed Karush–Kuhn–Tucker (KKT) or necessary conditions are given by

$$\hat{F}(\overline{x}_k, z, s, w) = \begin{pmatrix} F'(\overline{x}_k)^T(F'(\overline{x}_k)x + r(\overline{x}_k)) - \nabla g^T(\overline{x}_k)z \\ g(\overline{x}_k) - s \\ z - w \\ SWe - \mu e \end{pmatrix} = 0 \tag{18}$$

where

$$\hat{F} : \mathsf{R}^{n+2n+2n} \longrightarrow \mathsf{R}^{n+2n+2n} \quad S = diag(s_1, \ldots, s_{2n}),\ W = diag(w_1, \ldots, w_{2n})$$

and $e = (1, \ldots, 1) \in \mathsf{R}^{2n}$. It is easy to see that $z = w$, hence the perturbed KKT system (18) is rewritten as:

$$\hat{F}(x, z, s, w) = \begin{pmatrix} F'(\overline{x}_k)^T(F'(\overline{x}_k)x + r(\overline{x}_k)) - \nabla g^T(\overline{x}_k)z \\ g(\overline{x}_k) - s \\ SZe - \mu e \end{pmatrix} = 0, \tag{19}$$

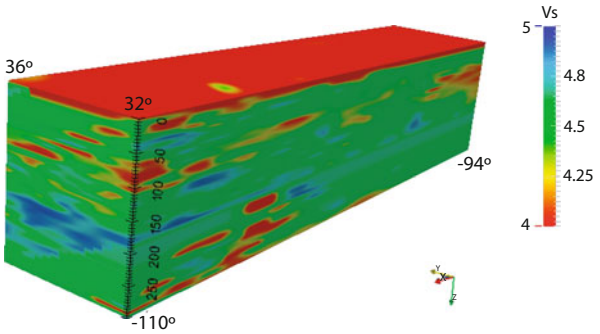


Fig. 7 3D shear wave model utilizing three geophysical data sets using multiobjective optimization (MOP) technique. *Blue* represents high velocities and *red* represents low velocities

thus the Jacobian associated to (19) is then computed as:

$$F'\begin{pmatrix} x \\ z \\ s \end{pmatrix} = \begin{bmatrix} F'(\overline{x}_k)^T F'(\overline{x}_k) & -\nabla g^T(\overline{x}_k) & 0_{nxn} \\ \nabla g(\overline{x}_k) & 0_{nxm} & -I_{mxm} \\ 0_{mxn} & S & Z \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta z \\ \Delta s \end{bmatrix} = - \begin{bmatrix} \nabla_x l(x,z,s) \\ g(\overline{x}_k) - s \\ SZe - \mu e \end{bmatrix} \tag{20}$$

System (20) can be reduced further by eliminating the third block of equations as follows. From the last block of equation in (20) we have

$$S\Delta z + Z\Delta s = -SZe + \mu e,$$

therefore

$$Z\Delta s = -SZe + \mu e - S\Delta z$$

$$\Delta s = -s + \mu Z^{-1}e - Z^{-1}S\Delta z,$$

and then

$$\begin{aligned} \nabla g^T(\overline{x}_k)\Delta x - \Delta s &= \nabla g^T(\overline{x}_k)\Delta x + s - \mu Z^{-1}e + Z^{-1}e + Z^{-1}S\Delta z \\ &= -\nabla g^T(\overline{x}_k)x + s \\ \nabla g^T(\overline{x}_k)\Delta x + Z^{-1}S\Delta z &= \mu Z^{-1}e - g(\overline{x}_k) \end{aligned}$$

which allow us to write the reduced linear system:

$$\begin{bmatrix} -F'(\overline{x}_k)^T F'(\overline{x}_k) & \nabla g^T(\overline{x}_k) \\ \nabla g(\overline{x}_k) & Z^{-1}S \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta z \end{bmatrix} = \begin{bmatrix} \nabla_x l(x,z,s) \\ Z^{-1}\mu e - g(\overline{x}_k) \end{bmatrix} \tag{21}$$

An example of a 3-D model using multiple geophysical datasets is shown in Fig. 7.

7 Conclusion

In summary, for this study we propose to utilize the MOP technique to perform joint inversion of multiple data sets (e.g., receiver functions, surface wave dispersion, and etc). We will incorporate different weights in the MOP inversion scheme in order to map the Pareto set (solution space) of receiver functions and surface wave dispersion measurements. We used the MOP technique to help characterize the crust and upper mantle of an ancient rift system in Texas using seismic data from USArray and EarthScope network. We will extend the PDIP algorithm with the MOP scheme in order to obtain high-resolution 3D imagery of Texas using teleseismic receiver functions, surface wave dispersion measurements, delay travel times, and gravity. We chose this optimization approach because we want to find the best possible solution for our nonlinear geophysics inverse problem. In geophysics, most inversion problems require finding some minimization. The optimization technique that we chose to solve our nonlinear inverse problem requires the search of the global minimum and this technique will be able to define the entire solution based from using different weights to map the Pareto set. From the Pareto set, the MOP technique performs a direct search method that basically selects the final solution from a set of alternative solutions from the model space [15, 27, 28]. For future work, we plan to incorporate gravity into our 3D model to be able to obtain a more constrained earth structure model of Texas, which will allow us to help answer questions such as if the rift system is still actively deforming and how does the rift influence the evolution of adjacent areas within the North American Plate.

Acknowledgment We would like to take the time to thank the computational science, mathematical science, and computer science departments from University of Texas at El Paso (UTEP). We would also like to thank Ezer Patlan, Dr. Anibal Sosa, Dr. Rodrigo Romero, Dr. Monica Maceira, and Azucena Zamora for all of their contributions to this work. This work was sponsored by the NSF CREST under Grant Cybershare HRD-0734825.

References

1. Astiz, L., Earle, P., Shearer, P.: Global stacking of broadband seismograms. Seis. Res. Lett. **67**, 8–18 (1996)
2. Bashir, L., Gao, S.S., Liu, K.H., Mickus, K.: Crustal structure and evolution beneath the Colorado Plateau and the southern Basin and Range Province: results from receiver function and gravity studies. Geochem. Geophys. Geosyst. **12**, Q06008 (2011). doi:10.1029/2011GC003563
3. Bailey, I.W., Miller, M.S., Liu, K., Levander, A.. Vs and density structure beneath the Colorado Plateau constrained by gravity anomalies and joint inversions of receiver function and phase velocity data. J. Geophys. Res. **117**, B02313 (2012). doi:10.1029/2011JB0085
4. Bodin, T., Sambridge, M., Tkalcic, H., Arroucau, P., Gallagher, K., Rawlinson, N.: Transdimensional inversion of receiver functions and surface wave dispersion. J. Geophys. Res. **117** (2012). doi:10.1029/2011JB008560

5. Cho, K.H., Herrmann, R.B., Ammon, C.J., Lee, K.: Imaging the upper crust of the Korean Peninsula by surface-wave tomography. Bull. Seismol. Soc. Am. **97**, 198–207 (2007)
6. Colombo, D., De Stefano, M.: Geophysical modeling via simultaneous joint inversion of seismic, gravity, and electromagnetic data: Application to prestack depth imaging. Leading Edge **26**, 326–331 (2007)
7. Dzierma, Y., Rabbel, W., Thorwart, M.M., Flueh, E.R., Mora, M.M., Alvarado, G.E.: The steeply subducting edge of the Cocos Ridge: evidence from receiver functions beneath the northern Talamanca Range, south-central Costa Rica. Geochem. Geophys. Geosyst. **12** (2011). doi:10.1029/2010GC003477
8. Gurrola, H., Baker, E.G., Minster, B.J.. Simultaneous time-domain deconvolution with application to the computation of receiver functions. Geophys. J. Int. **120**, 537–543 (1995)
9. Haber, E., Oldenburg, D.: Joint inversion: A structural approach. Inverse Probl. **13**, 63–77 (1997)
10. Hansen, P.C.: Discrete inverse problems: Insight and algorithms, 225 pp. Soc. Ind. Appl. Math. Philadelphia, Pa. (2010)
11. Hansen, S.M., Dueker, K.G., Stachnik, J.C., Aster, R.C., Karlstrom, K.E.: A rootless rockies - Support and lithospheric structure of the Colorado Rocky Mountains inferred from CREST and TA seismic data. Geochem. Geophys. Geosyst. **14**, 2670–2695 (2013). doi:10.1002/ggge.20143
12. Hackney, R.I., Featherstone, W.E.: Geodetic versus geophysical perspectives of the gravity anomaly. Geophys. J. Int. **154**(1), 35–43 (2003)
13. Heiskanen, W.A., Moritz, H.: Physical geodesy. W. H. Freeman and Company, San Francisco (1967)
14. Julia, J., Ammon, C.J., Hermann, R., Correig, M.: Joint inversion of receiver function and surface wave dispersion observations. Geophys. J. Int. **142**, 99–112 (2000)
15. Kozlovskaya, E.: An algorithm of geophysical data inversion based on non-probabilistic presentation of a-prior information and definition of pareto-optimality. Inverse Probl. **16**, 839–861 (2000)
16. Langston, C.A.: Evidence for the subducting lithosphere under southern Vancouver Island and western Oregon from teleseismic P wave conversions. J. Geophys. Res. **86**, 3857–3866 (1981)
17. Laske, G., Masters, G., Reif, C.: Crust 2.0. the current limits of resolution for surface wave tomography in North America. EOS Trans. AGU **81** F897 (2000). http://igpppublic.ucsd.edu/gabi/ftp/crust2/
18. Lees, J.M., Vandecar, J.C.: Seismic tomography constrained by bouguer gravity anomalies: Applications in western Washington. PAGEOPH. **135**, 31–52 (1991)
19. Lin, F.C., Schmandt, B., Tsai, V.C.: Joint inversion of Rayleigh wave phase velocity and ellipticity using USArray: Constraining velocity and density structure in the upper crust. Geophys. Res. Lett. **39**, L12303 (2012). doi:10.1029/2012GL052196
20. Lodge, A., Helffrich, G.: Grid search inversion of teleseismic receiver functions. Geophys. J. Int. **178**, 513–523 (2009)
21. Maceira, M., Ammon, C.J.: Joint inversion of surface wave velocity and gravity observations and its application to central Asian basins s-velocity structure. J. Geophys. Res. **114**, B02314 (2009). doi:10.1029/2007JB0005157.
22. Moorkamp, M., Jones, A.G., Fishwick, S.: Joint inversion of receiver functions, surface wave dispersion, and magnetotelluric data. J. Geophys. Res. **115**, B04318 (2010). doi:10.1029/2009JB0006369
23. Moorkamp, M., Heincke, B., Jegen, M., Roberts, A.W., Hobbs, R.W.: A framework for 3–D joint inversion of MT, gravity and seismic refraction data. Geophys. J. Int. **184**, 477–493 (2011)
24. Nocedal, J., Wright, S.: Numerical optimization. 2nd edn. Springer, New York (2006)
25. Obrebski, M., Kiselev, S., Vinnik, L., Montagner, J.P.: Anisotropic stratification beneath Africa from joint inversion of SKS and P receiver functions. J. Geophys. Res. **115**, B09313 (2010). doi:10.1029/2009JB006923

26. Owens, T.J., Crotwell, H.P., Groves, C., Oliver-Paul, P.: SOD: Standing order for data. Seismol. Res. Lett. **75**, 515–520 (2004)
27. Sambridge, M.: Geophysical inversion with a neighborhood algorithm I: Searching a parameter space. Geophys. J. Int. **138**, 479–494 (1999)
28. Sambridge, M.: Geophysical inversion with a neighborhood algorithm II: Appraising the ensemble. Geophys. J. Int. **138**, 727–746 (1999)
29. Shearer, P.M.: Introduction to Seismology, 2nd edn. Cambridge University Press, Cambridge (2009)
30. Shen, W., Ritzwoller, M.H., Schulte-Pelkum, V.: A 3-D model of the crust and uppermost mantle beneath the Central and Western US by joint inversion of receiver functions and surface wave dispersion. J. Geophys. Res. Solid Earth **118** (2013). doi:10.1029/2012JB009602
31. Sosa, A., Velasco, A.A., Velasquez, L., Argaez, M., Romero, R.: Constrained Optimization framework for joint inversion of geophysical data sets. Geophys. J. Int. **195**, 197–211 (2013)
32. Stein, S., Wysession, M.: An introduction to seismology, earthquakes, and earth structure. Blackwell, Maiden (2003)
33. Tikhonov, A.N., Arsenin, V.Y.: Solutions if Ill-posed Problems. Winston and Sons, Washington (1977)
34. Vogel, C.R.: Computational methods for inverse problems. SIAM FR23, Philadelphia, (2002)
35. Vozoff, K., Jupp, D.L.B.: Joint inversion of geophysical data. Geophys. J. R. Astr. Soc. **42**, 977–991 (1975)
36. Wilson, D.: Imagining crust and upper mantle seismic structure in the southwestern United States using teleseismic receiver functions. Leading Edge **22**, 232–237 (2003)
37. Wilson, D., Aster, R.: Seismic imaging of the crust and upper mantle using Regularized joint receiver functions, frequency-wave number filtering, and Multimode Kirchhoff migration. J. Geophys. Res. B05305 (2005). doi:10.1029/2004JB003430
38. Wilson, D., Aster, R., Ni, J., Grand, S., West, M., Gao, W., Baldridge, W.S., Semken, S.: Imaging the structure of the crust and upper mantle beneath the Great Plains, Rio Grande Rift, and Colorado Plateau using receiver functions. J. Geophys. Res. **110**, B05306 (2005). doi:10.1029/2004JB003492
39. Zhu, L., Kanamori, H.: Moho depth variation in southern California from teleseismic receiver functions. J. Geophys. Res. **105**, 2969–2980 (2000)

Multiobjective Semi-infinite Optimization: Convexification and Properly Efficient Points

Francisco Guerra-Vásquez and Jan-Joachim Rückmann

This work was partially supported by Sistema Nacional de Investigadores (SNI, México) under grant 14480.

Abstract This chapter deals with nonconvex semi-infinite optimization problems that are defined by finitely many objective functions and infinitely many inequality constraints in a finite-dimensional space. Under the reduction approach, it is shown that locally around an efficient point this problem can be transformed equivalently in such a way that the Lagrangian of the transformed weighted sum optimization problem becomes locally convex. Consequently, local duality theory and corresponding solution methods can be used after applying this convexification procedure. Furthermore, the strong relationship between properly efficient points of both the original and the transformed problems is discussed.

Keywords Semi-infinite optimization · Multiobjective optimization · Reduction approach · Convexification procedures · Properly efficient points

1 Introduction

In this chapter, we consider nonlinear multiobjective semi-infinite optimization problems (MOSIPs). *Semi-infinite* means that we have *infinitely* many constraints and finitely many objective functions defined on a *finite*-dimensional space. Semi-infinite optimization became a very vivid area of research over the past two decades; we refer

F. Guerra-Vázquez (✉)
Escuela de Ciencias, Universidad de las Américas, Puebla,
San Andrés Cholula, 72820 Puebla, México
e-mail: francisco.guerra@udlap.mx

J.-J. Rückmann
Department of Informatics, University of Bergen,
Postbox 7803, 5020 Bergen, Norway
e-mail: Jan-Joachim.Ruckmann@ii.uib.no

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_23

to some recent books and monographs [6, 24, 25, 27] as well as to the standard book [3] on multiobjective optimization.

As a starting point of this chapter, we define an MOSIP as follows:

$$\text{MOSIP "min" } f(x) \text{ s.t. } x \in M,$$

where $f = (f_1, \ldots, f_q)^\top$ is the vector of objective functions $f_i \in C^2(\mathbb{R}^n, \mathbb{R})$, $i = 1, \ldots, q$ ($C^k(\mathbb{R}^n, \mathbb{R})$ denotes the space of k-times continuously differentiable real-valued functions defined on $\mathbb{R}^n$) and

$$M = \left\{ x \in \mathbb{R}^n \mid g(x, y) \leq 0,\ y \in Y \right\}$$

is the feasible set. Here, $g \in C^2(\mathbb{R}^n \times \mathbb{R}^m, \mathbb{R})$ and $Y \subset \mathbb{R}^m$ is a compact—and, in general, infinite—index set. Note that each $\overline{y} \in Y$ represents a corresponding constraint $g(x, \overline{y}) \leq 0$.

Given a feasible point $\overline{x} \in M$, we define the set of active inequality constraints at $\overline{x}$ as:

$$Y_0(\overline{x}) = \{ y \in Y \mid g(\overline{x}, y) = 0 \}.$$

In order to recall the well-known notations of solutions for MOSIP, let the vector $c \in \mathbb{R}^n$ denote its components by $c_i, i = 1, \ldots, n$ and for $c, e \in \mathbb{R}^n$ let:

- $c \leqq e$, if $c_i \leq e_i, i = 1, \ldots, n$
- $c < e$, if $c_i < e_i, i = 1, \ldots, n$
- $c \leq e$, if $c_i \leq e_i, i = 1, \ldots, n$ and $c \neq e$

Definition 1

(i) A point $\overline{x} \in M$ is called *efficient* (for the problem MOSIP) if there does not exist any $x \in M$ with $f(x) \leq f(\overline{x})$.

(ii) A point $\overline{x} \in M$ is called *locally efficient* (for the problem MOSIP) on $B(\overline{x}, \varepsilon)$ if there is a real number $\varepsilon > 0$ and if there does not exist any $x \in B(\overline{x}, \varepsilon) \cap M$ with $f(x) \leq f(\overline{x})$ (here, $B(\overline{x}, \varepsilon) = \{ x \in \mathbb{R}^n \mid \|x - \overline{x}\| < \varepsilon \}$ and $\|\cdot\|$ denotes the Euclidean norm).

In this chapter, we will assume that at a point under consideration, say $\overline{x} \in M$, the so-called reduction approach (RA) holds. The basic definition of this generic property will be reviewed in Sect. 3; for more details we refer to [9, 11, 12, 14, 15, 28]. In particular, the RA implies that locally around $\overline{x}$ the feasible set can be described by *finitely* many C^2-inequality constraints $\widetilde{g}_j(x)$ ($= g(x, \overline{y}^j(x))$ in (9)), $j = 1, \ldots, s$ as

$$M \cap U = \{ x \in U \mid \widetilde{g}_j(x) \leq 0,\ j = 1, \ldots, s \}$$

where $U \subset \mathbb{R}^n$ is an appropriate neighbourhood of $\overline{x}$. Then, we define locally on U the Lagrangian for the so-called *weighted sum*optimization problem associated to MOSIP as

$$L(x, \lambda, \mu) = \sum_{i=1}^{q} \lambda_i f_i(x) + \sum_{j=1}^{s} \mu_j \widetilde{g}_j(x), \tag{1}$$

where $\lambda_i > 0$ are the positive weights for f_i, $i = 1, \ldots, q$ and $\mu \geq 0_s$ (where 0_s denotes the origin in $\mathbb{R}^s$).

Frequently, primal/dual solution methods and propositions within local duality theory use the assumption that the partial Hessian with respect to x of the Lagrangian (1) is positive definite. This is a very strong assumption which, in general, is not fulfilled for nonconvex problems. In this chapter, we apply the so-called p-power transformation to MOSIP where the functions of the original problem are substituted by their pth power; e.g. f_i by $(f_i)^p$, $i = 1, \ldots, q$. This technique was applied to standard optimization problems in [8, 18–20, 29].

The objective of this chapter is twofold. First, we will show that the p-power transformation is a convexification procedure for the Lagrangian associated to MOSIP. More precisely, assuming that the RA, an appropriate constraint qualification and an appropriate second-order condition hold at $\overline{x}$, we will show that for any sufficiently large power p the Lagrangian of the p-power transformation becomes convex locally around $\overline{x}$. Note that the original problem and the transformed problem have the same feasible set and several solution properties remain unchanged as well. Second, in multiobjective optimization, an efficient point can be further qualified as *properly efficient* point (in the sense of Geoffrion [5] or in the sense of Kuhn and Tucker [17]; see Sect. 2 for definitions). We will show that the property of being a properly efficient point is invariant under p-power transformation.

This chapter is organized as follows. Section 2 provides some basic results on locally properly efficient points and constraint qualifications. In Sect. 3, we recall the idea of the RA. Section 4 contains the main results on the relationship between the original problem and its p-power transformation locally around a particular point under consideration. An illustrating example is presented in Sect. 5, and Sect. 6 yields some conclusions.

Finally, we explain some notations. For a function $h \in C^1(\mathbb{R}^n, \mathbb{R})$, denote by the *row* vector $Dh(\overline{x})$ ($D_{x^1}h(\overline{x})$) the gradient of h (partial gradient of h with respect to the subvector x^1 of x) at $\overline{x} \in \mathbb{R}^n$. For $h \in C^2(\mathbb{R}^n, \mathbb{R})$, the second derivatives are analogously defined.

2 Some Basic Results

In this section we recall some basic properties of a *standard* multiobjective optimization problem (MOP); here, *standard* means that it has only finitely many inequality constraints. In particular, we recall definitions on properly efficient points as well as some constraint qualifications. Throughout this section, consider a standard MOP of the form

$$\text{MOP "min" } f(x) \text{ s.t. } x \in X$$

where $f = (f_1, \ldots, f_q)^\top$ is defined as in Sect. 1 and

$$X = \left\{ x \in \mathbb{R}^n \mid g_j(x) \leq 0,\ j = 1, \ldots, s \right\}$$

with $g_j \in C^1(\mathbb{R}^n, \mathbb{R})$, $j = 1, \ldots, s$. For $\overline{x} \in X$ define the set of active indices at $\overline{x}$ as

$$J_0(\overline{x}) = \{j \in \{1, \ldots, s\} \mid g_j(\overline{x}) = 0\}.$$

Locally Properly Efficient Points In the following definition of two kinds of locally properly efficient points, we assume that the points under consideration are locally efficient for the problem MOP where the notation is analogously defined as in Definition 1 for the problem MOSIP. The following definitions refer to properly efficient points that where introduced by *Geoffrion* in [5] as well as by *Kuhn* and *Tucker* in [17].

Definition 2 (See e.g. [3, 5, 17]).

(i) A point $\overline{x} \in X$ is called *locally properly efficient* (for the problem MOP) *in the sense of Geoffrion* (shortly: *G-locally properly efficient*) if there exists a real number $\varepsilon > 0$ such that
- $\overline{x}$ is locally efficient (for the problem MOP) on $B(\overline{x}, \varepsilon)$.
- There exists a real number $K > 0$ such that for each index $i \in \{1, \ldots, q\}$ and any $x \in B(\overline{x}, \varepsilon) \cap X$ with $f_i(x) < f_i(\overline{x})$, there exists an index $j \in \{1, \ldots, q\}$ such that $f_j(x) > f_j(\overline{x})$ and

$$\frac{f_i(\overline{x}) - f_i(x)}{f_j(x) - f_j(\overline{x})} \leq K.$$

(ii) A point $\overline{x} \in X$ is called *locally properly efficient* (for the problem MOP) *in the sense of Kuhn and Tucker* (shortly: *KT-locally properly efficient*) if there exists a real number $\varepsilon > 0$ such that
- $\overline{x}$ is locally efficient (for the problem MOP) on $B(\overline{x}, \varepsilon)$.
- The following system has no solution $d \in \mathbb{R}^n$:

$$\begin{aligned} &Df_i(\overline{x})d \leq 0, \ i = 1, \ldots, q, \\ &Df_k(\overline{x})d < 0, \ \text{for some } k \in \{1, \ldots, q\}, \\ &Dg_j(\overline{x})d \leq 0, \ j \in J_0(\overline{x}). \end{aligned}$$

The role of $K > 0$ in Definition 2 can be interpreted as follows. The ratio (which is also called *trade-off*, see e.g. [3]) between the improvement of one objective function ($f_i(x) < f_i(\overline{x})$) and the decrease of another objective function ($f_j(x) > f_j(\overline{x})$) is bounded by the finite number $K > 0$. Note that in Definition 2 for any such index $i \in \{1, \ldots, q\}$ and point x with $f_i(x) < f_i(\overline{x})$, there has to exist a corresponding index $j \in \{1, \ldots, q\}$ with $f_j(x) > f_j(\overline{x})$ since $\overline{x}$ is locally efficient.

The difference between these two concepts of locally properly efficient points can be characterized by the fact whether or not a corresponding constraint qualification is fulfilled, see for example the next lemma or, more general, [7]. We also refer to the (longer) Example 4.2 in [7] where both concepts are illustrated for several particular problems.

Constraint Qualifications We recall the following two well-known constraint qualifications:

The *Mangasarian–Fromovitz constraint qualification* (MFCQ; cf. [21, 22]) is said to hold at $\overline{x} \in X$ if there exists a vector $d \in \mathbb{R}^n$ such that

$$Dg_j(\overline{x})d < 0, \ j \in J_0(\overline{x}). \tag{2}$$

The *Kuhn–Tucker constraint qualification* (KTCQ; cf. [17]) is said to hold at $\overline{x} \in X$ if for any vector $d \in \mathbb{R}^n \backslash \{0_n\}$ satisfying

$$Dg_j(\overline{x})d \leq 0, \ j \in J_0(\overline{x}),$$

there exist real numbers $\overline{t} > 0, \alpha > 0$ and a continuously differentiable path

$$\vartheta : t \in [0, \overline{t}] \mapsto \vartheta(t) \in \mathbb{R}^n$$

such that $\vartheta(0) = \overline{x}$, $\vartheta'(0) = \alpha d$ and $\vartheta(t) \in X$ for all $t \in [0, \overline{t}]$.

The next lemma summarizes some known results.

Lemma 1

(i) *If MFCQ holds at $\overline{x} \in X$, then KTCQ holds at $\overline{x} \in X$ as well.*
(ii) *If KTCQ holds at $\overline{x} \in X$ and $\overline{x}$ is G-locally properly efficient, then $\overline{x}$ is KT-locally properly efficient.*
(iii) *If $\overline{x} \in X$ is KT-locally properly efficient, then there exist $\overline{\lambda} > 0_q$ and $\overline{\mu} \geqq 0_s$ such that*

$$\sum_{i=1}^{q} \overline{\lambda}_i Df_i(\overline{x}) + \sum_{j=1}^{s} \overline{\mu}_j Dg_j(\overline{x}) = 0_n, \ \sum_{j=1}^{s} \overline{\mu}_j g_j(\overline{x}) = 0. \tag{3}$$

(iv) *If $\overline{x} \in X$ is G-locally properly efficient and MFCQ holds at $\overline{x}$, then there exist $\overline{\lambda} > 0_q$ and $\overline{\mu} \geqq 0_s$ satisfying (3).*

Proof

(i) See [1, Lemma 5.2.1]
(ii) See [3, Theorem 2.51]
(iii) See [3, Theorem 3.25]
(iv) Consequence of (i), (ii) and (iii) $\triangle$

The following lemma is related to the local convexity of the Lagrange function. For the sake of completeness, we present its proof that contains parts of the proof of Theorem 3.11 in [3].

Lemma 2 (see [3, Theorem 3.11]). *Let $\overline{x} \in X$ and let $\overline{\lambda} > 0_q$ and $\overline{\mu} \geqq 0_s$ such that (3) is fulfilled. If the Lagrange function*

$$L(\cdot, \overline{\lambda}, \overline{\mu}) = \sum_{i=1}^{q} \overline{\lambda}_i f_i(\cdot) + \sum_{j=1}^{s} \overline{\mu}_j g_j(\cdot)$$

is convex on $B(\overline{x}, \varepsilon)$ for some $\varepsilon > 0$, then $\overline{x}$ is G-locally properly efficient.

Proof Since $L(\cdot, \overline{\lambda}, \overline{\mu})$ is convex on $B(\overline{x}, \varepsilon)$, we get from (3) that $\overline{x}$ is a global minimizer of $L(\cdot, \overline{\lambda}, \overline{\mu})_{|B(\overline{x},\varepsilon)}$, that is

$$L(\overline{x}, \overline{\lambda}, \overline{\mu}) \leq L(x, \overline{\lambda}, \overline{\mu}) \text{ for all } x \in B(\overline{x}, \varepsilon). \tag{4}$$

In order to prove that $\overline{x}$ is locally efficient, assume the contrary. Assume that there exists $\widehat{x} \in B(\overline{x}, \varepsilon) \cap X$ such that $f(\widehat{x}) \leq f(\overline{x})$. Then, by $\widehat{\lambda} > 0_q$, we obtain:

$$\sum_{i=1}^{q} \overline{\lambda}_i f_i(\widehat{x}) < \sum_{i=1}^{q} \overline{\lambda}_i f_i(\overline{x}) \text{ and}$$

$$\sum_{j=1}^{s} \overline{\mu}_j g_j(\widehat{x}) \leq \sum_{j=1}^{s} \overline{\mu}_j g_j(\overline{x}) \ (=0),$$

which contradicts (4).

In order to prove that $\overline{x}$ is G-locally properly efficient, assume the contrary. Choose

$$K = (q-1) \max_{i,j} \frac{\overline{\lambda}_j}{\overline{\lambda}_i}.$$

Then, there exist an index $i_0 \in \{1, \ldots, q\}$ and a point $x^0 \in B(\overline{x}, \varepsilon) \cap X$ such that

$$f_{i_0}(x^0) < f_{i_0}(\overline{x})$$

and for each $j \in \{1, \ldots, q\}$ with $f_j(x^0) > f_j(\overline{x})$, we have

$$f_{i_0}(\overline{x}) - f_{i_0}(x^0) > K \left(f_j(x^0) - f_j(\overline{x}) \right).$$

The latter inequality yields for all $j \in \{1, \ldots, q\} \backslash \{i_0\}$ that

$$f_{i_0}(\overline{x}) - f_{i_0}(x^0) > (q-1) \frac{\overline{\lambda}_j}{\overline{\lambda}_{i_0}} \left(f_j(x^0) - f_j(\overline{x}) \right)$$

and, hence

$$\frac{1}{q-1} \overline{\lambda}_{i_0} \left(f_{i_0}(\overline{x}) - f_{i_0}(x^0) \right) > \overline{\lambda}_j \left(f_j(x^0) - f_j(\overline{x}) \right), \ j \in \{1, \ldots, q\} \backslash \{i_0\}. \tag{5}$$

After summing up the $(q-1)$ inequalities in (5), we obtain

$$\sum_{i=1}^{q} \overline{\lambda}_i f_i(\overline{x}) > \sum_{i=1}^{q} \overline{\lambda}_i f_i(x^0)$$

and, therefore

$$L(\overline{x}, \overline{\lambda}, \overline{\mu}) > L(x^0, \overline{\lambda}, \overline{\mu})$$

which contradicts (4). This completes the proof. $\triangle$

In our previous chapter [7], we presented two approximation problems as applications of MOSIPs. As a motivation for this chapter, we quote one of these applications in the following.

Example 1 Simultaneous Chebyshev best approximation.

This example is completely taken from [7] *including the terminology*. The following simultaneous Chebyshev best approximation problem can be derived from an abstract characterization theory of efficiency and is a useful model for many practical applications (for more details and, in particular, necessary conditions, see [2]). Consider an interval $[a, b] \subset \mathbb{R}$ and a set of p ($p > 1$) continuous real-valued functions:

$$\psi_0^i : [a, b] \to \mathbb{R},\ i = 1, \ldots, p$$

as well as for each index $i_0 \in \{1, \ldots, p\}$ a corresponding family of n continuous real-valued functions:

$$\psi_k^{i_0} : [a, b] \to \mathbb{R},\ k = 1, \ldots, n.$$

Define the difference between ψ_0^i and a linear combination of the (approximation) functions ψ_k^i, $k = 1, \ldots, n$ as:

$$f_i(x) = \max_{y \in [a,b]} \left| \psi_0^i(y) - \sum_{k=1}^{n} x_k \psi_k^i(y) \right|,\ i = 1, \ldots, p,$$

where $x \in \mathbb{R}^n$ is varying in a given (feasible) set $M^1 \subseteq \mathbb{R}^n$. Then, the *simultaneous Chebyshev best approximation problem* is to solve the following non-differentiable MOP:

$$\text{“min”}\ \left(f_1(x), \ldots, f_p(x)\right)^\top \text{ s.t. } x \in M^1.$$

A standard *epigraph reformulation* of the latter problem provides (with auxiliary variables $q \in \mathbb{R}^p$)

$$\begin{aligned} &\text{“min”}\ \left(q_1, \ldots, q_p\right)^\top \text{ s.t.} \\ &\max\nolimits_{y \in [a,b]} \left| \psi_0^i(y) - \sum_{k=1}^{n} x_k \psi_k^i(y) \right| \le q_i,\ i = 1, \ldots, p. \end{aligned}$$

The latter problem can be rewritten as a differentiable MOSIP with $Y = [a, b]$ as follows:

$$\begin{aligned} &\underset{(x,q) \in M^1 \times \mathbb{R}^p}{\text{“min”}}\ \left(q_1, \ldots, q_p\right)^\top \text{ s.t.} \\ &\psi_0^i(y) - \sum_{k=1}^{n} x_k \psi_k^i(y) \le q_i,\ i = 1, \ldots, p,\ y \in Y, \\ &-\psi_0^i(y) + \sum_{k=1}^{n} x_k \psi_k^i(y) \le q_i,\ i = 1, \ldots, p,\ y \in Y. \end{aligned}$$

Note that this problem has a more general form than MOSIP since it contains more than one (but finitely many) inequality constraints of the form $g(x, y) \leq 0$, $y \in Y$. However, all results presented in this chapter can be generalized straightforwardly to this more general case. This completes this example.

3 The Reduction Approach

In this section, we return to the MOSIP. We will recall the so-called RA; if this approach holds at a point $\overline{x} \in M$, then the problem MOSIP can, locally around $\overline{x}$, be described by *finitely* many continuously differentiable constraints. This latter property refers to an important feature of solution methods for semi-infinite problems where the original problem with *infinitely* many constraints has to be transformed (locally or as an approximation) into one with *finitely*many constraints. We will mention at the end of this section that the RA is a generic property; for a detailed study on this topic, see e.g. [9, 11, 12, 14, 15, 28]. Consider MOSIP and assume that the index set $Y \subset \mathbb{R}^m$ is given as

$$Y = \{y \in \mathbb{R}^m \mid u_l(y) = 0,\ l \in A,\ v_k(y) \leq 0,\ k \in B\}$$

where $A = \{1, \ldots, a\}$, $a < m$, $B = \{1, \ldots, b\}$ and $u_l, v_k \in C^2(\mathbb{R}^m, \mathbb{R})$, $l \in A$, $k \in B$. Furthermore, assume throughout this section the generic property that the linear independence constraint qualification (LICQ) holds at each $\overline{y} \in Y$, that is, the gradients $Du_l(\overline{y})$, $Dv_k(\overline{y})$, $l \in A$, $k \in B_0(\overline{y})$ are linearly independent where

$$B_0(\overline{y}) = \{k \in B \mid v_k(\overline{y}) = 0\}.$$

Obviously, for $\overline{x} \in M$, each index $\overline{y} \in Y_0(\overline{x})$ is a global maximizer of the so-called lower level problem (which depends on the parameter $\overline{x}$)

$$LL(\overline{x}) \max\ g(\overline{x}, y) \text{ s.t. } y \in Y \tag{6}$$

and, by LICQ, there exist uniquely determined multipliers $\overline{\alpha}_l$, $l \in A$, $\overline{\delta}_k \geq 0$, $k \in B_0(\overline{y})$ such that

$$D_y g(\overline{x}, \overline{y}) - \sum_{l \in A} \overline{\alpha}_l Du_l(\overline{y}) - \sum_{k \in B_0(\overline{y})} \overline{\delta}_k Dv_k(\overline{y}) = 0_m. \tag{7}$$

We recall that the *strong second-order sufficient condition* (SSOSC) is said to hold at the global maximizer $\overline{y} \in Y_0(\overline{x})$ of (6) if the matrix $V^\top H V$ is positive definite where

$$H = -D_y^2 g(\overline{x}, \overline{y}) + \sum_{l \in A} \overline{\alpha}_l D^2 u_l(\overline{y}) + \sum_{k \in B_0(\overline{y})} \overline{\delta}_k D^2 v_k(\overline{y})$$

and V is a matrix whose columns form a basis of the subspace

$$\{y \in \mathbb{R}^m \mid Du_l(\overline{y})y = 0,\ Dv_k(\overline{y})y = 0,\ l \in A,\ k \in B_+(\overline{y})\}$$

with

$$B_+(\overline{y}) = \{k \in B_0(\overline{y}) \mid \overline{\delta}_k > 0\}.$$

If the latter tangent space is trivial (that is, $\{0\}$), then SSOSC is, by definition, fulfilled as well. The following lemma is a straightforward conclusion from the implicit function theorem.

Lemma 3 (cf. e.g. [9, 10]). *Let $\overline{x} \in M$ and $\overline{y} \in Y_0(\overline{x})$. Assume that SSOSC holds at $\overline{y}$ such that $B_0(\overline{y}) = B_+(\overline{y})$. Then, there exist an open neighbourhood $\overline{U}$ of $\overline{x}$ and a uniquely determined continuously differentiable function*

$$\widetilde{y} : x \in \overline{U} \mapsto \widetilde{y}(x) \in \mathbb{R}^m$$

such that $\widetilde{y}(\overline{x}) = \overline{y}$, the point $\widetilde{y}(x)$ is a local maximizer of $LL(x)$ for each $x \in \overline{U}$, $g(\cdot, \widetilde{y}(\cdot)) \in C^2(\overline{U}, \mathbb{R})$, and $\widetilde{y}(x)$ is a locally unique maximizer of $LL(x)$ around $\overline{y}$. △

Now, we are able to recall the following. The RA is said to hold at $\overline{x} \in M$ if $B_0(y) = B_+(y)$ for all $y \in Y_0(\overline{x})$ and SSOSC holds at all $y \in Y_0(\overline{x})$. As a consequence of the compactness of the index set Y and Lemma 3, we obtain the following corollary:

Corollary 1 *Let $\overline{x} \in M$ and assume that RA holds at $\overline{x}$. Then, we have:*

- *$Y_0(\overline{x})$ is a finite set, say $Y_0(\overline{x}) = \{y^1, \ldots, y^s\}$.*
- *There exist an open neighbourhood U of $\overline{x}$ and uniquely determined continuously differentiable functions*

$$\overline{y}^j : x \in U \mapsto \overline{y}^j(x) \in \mathbb{R}^m, \ j = 1, \ldots, s \tag{8}$$

such that $\overline{y}^j(\overline{x}) = y^j$, $g(\cdot, \overline{y}^j(\cdot)) \in C^2(U, \mathbb{R})$, $j = 1, \ldots, s$ and

$$M \cap U = \{x \in U \mid g(x, \overline{y}^j(x)) \leq 0, \ j = 1, \ldots, s\}. \tag{9}$$

△

Assuming that RA holds at $\overline{x} \in M$, the statement (9) means that the feasible set M of MOSIP can be described locally around $\overline{x}$ by finitely many continuously differentiable inequality constraints. Although the functions in (8) are only implicitly known, a short calculation shows that for $x \in U$, we have

$$Dg(x, \overline{y}^j(x)) = D_x g(x, y)_{|y=y^j(x)}, \ j = 1, \ldots, s. \tag{10}$$

In the following, we will assume that RA holds at our point under consideration.

4 Main Results

4.1 The Setting

In the remainder of this chapter, let $\overline{x} \in M$ be our point under consideration and assume the following:

- RA holds at $\overline{x}$; we will use the notations from Corollary 1 which means in particular that locally around $\overline{x}$, the feasible set M can be described as stated in (9).
- The extended MFCQ (EMFCQ; (cf. [13]) holds at $\overline{x}$, that is, there exists $d \in \mathbb{R}^n$ satisfying

$$D_x g(\overline{x}, y)d < 0, \ y \in Y_0(\overline{x}).$$

- There exist (fixed) $\overline{\lambda} > 0_q$ as well as multipliers $\mu_j \geq 0$, $j = 1, \ldots, s$ such that

$$\sum_{i=1}^{q} \overline{\lambda}_i Df_i(\overline{x}) + \sum_{j=1}^{s} \mu_j D_x g(\overline{x}, y^j) = 0_n, \ \mu \geq 0_s. \tag{11}$$

  Then, by EMFCQ, the set

$$\mathcal{M} = \{\mu \in \mathbb{R}^s \mid \mu \text{ is a solution of (11)}\}$$

  is compact (cf. [4, 16]).
- The following extended SSOSC (ESSOSC; cf. [26]) holds at $\overline{x}$: For each $\mu \in \mathcal{M}$ the matrix

$$\sum_{i=1}^{q} \overline{\lambda}_i D^2 f_i(\overline{x}) + \sum_{j=1}^{s} \mu_j D^2 g(\overline{x}, y^j(\overline{x}))$$

  is positive definite on the subspace

$$T(\mu) = \{w \in \mathbb{R}^n \mid D_x g(\overline{x}, y^j)w = 0, \ j \in \{\nu \in \{1, \ldots s\} \mid \mu_\nu > 0\}\}.$$

By [26], these assumptions imply that $\overline{x}$ is a local minimizer of the—locally defined—problem

$$\min_{x \in U} \sum_{i=1}^{q} \overline{\lambda}_i f_i(x) \text{ s.t. } g(x, \overline{y}^j(x)) \leq 0, \ j = 1, \ldots, s.$$

Furthermore, as a consequence of Lemma 2, we obtain the following.

Corollary 2 *If the (Lagrange) function*

$$\sum_{i=1}^{q} \overline{\lambda}_i f_i(x) + \sum_{j=1}^{s} \overline{\mu}_j g(x, \overline{y}^j(x)) \tag{12}$$

is convex on $B(\overline{x}, \varepsilon)$ for some $\varepsilon > 0$ with $B(\overline{x}, \varepsilon) \subset U$ and some $\overline{\mu} \in \mathcal{M}$, then $\overline{x}$ is G-locally properly efficient for the problem

$$\text{“}\min_{x \in U}\text{”} f(x) \ \ s.t. \ g(x, \overline{y}^j(x)) \leq 0, \ j = 1, \ldots, s. \tag{13}$$

However, in general, ESSOSC does not imply that the Lagrange function (12) is convex on a neighbourhood of $\overline{x}$. In the following, we will show how the so-called p-power transformation can overcome this disadvantage.

4.2 The p-Power Transformation

In the remainder of this section we assume for all $x \in cl\ U$ (where cl denotes closure) that

$$f_i(x) > 0,\ i = 1, \ldots, q$$

and that $g(x, y)$ can be written as the difference of a C^2-function and a positive constant as

$$g(x, y) = G(x, y) - r$$

with $G(x, y) > 0$, $r > 0$ for all $(x, y) \in cl\ U \times Y$. This can be assumed without loss of generality since it can be fulfilled by an equivalence transformation, e.g. by exponential transformation $\left(f_i(x) \to e^{f_i(x)}\right)$ or by adding a sufficiently large constant $(f_i(x) \to f_i(x) + c, c > 0)$. Thus, our problem (13) can be written as

$$\text{“}\min_{x \in U}\text{”}(f_1(x), \ldots, f_q(x)) \text{ s.t. } G(x, \overline{y}^j(x)) \leq r,\ j = 1, \ldots, s. \tag{14}$$

For a real number $p > 0$, we define now the so-called *p-power transformation* of (14) by substituting the original functions by their pth power (we write $f_1^p(x)$ for $(f_1(x))^p$)

$$\text{“}\min_{x \in U}\text{”}(f_1^p(x), \ldots, f_q^p(x)) \text{ s.t. } G^p(x, \overline{y}^j(x)) \leq r^p,\ j = 1, \ldots, s. \tag{15}$$

In the subsequent two theorems we will show:

- that the feasible sets of the problems (14) and (15) are identic.
- that efficient and (KT- and G-) properly efficient points for (14) and (15) are closely related.
- that for sufficiently large powers p in (15), the corresponding Lagrange function is convex locally around $\overline{x}$.

For more details on p-power transformations for standard finite and semi-infinite optimization problems, we refer to [8, 18–20, 29].

Theorem 1

(i) *The feasible sets of the problems (14) and (15) are identic.*
(ii) *A point $\widetilde{x} \in U$ is locally efficient for (14) if and only if $\widetilde{x} \in U$ is locally efficient for (15).*
(iii) *A point $\widetilde{x} \in U$ is G-locally properly efficient for (14) if and only if $\widetilde{x} \in U$ is G-locally properly efficient for (15).*
(iv) *A point $\widetilde{x} \in U$ is KT-locally properly efficient for (14) if and only if $\widetilde{x} \in U$ is KT-locally properly efficient for (15).*

Proof The statements (i) and (ii) follow immediately from the positivity of p (and, thus, of $1/p$) and the positivity of the functions f_i, $i = 1, \ldots, q$, $G(\cdot, \overline{y}^j(\cdot))$, $j = 1, \ldots, s$ for all $x \in U$.

(iii) Let $\overline{x} \in U$ be G-locally properly efficient for (14) but not G-locally properly efficient for (15). Then, there exist infinite sequences of positive numbers $\{\varepsilon^\nu\}$, $\{K^\nu\}$ (the index ν is always varying in the whole set $\mathbb{N}$ of natural numbers) with $\varepsilon^\nu \to 0$, $K^\nu \to \infty$, an index sequence $\{i^\nu\} \subset \{1, \ldots, q\}$—let, after shrinking to a subsequence, $i^\nu = 1$, $\forall \nu \in \mathbb{N}$—and a sequence of points $\{x^\nu\} \subset U$ with $x^\nu \to \widetilde{x}$ ($x^\nu \in B(\widetilde{x}, \varepsilon^\nu)$),

$$f_1^p(x^\nu) < f_1^p(\widetilde{x}) \tag{16}$$

such that for each $\nu \in \mathbb{N}$, we have

$$\frac{f_1^p(\widetilde{x}) - f_1^p(x^\nu)}{f_{j^\nu}^p(x^\nu) - f_{j^\nu}^p(\widetilde{x})} > K^\nu \tag{17}$$

for all indices $j^\nu \in \{1, \ldots, q\}$ with

$$f_{j^\nu}^p(x^\nu) > f_{j^\nu}^p(\widetilde{x}). \tag{18}$$

Since $\widetilde{x}$ is G-locally properly efficient for (14), we get, from (16), that for each $\nu \in \mathbb{N}$, there exists an index $l^\nu \in \{1, \ldots, q\}$—let, after shrinking to a subsequence $l^\nu = 2$, $\forall \nu \in \mathbb{N}$—with

$$f_2(x^\nu) > f_2(\widetilde{x}) \tag{19}$$

such that for some $K > 0$, we get

$$\frac{f_1(\widetilde{x}) - f_1(x^\nu)}{f_2(x^\nu) - f_2(\widetilde{x})} \leq K. \tag{20}$$

In particular, (17), (18) and (19) yield

$$\frac{f_1^p(\widetilde{x}) - f_1^p(x^\nu)}{f_2^p(x^\nu) - f_2^p(\widetilde{x})} > K^\nu. \tag{21}$$

By the mean value theorem, there exist

$$\Delta_1^\nu,\ \Delta_2^\nu \in \{\rho\widetilde{x} + (1-\rho)x^\nu \mid \rho \in [0,1]\}$$

such that (21) implies

$$\frac{-pf_1^{p-1}(\Delta_1^\nu)Df_1(\Delta_1^\nu)(x^\nu - \widetilde{x})}{pf_2^{p-1}(\Delta_2^\nu)Df_2(\Delta_2^\nu)(x^\nu - \widetilde{x})} > K^\nu.$$

Since $\Delta_1^\nu \to \widetilde{x}$, $\Delta_2^\nu \to \widetilde{x}$ and $f_1(x) > 0$, $f_2(x) > 0$ for all $x \in cl\ U$, the latter inequality and $K^\nu \to \infty$ yield

$$\frac{\|x^\nu - \widetilde{x}\|}{\|x^\nu - \widetilde{x}\|} \frac{(-Df_1(\Delta_1^\nu)(x^\nu - \widetilde{x}))}{Df_2(\Delta_2^\nu)(x^\nu - \widetilde{x})} \to \infty \text{ as } \nu \to \infty. \tag{22}$$

Perhaps after shrinking to a subsequence, let

$$\frac{x^\nu - \widetilde{x}}{\|x^\nu - \widetilde{x}\|} \to x^0 \text{ as } \nu \to \infty.$$

Then, (22) implies after taking the limit that

$$Df_2(\widetilde{x})x^0 = 0. \tag{23}$$

By using an analogous argument, the mean value theorem and (20) deliver for $\upsilon \to \infty$ that

$$\frac{-Df_1(\widetilde{x})x^0}{Df_2(\widetilde{x})x^0} \leq K$$

which contradicts (23). Therefore, $\widetilde{x} \in U$ is G-locally properly efficient for (15). The other direction is proved analogously by considering the power $1/p$.
(iv) Let $\widetilde{x} \in U$ be KT-locally properly efficient for (14); that is, the following system has no solution $d \in \mathbb{R}^n$ (where, for sake of simplicity, we shorten the notations and substitute $f_i(\widetilde{x})$ by f_i and $G(\widetilde{x}, \overline{y}^j(\widetilde{x}))$ by G^j)

$$\begin{aligned} &Df_i d \leq 0,\ i = 1, \ldots, q, \\ &Df_k d < 0, \text{for some } k \in \{1, \ldots, q\}, \\ &D(G^j - r)d \leq 0,\ j = 1, \ldots, s. \end{aligned}$$

For the derivatives of the functions from (15), we get:

$$\begin{cases} D(f_i^p) = pf_i^{p-1} Df_i,\ i = 1, \ldots, q. \\ D((G^j)^p - r^p) = p(G^j)^{p-1} DG^j,\ j = 1, \ldots, s. \end{cases} \tag{24}$$

Since $pf_i^{p-1} > 0$, $i = 1, \ldots, q$ and $p(G^j)^{p-1} > 0$, $j = 1, \ldots, s$, we obtain from (24) for any $d \in \mathbb{R}^n$ that:

$$\begin{aligned} Df_i d \leq 0 &\longleftrightarrow D(f_i^p)d \leq 0,\ i = 1, \ldots, q, \\ Df_i d < 0 &\longleftrightarrow D(f_i^p)d < 0,\ i = 1, \ldots, q, \\ D(G^j - r)d \leq 0 &\longleftrightarrow D((G^j)^p - r^p)d \leq 0,\ j = 1, \ldots, s. \end{aligned}$$

Consequently, $\widetilde{x} \in U$ is KT-locally properly efficient for (15). This completes the proof of Theorem 1. $\triangle$

We recall the Lagrangian of the problem (14) with the fixed $\overline{\lambda} > 0_q$ as well as the compact set $\mathcal{M}$ representing the solution set of (11) (here (10) comes into play). In the next theorem, we will show that for a sufficiently large power p, the Lagrangian of the problem (15) that corresponds to $\overline{\lambda}$ is convex on a neighbourhood of $\overline{x}$. This Lagrangian is given for $x \in U$ as follows:

$$L_p(x, \overline{\beta}, \gamma) = \sum_{i=1}^{q} \overline{\beta}_i f_i^p(x) + \sum_{j=1}^{s} \gamma_j (G^p(x, \overline{y}^j(x)) - r^p),$$

where $\gamma \in \mathbb{R}^s$ and $\overline{\beta} \in \mathbb{R}^q$ are fixed with

$$\overline{\beta}_i = \frac{\overline{\lambda}_i}{f_i^{p-1}(\overline{x})},\ i = 1, \ldots, q. \tag{25}$$

Note that it is $\overline{\beta} > 0_q$ and that $\overline{\beta}$ depends on p. By (11), a short calculation shows that the compact set of solutions $\gamma \in \mathbb{R}^s$ satisfying

$$D_x L_p(\overline{x}, \overline{\beta}, \gamma) = 0,\ \gamma \geq 0_s,$$

is

$$\Gamma_p = \left\{ \gamma \in \mathbb{R}^s \mid \gamma_j = \frac{\mu_j}{G^{p-1}(\overline{x}, \overline{y}^j(\overline{x}))},\ j = 1, \ldots, s,\ \mu \in \mathcal{M} \right\}. \tag{26}$$

Now, we will show that $D_x^2 L_p(\overline{x}, \overline{\beta}, \gamma)$ is positive definite for all sufficiently large chosen powers p and all $\gamma \in \Gamma_p$ where $\overline{\beta} > 0_q$ is chosen as in (25).

Theorem 2 *There exists a power $\overline{p} > 0$ such that the Hessian $D_x^2 L_p(\overline{x}, \overline{\beta}, \gamma)$ is positive definite for all $\gamma \in \Gamma_p$ whenever $p > \overline{p}$.*

Proof Throughout this proof we substitute $f_i(\overline{x})$ by $f_i, i = 1, \ldots, q$ and $G(\overline{x}, \overline{y}^j(\overline{x}))$ by G^j, $j = 1, \ldots, s$. We obtain for f_i (and analogously for G^j):

$$\begin{aligned} Df_i^p &= p f_i^{p-1} Df_i \text{ and} \\ D^2 f_i^p &= p(p-1) f_i^{p-2} Df_i^\top Df_i + p f_i^{p-1} D^2 f_i. \end{aligned}$$

Then, the Hessian $D_x^2 L_p(\overline{x}, \overline{\beta}, \gamma)$ with $\overline{\beta} > 0_q$ satisfying (25) and $\gamma \in \Gamma_p$ is

$$\begin{aligned} D_x^2 L_p(\overline{x}, \overline{\beta}, \gamma) &= \sum_{i=1}^q \overline{\beta}_i D^2 f_i^p + \sum_{j=1}^s \gamma_j D^2((G^j)^p - r^p) \\ &= \sum_{i=1}^q \left(p(p-1) \frac{\overline{\lambda}_i}{f_i} Df_i^\top Df_i + p\overline{\lambda}_i D^2 f_i \right) \\ &\quad + \sum_{j=1}^s \left(p(p-1) \frac{\mu_j}{G^j} (DG^j)^\top DG^j + p\mu_j D^2 G^j \right) \qquad (27) \\ &= p D_x^2 L(\overline{x}, \overline{\lambda}, \mu) + p(p-1) \left(\sum_{i=1}^q \frac{\overline{\lambda}_i}{f_i} Df_i^\top Df_i \right. \\ &\quad \left. + \sum_{j=1}^s \frac{\mu_j}{G^j} (DG^j)^\top DG^j \right) \end{aligned}$$

(γ and μ are related according to (26)).

Now assume that the statement of Theorem 2 would not be true. Then, there exist sequences $\{p^\nu\} \subset \mathbb{R}$, $\{\gamma^\nu\} \subset \Gamma_{p^\nu}$, $\{w^\nu\} \subset \mathbb{R}^n$ such that:

- $p^\nu > 0$, $p^\nu \to \infty$.
- $\gamma_j^\nu = \frac{\mu_j^\nu}{(G^j)^{p^\nu - 1}}$ (as in (26)), $\mu^\nu \to \overline{\mu}$ ($\mathcal{M}$ is a compact set).
- $\|w^\nu\| = 1$, $w^\nu \to \overline{w}$ and

$$(w^\nu)^\top D_x^2 L_{p^\nu}(\overline{x}, \overline{\beta}, \gamma^\nu) w^\nu \leq 0. \tag{28}$$

After dividing (27) by p, we get from (28) that

$$(w^\nu)^\top D_x^2 L(\overline{x}, \overline{\lambda}, \mu^\nu) w^\nu + (p^\nu - 1)(w^\nu)^\top \left[\sum_{i=1}^{q} \frac{\overline{\lambda}_i}{f_i} Df_i^\top Df_i + \sum_{j=1}^{s} \frac{\mu_j^\nu}{G^j} (DG^j)^\top DG^j \right] w^\nu \leq 0. \tag{29}$$

Dividing (29) by $(p^\nu - 1)$ ($p^\nu > 1$ for ν sufficiently large) and taking the limit yields

$$\overline{w}^\top \left[\sum_{i=1}^{q} \frac{\overline{\lambda}_i}{f_i} Df_i^\top Df_i + \sum_{j=1}^{s} \frac{\overline{\mu}_j}{G^j} (DG^j)^\top DG^j \right] \overline{w} \leq 0.$$

Since $\frac{\overline{\lambda}_i}{f_i} > 0$, $\frac{\overline{\mu}_j}{G^j} \geq 0$, $i = 1, \ldots, q$, $j = 1, \ldots, s$ we obtain $\overline{w} \in T(\overline{\mu})$ and, by ESSOSC,

$$\overline{w}^\top D_x^2 L(\overline{x}, \overline{\lambda}, \overline{\mu}) \overline{w} > 0$$

and, therefore for sufficiently large ν, we get

$$(w^\nu)^\top D_x^2 L(\overline{x}, \overline{\lambda}, \mu^\nu) w^\nu > 0.$$

The latter inequality and

$$(p^\nu - 1)(w^\nu)^\top \left[\sum_{i=1}^{q} \frac{\overline{\lambda}_i}{f_i} Df_i^\top Df_i + \sum_{j=1}^{s} \frac{\mu_j^\nu}{G^j} (DG^j)^\top DG^j \right] w^\nu \geq 0$$

provide for sufficiently large ν a contradiction with (29). This completes the proof. $\triangle$

As a consequence of the latter theorem, we get the following two corollaries.

Corollary 3 *There exists $\overline{p} > 0$ such that $\overline{x}$ is G-locally properly efficient and KT-locally properly efficient for (15) for all $p \geq \overline{p}$.*

Proof According to Theorem 2, there exists $\overline{p} > 0$ such that the Hessian $D_x^2 L_p(\overline{x}, \overline{\beta}, \gamma)$ of (15) is positive definite for all $\gamma \in \Gamma_p$ whenever $p > \overline{p}$. Let $\widetilde{p} > \overline{p}$ be arbitrarily chosen and fixed. By continuity, for a fixed $\overline{\gamma} \in \Gamma_{\widetilde{p}}$, there exists $\overline{\varepsilon} > 0$ such that $D_x^2 L_{\widetilde{p}}(\overline{x}, \overline{\beta}, \overline{\gamma})$ is positive definite for all $x \in B(\overline{x}, \overline{\varepsilon})$ and, therefore,

$L_{\widetilde{p}}(\cdot, \overline{\beta}, \overline{\gamma})$ is convex on $B(\overline{x}, \overline{\varepsilon})$. Then, Lemma 2 yields that $\overline{x}$ is G-locally properly efficient for (15) (with $p = \widetilde{p}$) and, by EMFCQ and Lemma 1 (i), (ii), $\overline{x}$ is KT-locally properly efficient for (15) (with $p = \widetilde{p}$). △

The next corollary is an immediate consequence of Theorem 1 and Corollary 3 and it uses the properties of p-power transformation only in its proof. Since it has its own significance (as explained in the subsequent remark), we recall all assumptions on the point $\overline{x}$ in this corollary.

Corollary 4 *Let $\overline{x} \in M$ and assume that the following assumptions are fulfilled at $\overline{x}$ (as described in more details in the beginning of this section):*

- *RA holds at $\overline{x}$.*
- *EMFCQ holds at $\overline{x}$.*
- *The condition (11) holds with corresponding multipliers, and*
- *ESSOSC holds at $\overline{x}$.*

Then, the point $\overline{x}$ is G-locally properly efficient and KT-locally properly efficient for (14).

Proof This follows directly from Theorem 1 (iii), (iv) and Corollary 3. △

Remark 1 The important message of the latter corollary is that the property that $\overline{x}$ is G-locally properly efficient and KT-locally properly efficient for (14) holds *without assuming that the Lagrange function related to (14) is locally convex.* In other words, if a feasible point $\overline{x} \in M$ fulfills the assumptions defined in Sect. 4.1, then $\overline{x}$ is G-locally properly efficient and KT-locally properly efficient for (14). In particular, under the assumptions of this corollary and having in mind that under RA a semi-infinite problem can be described locally as a finite problem, the basic result for finite optimization problems as stated in Lemma 2 holds *without assuming that its Lagrange function is locally convex.*

5 An Illustrating Example

In this section, we present an illustrating example of the p-power transformation of MOSIP locally around a G-locally properly efficient feasible point.

Example 2 *Let $n = 2, q = 2, m = 1$ and*

$$f_1(x) = \tfrac{1}{90}\left[\left(x_1 - 2x_2 + 5x_1^2 - x_1^2x_2 - 13\right)^2 + \left(x_1 - 14x_2 + x_2^2 + x_2^3 - 29\right)^2\right],$$
$$f_2(x) = 4\,(x_1 - 2)^2 - 16\,(x_1 - 2)\,(x_2 - 2) + 11,$$
$$g(x, y) = \left(1 - x_1^2y^2\right)^2 - x_1y^2 - x_2^2 + x_2,$$

$Y = [0, 1]$ *and let the feasible set be given as*

$$M = \{x \in \mathbb{R}^2 \mid g(x, y) \leq 0,\ y \in Y,\ -1.1 \leq x_1 \leq -0.5,\ 1.6 \leq x_2 \leq 1.7\}.$$

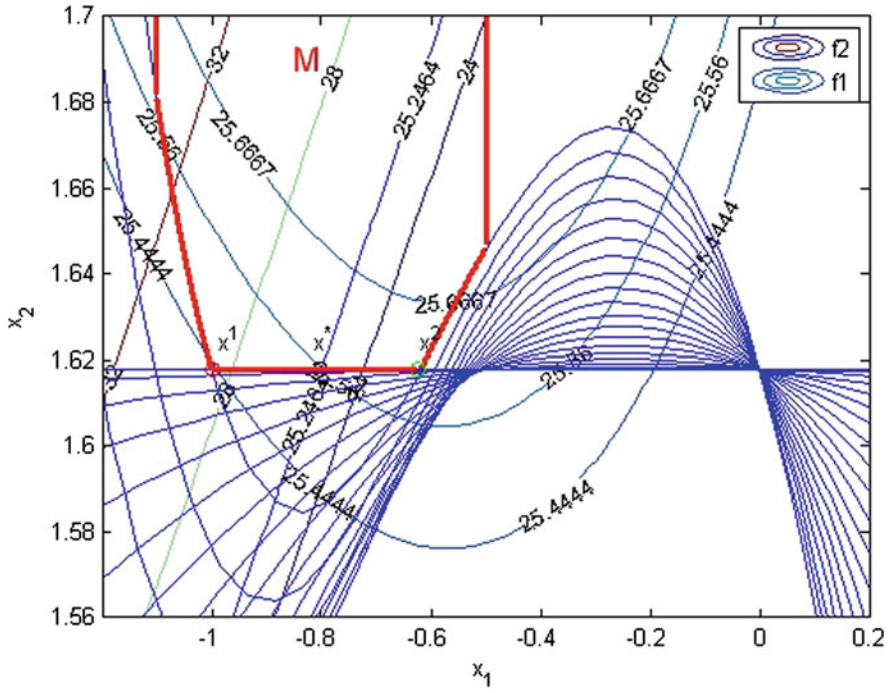


Fig. 1 Level lines of f_1, f_2, g, the feasible set M and the efficient point x^*

For the sake of illustration in Fig. 1, we added box constraints for x_1 and x_2 to the description of M. The functions f_1 and g are taken from [23]. In Fig. 1, level lines of f_1 and f_2 are shown. Furthermore, it contains curves corresponding to $g(x, y)$ for several values of $y \in Y$. The feasible point $x^* = (-0.8, 1.618)^\top$ is efficient.

Since

$$\max \left\{ \frac{f_1(x^*) - f_1(x)}{f_2(x) - f_2(x^*)} \middle| \, f_1(x) < f_1(x^*),\ x_1 \geq -0.9,\ x \in M \right\} = 516$$

and

$$\max \left\{ \frac{f_2(x^*) - f_2(x)}{f_1(x) - f_1(x^*)} \middle| \, f_2(x) < f_2(x^*),\ x_1 \leq -0.7,\ x \in M \right\} = 4887,$$

the point x^* is G-locally properly efficient. The following can easily be obtained:

- $Y_0(x^*) = \{0\}$.
- $D_x g(x^*, 0) = (3.2, -2.236)$.
- $\overline{y}^1(x) = 0$ for x near x^*.
- RA holds at x^*.

In the remainder of this section, we will use the following notation from multiobjective optimization: If x^* is a (G-locally) properly efficient solution, then $f(x^*)$ is

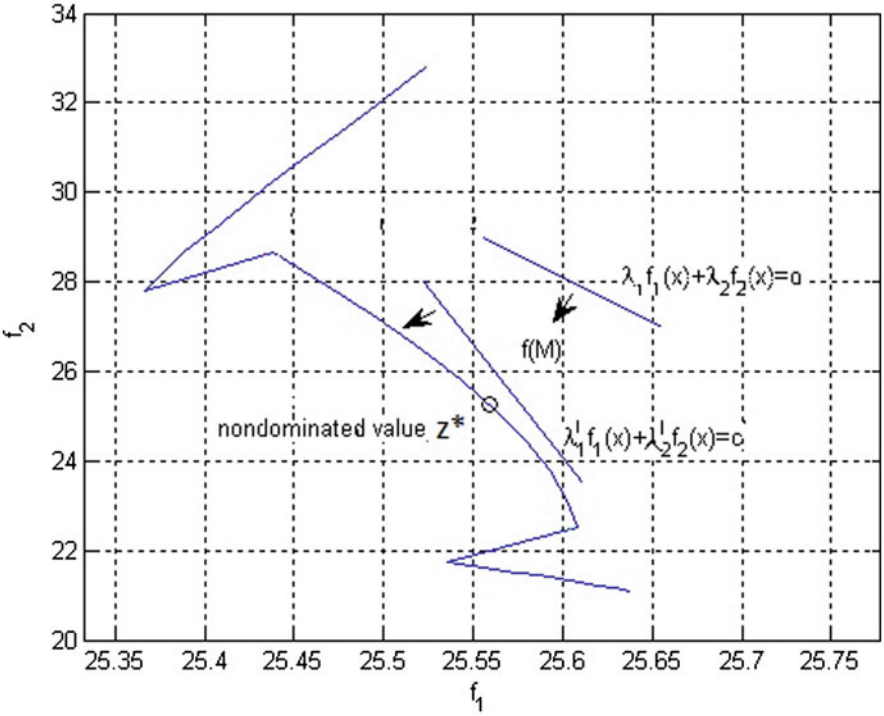


Fig. 2 $z^* = f(x^*)$ is properly nondominated in the objective space $f(M)$

properly nondominated in the *objective space* $f(M) = \{f(x) \mid x \in M\}$. For the Lagrangian

$$L(x, \lambda, \mu) = \lambda_1 f_1(x) + \lambda_2 f_2(x) + \mu g(x, \overline{y}^1(x)),$$

we obtain for any choice $\overline{\lambda} > 0_2$, $\overline{\mu} \geq 0$ satisfying (11) (with $\overline{x} = x^*$) that the Hessian $D_x^2 L(x^*, \overline{\lambda}, \overline{\mu})$ has at least one negative eigenvalue. Figure 2 illustrates this situation geometrically in the objective space $f(M)$. There we have sketched for two different pairs of positive weights, a corresponding level line of $\lambda_1 f_1(x) + \lambda_2 f_2(x)$ (and $\lambda_1' f_1(x) + \lambda_2' f_2(x)$). Although $z^* = f(x^*)$ is properly nondominated in $f(M)$, none of its preimages can be a minimizer of

$$\min \lambda_1 f_1(x) + \lambda_2 f_2(x) \text{ s.t. } x \in M$$

for any choice $\lambda_1 > 0$, $\lambda_2 > 0$.

For applying the p-power transformation, we add a constant to $g(x, y)$ and obtain as new constraint

$$G(x, y) = g(x, y) + 7, \; r = 7.$$

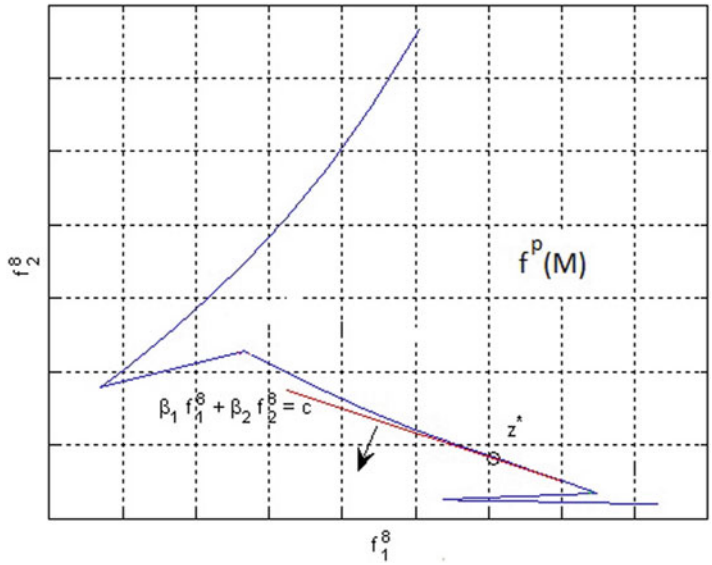


Fig. 3 The objective space after p-power transformation

Note that $f_1(x)$ and $f_2(x)$ are positive for all $x \in M$. Then, the p-power transformation of the original problem becomes

$$\text{“min”} \ \left(f_1^p(x), f_2^p(x)\right) \ \text{s.t.} \ \left(g(x, \overline{y}^1(x)) + 7\right)^p \leq 7^p$$

and it can easily be seen that the corresponding Hessian of the Lagrangian is positive definite for $p > 7.7$ and an appropriate choice of $\overline{\lambda} > 0_2$ (respectively, $\beta > 0_2$ according to (25)). Figure 3 (appropriately scaled) illustrates for $p = 8$ the existence of appropriate $\beta_1 > 0$, $\beta_2 > 0$ such that x^* is a local minimizer of the problem

$$\min \beta_1 f_1^p(x) + \beta_2 f_2^p(x) \ \text{s.t.} \ x \in M.$$

In particular, we get $\beta_1 = 1$ and $\beta_2 = 0.03$.

6 Conclusions

In this chapter, we applied a convexification procedure, called p-power transformation, to the setting of a nonconvex MOSIP. Under the assumptions presented in Sect. 4.1, we have seen that for sufficiently large powers p, the Lagrangian of a transformed weighted sum optimization problem becomes convex locally around the efficient point under consideration. Since convexity of the Lagrangian is an essential requirement for the application of duality theory and corresponding solution methods, the results in this chapter allow the use of these solution methods to a

broader class of optimization problems. We also have shown that the property of being a locally properly efficient point (in the sense of Geoffrion or in the sense of Kuhn and Tucker) is invariant under this convexification procedure.

Acknowledgement The authors thank both referees for their careful reading and substantial critical remarks which improved essentially the quality of this chapter.

References

1. Bazaraa, M.S., Sherali, H.D., Shetty, C.M.: Nonlinear Programming: Theory and Algorithms. Wiley, Chichester (2006)
2. Censor, Y.: Necessary conditions for Pareto optimality in simultaneous Chebyshev best approximation. J. Approx. Theory **27**, 127–134 (2006)
3. Ehrgott, M.: Multicriteria optimization. Springer, Berlin (2005)
4. Gauvin, J.: A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming. Math. Progr. **12**, 136–138 (1977)
5. Geoffrion, A.: Proper efficiency and the theory of vector maximization. J. Math. Anal. Appl. **22**, 618–630 (1968)
6. Goberna, M.A., López, M.A. (eds.): Semi-Infinite Programming—Recent Advances. Kluwer, Boston (2001)
7. Guerra-Vázquez, F., Rückmann, J.-J.: On proper efficiency in multiobjective semi-infinite optimization. In: Xu, H. et al. (eds.) Advances in Optimization and Control with Applications. Springer, Berlin
8. Guerra-Vázquez, F., Rückmann, J.-J., Werner, R.: On saddle points in nonconvex semi-infinite programming. J. Global Optim. **54**(3), 433–447 (2012)
9. Hettich, R., Jongen, H.Th.: Semi-infinite programming: conditions of optimality and applications. In: Stoer, J. (ed.) Optimization Techniques, vol. 2, pp. 1–11. Springer, Berlin (1978)
10. Hettich, R., Kortanek, K.O.: Semi-infinite programming: theory, methods and applications. SIAM Rev. **35**, 380–429 (1993)
11. Hettich, R., Still, G.: Second order optimality conditions for generalized semi-infinite programming problems. Optimization **34**, 195–211 (1995)
12. Jongen, H.Th., Wetterling, W.W.E., Zwier, G.: On sufficient conditions for local optimality in semi-infinite optimization. Optimization **18**, 165–178 (1987)
13. Jongen, H.Th., Twilt, K., Weber, G.W.: Semi-infinite optimization: structure and stability of the feasible set. J. Optim. Theory Appl. **72**, 529–552 (1992)
14. Klatte, D.: Stability of stationary solutions in semi-infinite optimization via the reduction approach. In: Oettli, W., Pallaschke, D. (eds.) Advances in Optimization, pp. 155–170. Springer, Berlin (1992)
15. Klatte, D.: Stable local minimizers in semi-infinite optimization: regularity and second-order conditions. J. Comput. Appl. Math. **56**, 137–157 (1994)
16. Klatte, D.: On regularity and stability in semi-infinite optimization. Set-Valued Anal. **3**, 101–111 (1995)
17. Kuhn, H., Tucker, A.: Nonlinear programming. In: Newman, J. (ed.) Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, pp. 481–492. University of California Press, Berkeley, (1951)
18. Li, D.: Zero duality gap for a class of nonconvex optimization problems. J. Optim. Theory Appl. **85**, 309–324 (1995)
19. Li, D.: Saddle-point generation in nonlinear nonconvex optimization. Nonlinear Anal. **30**, 4339–4344 (1997)

20. Li, D., Sun, X.L.: Local convexification of the Lagrangian function in nonconvex optimization. J. Optim. Theory Appl. **104**, 109–120 (2000)
21. Mangasarian, O.L.: Nonlinear Programming. McGraw-Hill, New York (1969). (Reprinted by SIAM Publications, 1995)
22. Mangasarian, O.L., Fromovitz, S.: The Fritz John necessary optimality conditions in the presence of equality and inequality constraints. J. Math. Anal. Appl. **17**, 37–47 (1967)
23. Mitsos, A.: A test set of semi-infinite programs—MIT. web.mit.edu/mitsos/www/pubs/siptestset.pdf (2009)
24. Polak, E.: Optimization. Algorithms and Consistent Approximations. Springer, New York (1997)
25. Reemtsen, R., Rückmann, J.-J. (eds.): Semi-Infinite Programming. Kluwer, Boston (1998)
26. Rückmann, J.-J.: On existence and uniqueness of stationary points in semi-infinite optimization. Math. Progr. **86**, 387–415 (1999)
27. Stein, O.: Bi-Level Strategies in Semi-Infinite Programming. Kluwer, Boston (2003)
28. Wetterling, W.W.E.: Definitheitsbedingungen für relative Extrema bei Optimierungs- und Approximationsaufgaben. Numer. Math. **15**, 122–136 (1970)
29. Xu, Z.K.: Local saddle points and convexification for nonconvex optimization problems. J. Optim. Theory Appl. **94**, 739–746 (1997)

Qualitative Analysis of Climate Seasonality Effects in a Model of National Electricity Market

Johnny Valencia, Gerard Olivar, Carlos Jaime Franco and Isaac Dyner

Abstract In the following chapter, we present a model for the supply and demand of electricity in a domestic market based on system dynamics. Additionally, the model shows piecewise smooth differential equations arising from the diagram of flows and levels, using dynamical systems theory for the study of stability of the equilibrium points that have such a system. We also present simulations, nonlinear numerical analysis, and qualitative analysis to the system of differential equations obtained, which is characterized by dynamics not smooth, due to the way decisions are made in that market. Using the software package Vensim and event-based scheme implementation in Matlab are verified as different saturation phenomena, oscillations, fixed points, among others, and the relationship between leverage points and stability of equilibrium points. Finally, we conclude the effects of climate seasonality in the market. Furthermore, we show that the system becomes periodically forced due to the external variable.

Keywords Electricity market · Applied mathematics · Nonsmooth dynamical system · Numerical analysis · Nonlinear system · Stability analysis · Simulation

J. Valencia (✉) · C. J. Franco
Department of Computer Science and Decision, School of Mines,
Universidad Nacional de Colombia, Medellín, Colombia
e-mail: jovalenciacal@unal.edu.co

C. J. Franco
e-mail: cjfranco@unal.edu.co

G. Olivar
Department of Electric, Electronic and Computer Science,
Universidad Nacional de Colombia, Manizales, Colombia
e-mail: golivart@unal.edu.co

I. Dyner
Universidad Nacional de Colombia, Bogotá, Colombia
Universidad Jorge Tadeo Lozano, Bogotá, Colombia
e-mail: idyner@unal.edu.co

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_24

1 Introduction

Many research efforts have, focused on the analysis and classification of modeling and simulation schema [3, 16, 20]. Some models, might be more appropriate than others for decision-making purposes within a particular framework in society [12]. In the case of this chapter, an electricity market is studied based on system dynamics (SD) [10, 11, 15]. We present the role of mathematical analysis that arises in dynamic systems.

The concept of quality, has had different connotations in the field. One would expect that the information contained in the causal diagrams has this nature [3]. The qualitative term refers to various phenomena which the system exhibits, i.e., its dynamic trends of flow over time, for example, when the system-level variable increases, decreases, or has oscillations reaching a point of equilibrium. It is determined by a parameter (leverage points) or a set of decision rules [9]. SD is structured so that it has a part of mathematical modeling and numerical simulation methods. A model based on SD, is a mathematical object [4], SD is a mathematical technique, requires a mathematical analysis of its dynamics. J. Redondo [19] shows that SD model may be often represented as a model by a piecewise smooth system. The Colombian electricity market explores the model equations. We reinforce this approach. It is potential to see other fields like social systems, such as hybrid and piecewise smooth dynamical systems are increasingly used [13, 17], economic systems, social systems, more generally all devices and systems whose dynamics is affected by the occurrence of discrete events on a microscopic time scale [2, 5].

It has been noted that the piecewise smooth systems can exhibit a wide range of nonlinear phenomena, including bifurcations and chaos. Under parameter variation, classical bifurcations can occur, such as bending, Hopf bifurcation, among others. Besides, discontinuity induces bifurcations [6]. When this occurs, the system may show a dramatic transition attractor to another, often including sudden transitions, experimentally demonstrated only for physical systems [7, 8]. The literature has not reported analysis for an electricity market model such as the one discussed in this chapter. Bifurcation analysis can be used to get models of real physical systems with specific characteristics and the best they can offer other designs. Besides, it is possible to implement innovative control strategies for defining paths for certain values of parameters in a dynamic system [1, 21].

Thus, it is possible to use nonlinear modeling schemes, piecewise smooth or batch systems, a high degree representing phenomena present in real models. Consequently, SD allows high level of aggregation substantially studying complex systems, which ultimately are sets of piecewise smooth systems, carefully interconnected by functions or mathematical laws. Models based on SD translate a type of mental model in the language of dynamical systems [3].

This chapter is developed as follows: in section 2, we show the Forrester diagram and the equations. We show in subsection 3 the numerical results. In subsection 4, we show the stability analysis. In section 3, we add an external variable that disturbs the system similar to the seasonal climate. We perform numerical analysis again with the effects of weather. Finally, we conclude about the dynamics of the market.

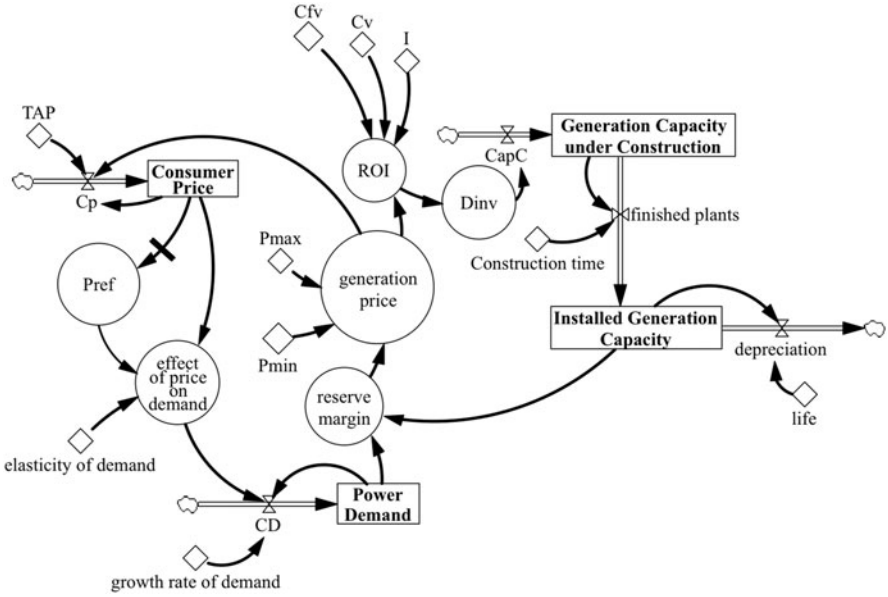


Fig. 1 Stock and flows diagram for a national electricity market

2 National Electricity Market Model

2.1 SD Model

The complementarity between SD and mathematical analysis, with the help of the modern theory of nonlinear dynamical systems allows establishing the qualitative behavior. However, it is difficult to distinguish between qualitative and quantitative. Therefore, starting from a mindset that is basically the root of the SD, it is clearly possible to achieve quantitative models describing the system. In Fig. 1, the electrical market model is shown. Figure 1 shows the main elements of the SD model, using stocks and flows as this is the standard in the SD literature, where the level variables are given by the *power demand*, the *consumer price*, *generation capacity under construction,* and *installed generation capacity*. This can be easily translated into a set of differential equations.

2.2 Mathematical Model

For the study of piecewise smooth dynamical systems, different analytical, there are numerical and experimental tools [21]. Hence, the verification and comparison of results between each of the strategies are necessary.

Table 1 Variables and parameters description

	Equivalence	Units	Name
x_1	GI	MW	Installed generation capacity
x_2	GC	MW	Generation capacity under construction
x_3	D	MW	Power demand
x_4	P_c	$\frac{\$}{\text{KWh}}$	Consumer price
r	frac1VU	years	Generation plant life
q	$\frac{1}{\text{TEC}}$	years	Construction plant time
s	$\frac{1}{\text{TAP}}$	years	Price delay
a	$\Delta P_{\max}$	$\frac{\$}{\text{KWh}}$	Maximum generation price
P_{ref}		$\frac{\$}{\text{KWh}}$	Reference consumer price
Cfv		$\frac{\$}{\text{KWh}}$	Fixed cost variability
Cv		$\frac{\$}{\text{KWh}}$	Fixed cost
I		$\frac{\$}{\text{KWh}}$	Incentives
ε		%	Elasticity of demand
b	$P_{\min}$	$\frac{\$}{\text{KWh}}$	Minimum generation price
k		$\frac{1}{\text{year}}$	Grow rate of demand

It is very common to use continuous models to describe discontinuous dynamical systems. However, such continuous models cannot provide adequate predictions of discontinuous dynamics. To better understand discontinuous systems, you should be aware that discontinuous models provide adequate and actual prediction. Therefore, consider that a global system is discontinuous having several continuous subsystems in different domains. Each continuous subsystem has different dynamic properties, i.e., rules for each adjacent continuous evolution subsystem. The laws of transition between the borders should be studied in more detail. Such variations can lead to dramatic changes in the dynamic behavior of the system. It has been shown, for example, that the transition to chaos is often due to bifurcations induced by discontinuities (abrupt changes) on the borders with which the system is modeled.

According to the data of Table 1 and Fig. 1, by algebraic manipulation it can reach the system of equations shown in (1):

$$\begin{aligned} \dot{x_1} &= -rx_1 + qx_2 \\ \dot{x_2} &= -qx_2 + B \\ \dot{x_3} &= kAx_3 \\ \dot{x_4} &= s\left(\left(\frac{a}{1+e^{MR}} + b\right) - x_4\right), \end{aligned} \tag{1}$$

where B is the construction capacity and it is given by (2).

$$B = \begin{cases} 0 & \text{if} \quad \text{Dinv} \leq 0 \\ 500 & \text{if} \quad 0 < \text{Dinv} \leq 0.1 \\ 2500 & \text{if} \quad \text{Dinv} > 0.1, \end{cases} \tag{2}$$

the investment decision Dinv is defined as $\text{Dinv} = \max\{0, \text{ROI}\}$ and the return over investment ROI is given by the nonlinear function shown in (3).

$$\text{ROI} = \frac{\left(\left(\frac{a}{1+e^{MR}}\right) - Cv + I\right)}{Cfv} \times 100. \tag{3}$$

MR is known as a reserve margin given by (4).

$$\text{MR} = \begin{cases} 10 & \text{if} \quad x_3 = 0 \\ \frac{x_1 - x_3}{x_3} & \text{if} \quad x_3 \neq 0. \end{cases} \tag{4}$$

Finally, the price effect on demand is shown in (5).

$$A = \begin{cases} 1 & \text{if} \quad P = 0 \\ \left(\frac{x_1 - x_3}{x_3}\right)^{\varepsilon} & \text{if} \quad P \neq 0. \end{cases} \tag{5}$$

A given piecewise smooth system can be classified according to its degree of discontinuity through the set of discontinuities, which divide a boundary of another. That is, you may have dashed paths through staple varieties as in the case of systems with impacts, but discontinuous or continuous state vector fields as in the case of so-called Filippov systems. In addition, a system can be continuous piecewise smooth, in the sense that its continuous states and vector fields through the borders of the state space, but with a possible discontinuous Jacobian [5, 6].

According to the classification mentioned above, we note that the model has two types of discontinuities. Equations 4 and 5 show a piecewise smooth continuous system, while Eq. 2 is a Fillipov discontinuity, because vector fields associated with each of the conditions are different.

2.3 Stability Analysis

From a mathematical point of view, it is necessary to analyze the possible scenarios that the system of equations presented. This allows you to set the operating range and robustness of the model, this being a fundamental part of the validation.

The system of equations (1) allows us to analyze the following cases:

2.3.1 Case 1: if $P = 0$ and $x_3 = 0$

This case only allows us to assess the robustness of the model, since the reference price and demand do not reach zero in the real market. Then, we find the equilibrium point associated with these conditions and perform stability analysis as follows:

$$x^* = \left(\frac{B}{r}, \frac{B}{q}, 0, \frac{a}{1+e^{10}} + b\right). \tag{6}$$

After linearizing the system, we find the eigenvalues that determine the stability of the equilibrium point as:

$$\lambda_1 = k$$
$$\lambda_2 = -q$$
$$\lambda_3 = -r$$
$$\lambda_4 = -s$$

2.3.2 Case 2: if $P = 0$ and $x_3 \neq 0$

There is no explicit solution that allows us to find a real equilibrium point.

2.3.3 Case 3: if $P \neq 0$ and $x_3 = 0$

In this case, the reference price is different and greater than zero, while the power demand is zero. Then, we can consider the hypothetical case that there is no demand at any given time in the market. This ensures that the model and the system of equations representing much of the market dynamics:

$$x^* = \left(\frac{B}{r}, \frac{B}{q}, 0, \frac{a}{1+e^{10}} + b\right). \tag{7}$$

with the eigenvalues:

$$\lambda_1 = k\left(\frac{b + \frac{a}{1+e^{10}}}{P}\right)^{\varepsilon}$$

$$\lambda_2 = -q$$
$$\lambda_3 = -r$$
$$\lambda_4 = -s$$

2.3.4 Case 4: if $P \neq 0$ and $x_3 \neq 0$

Finally, this case represents the real market. Considering the above cases, we can have a recognition of present or not on the market dynamics. We find the equilibrium point and similarly in previous cases analytically determine the stability of the point:

$$x^* = \left(\frac{B}{r}, \frac{B}{q}, 0, b\right), \tag{8}$$

with the eigenvalues:

$$\lambda_1 = k\left(\frac{b}{P}\right)^{\varepsilon}$$

$$\lambda_2 = -q$$
$$\lambda_3 = -r$$
$$\lambda_4 = -s$$

Now let us look at the results, it looks like for each of the cases the eigenvalue associated with the power demand, λ_1, changes, and depends on the growth rate of demand.

From the classical theory of dynamical systems, it can be said that if all the eigenvalues are negative, the system is stable, but on the contrary, only one of the eigenvalues is positive the system is unstable [14]. Note that for the cases studied above, the stability depends on the growth rate of demand k. The first results are shown in Fig. 2, in which the behavior of each of the level variables seen with $k = 0.03$. The system is oscillatory. On the other hand, if $k = 0.08$ demand is growing fast, and the market is unstable. See Fig. 3. Finally, we show in Fig. 4, what happens when $k < 0$, the system is stable and tends to the fixed point.

3 Climate Seasonality Effects

Due to the Colombian topography and abundant water resources, hydropower has been the most attractive way to meet the demand for electricity in the country. Hence, the generation in Colombian market is characterized by being highly conditional on water resources, being referred to by some authors as dependent on hydroelectricity or a hydro-dominated market [18].

For this case, we perturb the market affecting the price of generation. We model the seasonality of climate as a parameter that emulates sine periods of rain and sunny periods. Periods of sun and rain affect the price of generation in the market; see Fig. 5.

As shown in Fig. 6, it is possible to see the effects of seasonality on the market. The generation price perturbed by an external variable tends to behave according to

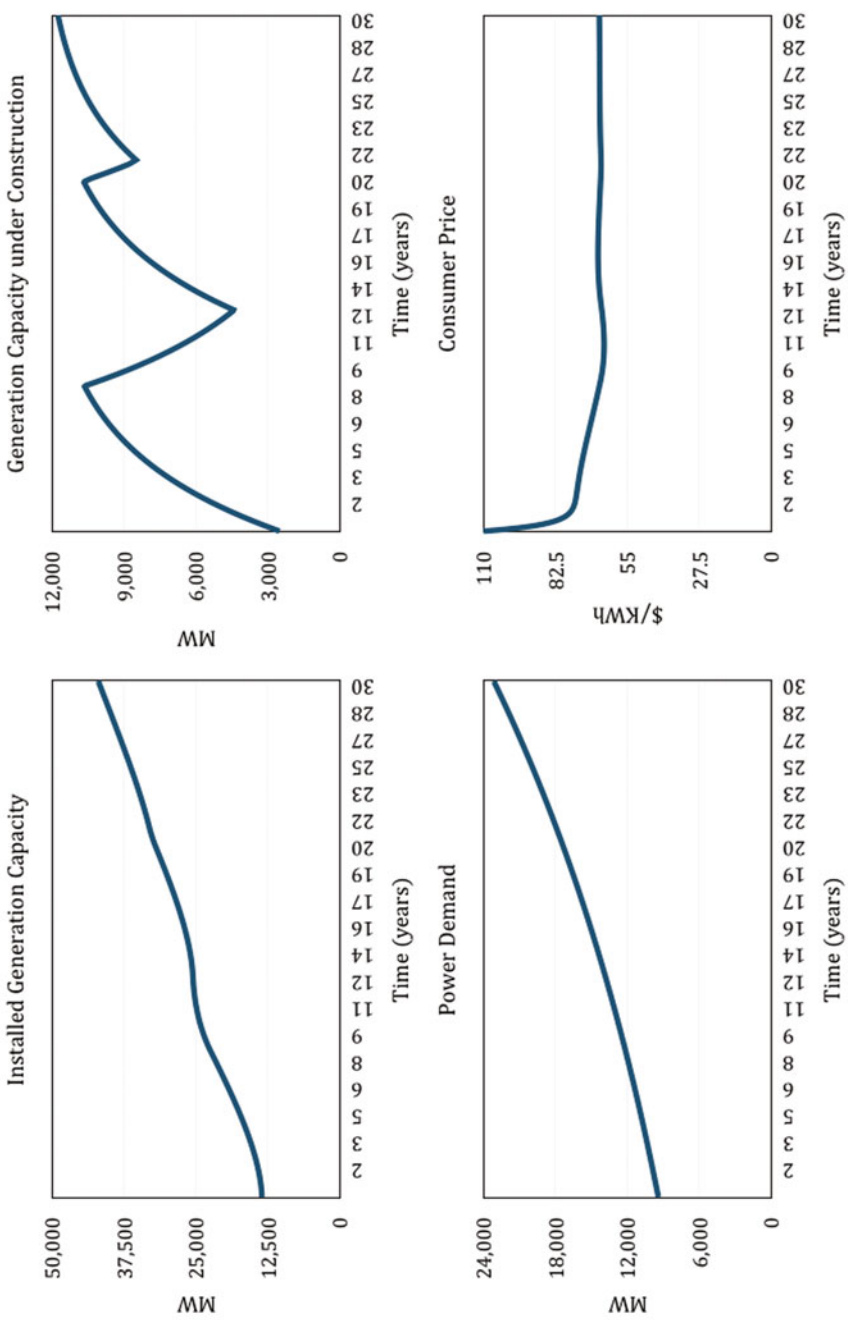


Fig. 2 Simulations results for a grow rate demand $k = 0.03$

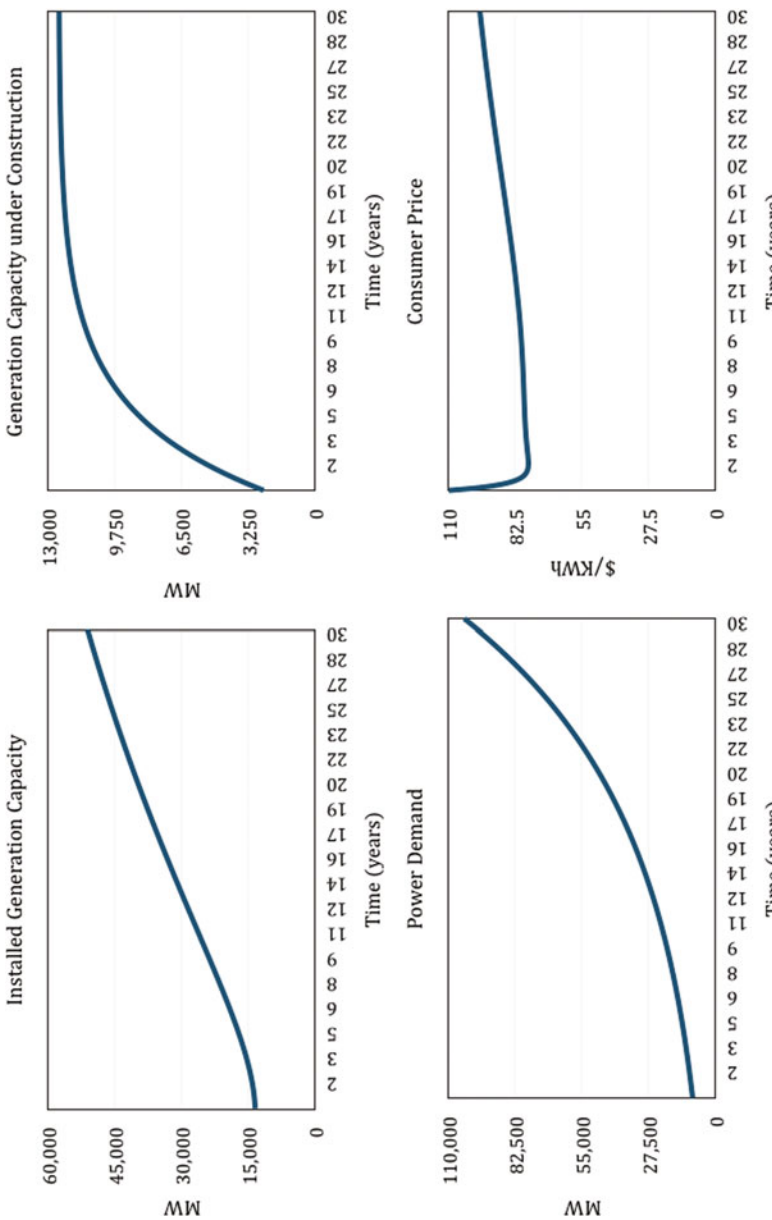


Fig. 3 Simulations results for a grow rate demand $k = 0.008$

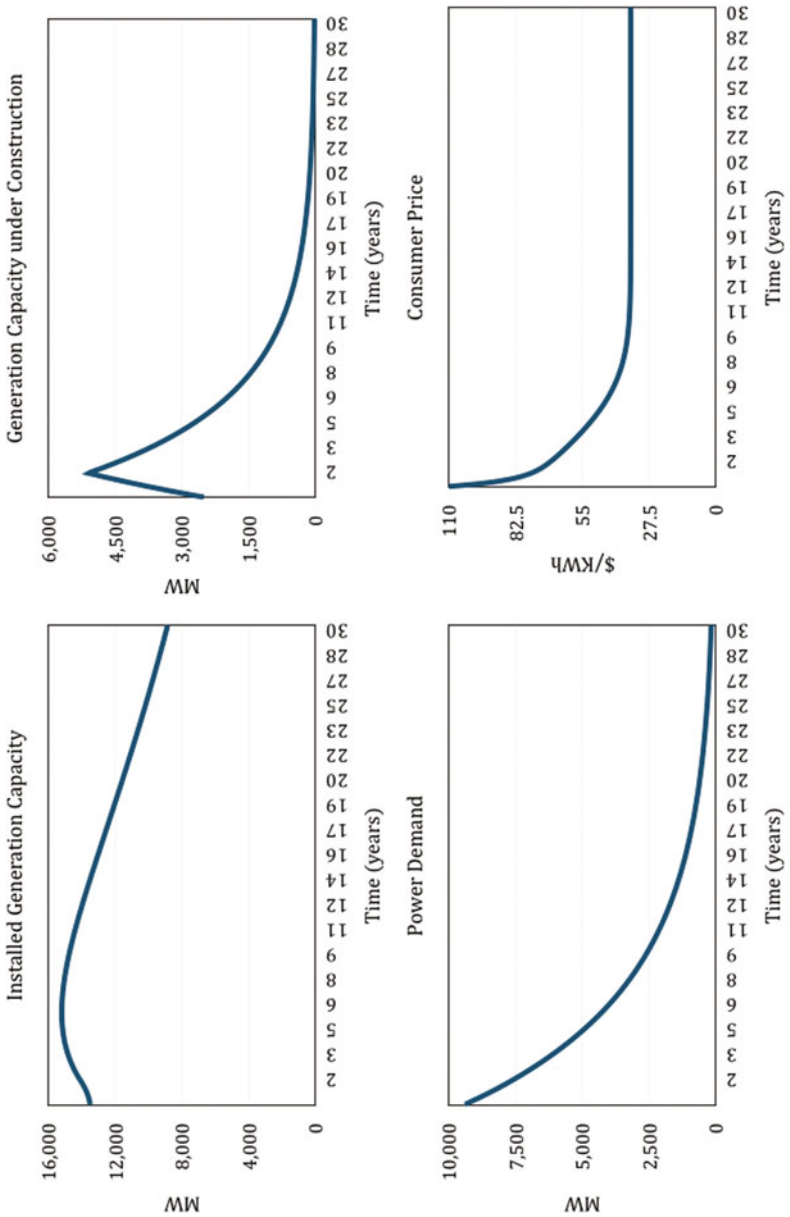


Fig. 4 Simulations results for a grow rate demand $k = -0.005$

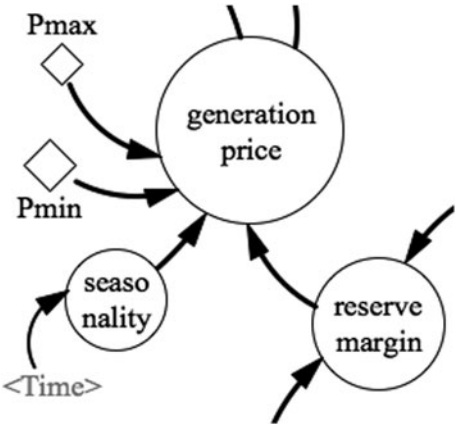


Fig. 5 Exogenous climate perturbation

the dynamics of the disturbance. The dynamics of the electricity market is preserved. The stability is not affected. The perturbed system is a periodically forced system.

Moreover, in Fig. 6, you can see how the oscillations are stronger for generation capacity under construction and consumer prices. Therefore, using the sensitivity analysis Vensim toolbox, is achieved demonstrate the oscillations of these variables for a specific range of the rate of demand growth. See the results of this analysis in Fig. 7. Finally, we see what the dynamic tendency and the range of solutions. Furthermore, we see that the perturbed dynamics is preserved.

4 Conclusions

Specifically, we have studied the model of a national electricity market, making this nonlinear analysis and simulation. We conclude that the nonlinear analysis is a useful tool in characterizing the various phenomena that such systems may exhibit. The scheme under which the numerical integration of solutions is performed is appropriate.

The mathematical analysis that sheds the study of market models, to determine what are the leverage points of the system, and find the reasons for some parameter values the dynamics of the system is unstable. Clearly evidenced as making systematic use of SD to extract and represent mental models using differential equations, gains access to the root of the dynamic behavior of the system.

Finally, it has been shown that by using a set of numerical and analytical tools used in this chapter, can be studied systems that model the supply and demand of electricity in a national market. Combining numerical, analytical tools and classifying nonsmooth phenomena, we can identify the most important behaviors in the system.

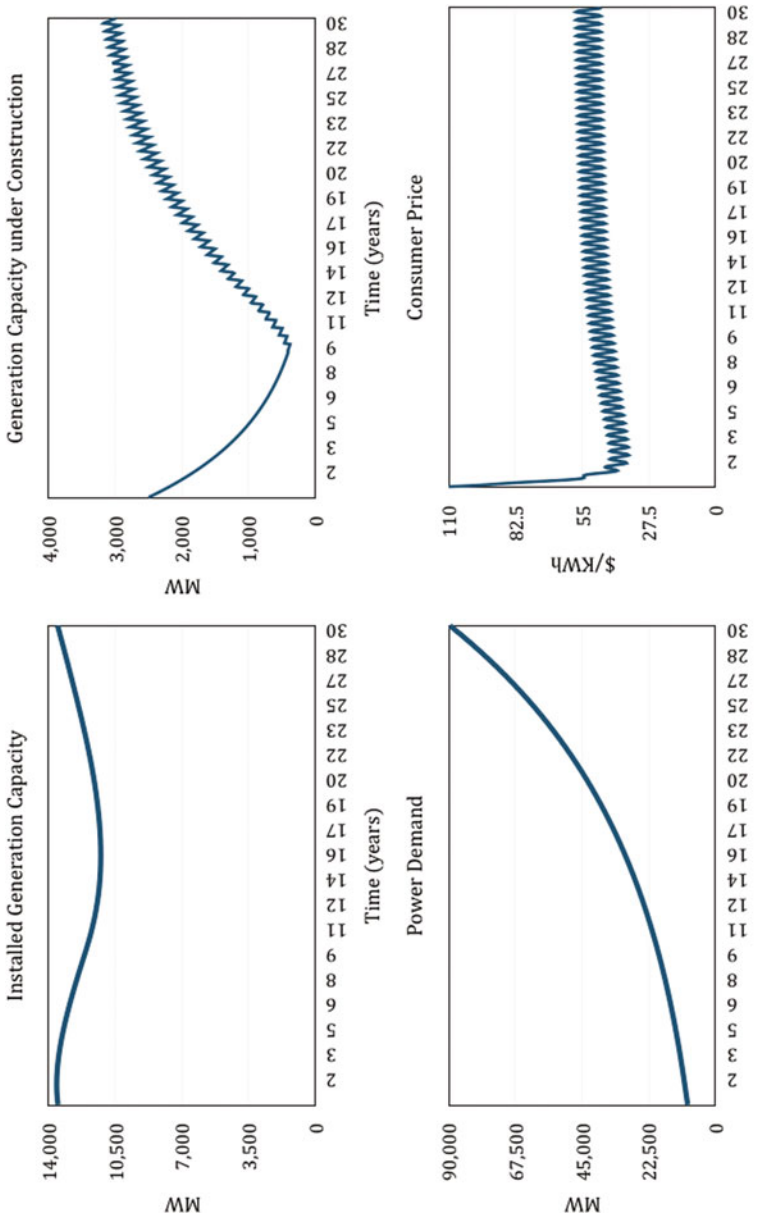


Fig. 6 Simulations results for a grow rate demand $k = 0.03$ with a climate perturbation

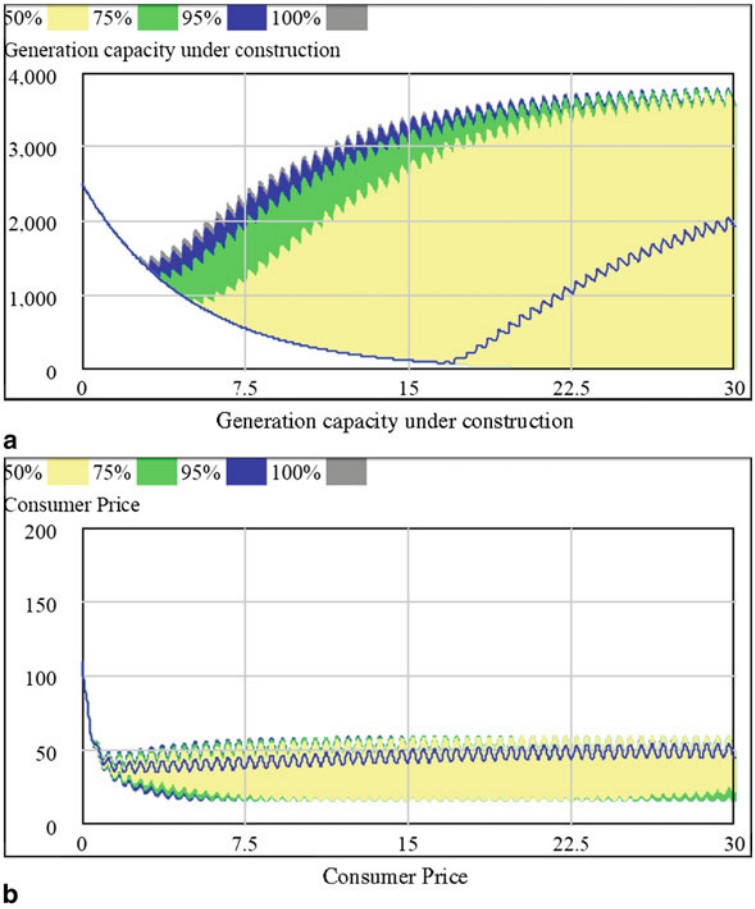


Fig. 7 Sensibility analysis results for a grow rate demand −0.03 to 0.03 with a climate perturbation

Acknowledgment The authors thank the National University of Colombia and Colciencias for the economic support in this research.

References

1. Alzate, R., Bernardo, M., Giordano, G., Rea, G., Santini, S.: Experimental and numerical investigation of coexistence novel bifurcations and chaos in a cam-follower systems. SIAM J. Appl. Dyn. Syst. **8**(2), 592–623 (2009)
2. Angulo, F., Olivar, G., Osorio, G., Escobar, C., Ferreira, J., Redondo, J.: Bifurcation in non-smooth systems. Nonlinear Sci. Numer. Simul. **16**, 1783–1786 (2011)
3. Aracil, J.: On the qualitative properties in system dynamics models. Eur. J. Econ. Soc. Syst. **13**(1), 1–18 (1999)

4. Aracil, J., Toro, M.: Métodos Cualitativos en dinámica de sistemas. Secretariado de publicaciones de la Universidad de Sevilla. Editorial Kronos s.a. Sevilla. España. (1993)
5. Bernardo, M., Budd, C., Champneys, A.R., Kowalczyk, P.: Piecewise-smooth dynamical systems: Theory and applications, vol. 163. Springer, London (2008)
6. Bernardo, M., Nordmarkc, A., Olivard, G.. Discontinuity-induced bifurcations of equilibria in piecewise-smooth and impacting dynamical systems. Physica D. **237**(1), 119–136 (2008)
7. Budd, F., Piiroinen, P.: Corner bifurcations in non-smoothly forced impact oscillators. Physica D Nonlinear Phenom. **220**, 127–145 (2006)
8. Colombo, A., Bernardo, M., Hogan, S.J., Jeffrey, M.R.: Bifurcations of piecewise smooth flows: Perspectives, methodologies and open problems. Physica D. **271**, 32–47 (2011)
9. Ford, A., Wright, J., Prize, F.. System dynamics and the electric power industry. **13**(1), 57–85 (1997)
10. Forrester, J.W.: IFORS' Operational Research Hall of Fame Jay Wright Forrester. Int. Trans. Oper. Res. **13**, 483–492 (2006)
11. Forrester, J.W.: System dynamics. **23**(2), 359–370 (2007)
12. Ghaffarzadegan, N., Lyneisb, J., Richardsona, J.: How small system dynamics models can help the public policy process. Syst. Dyn. Rev. **27**, 22–44 (2011)
13. Koster, M.: Vibrations of cam mechanisms. Phillips technical library series. Macmillan, London (1974)
14. Kustetsov, Y.: Elements of applied bifurcation theory. Applied Mathematic Sciences, vol. 112, 3rd edn. Springer, New York (2010)
15. Lane, D.C., Sterman, J.D., Forrester, J.W.: Profiles in Operations Research : Jay Wright Forrester. Springer (2011)
16. Londoño, S., Lozano, C.: Revisión de herramientas aplicadas al modelamiento de mercados de electricidad. Rev. Ing. Investig. **3**, 67–73 (2009)
17. Norton, R.: Cam Design and Manufacturing Handbook. Industrial Press Inc., New York (2002)
18. Quintero, M., Isaza, F.: Dependencia hidrológica y regulatoria en la formación de precio de la energía en un ssistema hidrodominado: Caso sistema eléctrico colombiano. Rev. Ing. Univ. Medellín **12**(22), 85–96 (2013) (ISSN 1692-3324)
19. Redondo, J.: Modelo de oferta y demanda de un mercado nacional de electricidad. Encuentro Colombiano de Dinámica de Sistemas Eudii. En Su Vigésimo Tercera Versión. Problematicos Complejas (2012)
20. Sterman, J.D.: Business Dynamics: Systems Thinking and Modeling for a Complex World. McGraw Hill, Boston (2000)
21. Valencia, J., Osorio, G.: Nonlinear numerical analysis of camfollower impacting systems. In: LASCAS2011 - Latin American Symposium on Circuits and Systems (2011)

Numerical Simulation Analysis of a Traffic Model

Mónica Jhoana Mesa Mazo, Johnny Valencia and Gerard Olivar Tost

Abstract In this chapter, we present an overview of the piecewise smooth model and simulation of a traffic system, characterized by a single vehicle traveling through a sequence of traffic lights that turn "on" and "off" with a specific frequency and a phase. The model includes three main dynamical modes: accelerated, decelerated, and null state. We show the description of the mathematical modeling used to simulate the system. The simulation was developed under an event-driven strategy and implemented in Matlab. Regarding the numerical analysis, we built a bifurcation diagram where the parameter under variation is the cycle of traffic lights. As a principal result, we evidence the effects of the cycle of traffic light in the dynamical behavior of the system.

Keywords Piecewise smooth dynamical systems · Vehicular traffic · Nonlinear numerical analysis · Bifurcations · Chaos

1 Introduction

In recent decades, large cities have faced many difficulties; among them, we can highlight traffic congestion. The book *El Libro Verde* [1] by The Commission of European Communities addresses the impact of transport on the environment. In this book, traffic congestion is defined as a temporary phenomenon, which occurs frequently with variable duration. This effect is caused by the imbalance between

M. J. Mesa Mazo (✉)
Universidad del Quindío, Quindío, Colombia
e-mail: mjmesa@uniquindio.edu.co

J. Valencia
Department of Computer Science and Decision, School of Mines, Universidad Nacional de Colombia, Campus Medellín, Medellín, Colombia
e-mail: jovalenciacal@unal.edu.co

G. O. Tost
Department of Electric, Electronic and Computer Science, Universidad Nacional de Colombia, Campus Manizales, Manizales, Colombia
e-mail: golivart@unal.edu.co

G.O. Tost, O. Vasilieva (eds.), *Analysis, Modelling, Optimization, and Numerical Techniques,* Springer Proceedings in Mathematics & Statistics 121,
DOI 10.1007/978-3-319-12583-1_25

supply, demand, and the capacity of transport infrastructure. In addition, traffic congestion appears in a place and in a given time, where traffic demand exceeds the capacity of the roads.

Population and number of vehicles are dramatically increasing in the cities of Colombia, which generate huge difficulties to its inhabitants and the environment. For instance, it is estimated that by 2015, urban areas will cause 80 % of CO_2 emissions [3]. There are numerous researchers and different traffic studies devoted to creating a model that would help to control and analyze urban traffic [2, 4, 6].

Toledo's model [5] is used to study the vehicular traffic in this research proposal. This model considers the dynamics of one vehicle moving through a sequence of traffic lights. The separation between the nth and $(n+1)$th traffic light is L_n. The nth light is green if $\sin(\omega_n t + \varphi_n) > 0$, otherwise the light will be red, where ω_n is the change frequency of the light and φ_n is the phase at the nth traffic light. These two parameters are important because they propose different control strategies based on system performance and improve the flow of traffic in the city.

2 Model Description

A car in this sequence of traffic lights can have the following situations:

- A positive acceleration a_+ until its velocity reaches the cruising speed $v_{\max}$.
- A constant speed $v_{\max}$ with zero acceleration.
- A negative acceleration a_- until either the car stops or the car accelerates again.
- A zero velocity when the car is stopped in the red-light traffic.

Therefore, we can summarize the equations of motion for the vehicle as:

Accelerated State This state is when the driver increases speed constantly, i.e., the car has a constant acceleration and positive a_+ until the car reaches cruising speed $v_{\max}$ allowed on the road. In this way, the system is as follows:

$$\begin{cases} \dot{x_1} = x_2 \\ \dot{x_2} = a_+ \end{cases} \tag{1}$$

Null State This mode can occur in two situations. The first is when the vehicle reaches the speed limit on the road. Therefore, the car must maintain this speed then its acceleration is zero. Next, the system of equations is determined as follows:

$$\begin{cases} \dot{x_1} = x_2 \\ \dot{x_2} = 0 \end{cases} \tag{2}$$

The second case is when the vehicle is at rest in the position of a traffic light, waiting for it to turn green. Under these conditions, the system of equations is as follows:

$$\begin{cases} \dot{x_1} = 0 \\ \dot{x_2} = 0 \end{cases} \tag{3}$$

Table 1 Normalized variables

Name	Variables
Velocity	$u = \frac{x_2}{v_{\max}}$
Distance	$y = \frac{x_1}{L}$
Time	$\tau = \frac{t}{T_c}$
Cruise time	$T_c = \frac{L}{v_{\max}}$
Cycle	$T_s = \omega_n T_c$

Decelerated State In this mode, the vehicle is forced to slow down its velocity, because the traffic light is on red. Then, the vehicle has a negative acceleration $-a_-$. After that, the equations associated with this state are:

$$\begin{cases} \dot{x_1} = x_2 \\ \dot{x_2} = -a_- \end{cases} \tag{4}$$

Under the conditions above, when the car approaches the nth traffic light with velocity $v_{\max}$, the driver must decide depending on what sign the traffic light is showing (to step on the brakes or not) at the distance $d = v_{\max}^2/2a_-$ (the last stopping point to arrive with null velocity at the traffic light).

Also, if $\sin(\omega_n t + \varphi_n) > 0$, the driver continues through the traffic light at speed $v_{\max}$. If $\sin(\omega_n + \varphi_n) \leq 0$, the driver starts braking with a_- until the car reaches the traffic light with speed $v = 0$ and waits for the next green light, or until the light turns green again with $v \neq 0$, at which point it starts accelerating with a_+.

It is very important to know that it is convenient to normalize the previous model so that the parameters are reduced, and the system units are removed.

To find the normalized model it is necessary to define new variables. Those variables shown in Table 1.

3 First Model

It is very important to understand the dynamics of a single vehicle because this will aid in the comprehension of the complex problem of the interaction between several cars traveling through a sequence of traffic lights.

In this section, we present a first model, characterized by a single vehicle traveling through a sequence of traffic lights that turn on and off where all lights have equal frequency $\omega_n = \omega$, null phase $\varphi_n = 0$. Also, the separation between the nth and $(n + 1)$th traffic light is L_n. The nth light is green if $\sin(\omega_n t + \varphi_n) > 0$, otherwise the light will be red.

Table 2 Parameters

Name	
Change frequency	$\omega_n = \omega$
Cycle	$T_n = \frac{2\pi L}{T_s v_{max}}$
Phase	$\varphi_n = 0$
Cruising speed	$v_{max} = 14$ m/s
Positive acceleration	$a_+ = 2$ m/s^2
Negative acceleration	$a_- = 6$ m/s^2

3.1 Numerical Simulation Analysis

The main strategies for numerical integration of solutions of piecewise smooth systems are event-based schemes [7] and fixed time step. The first is based on a hybrid formulation, while the second is based on problem solving with complementary variables. For the event-based scheme, under which the numerical integration is performed for this case study, there are three main dynamic states, mentioned above:

To know about this dynamic behavior, many numeric simulations were made, and the subsequent suppositions were assumed. In addition, the following data were collected from secondary sources [8]. The values of these parameters are shown in Table 2.

Additionally, the distances between traffic lights were measured at the 19th Street of Armenia in Colombia. These are shown in Table 3).

In the bifurcation diagram shown in Fig. 1, the normalized cycle of traffic light T_s is taken as a parameter of bifurcation and is shown on the horizontal axis. On the other hand, the vertical axis shows the normalized distance of the car.

Within the values between 0.95 and 1.0, there are accumulation lines. Furthermore, it is shown that there is a period doubling to chaos. This behavior is truncated because of the appearance of an orbit of 3T-periodic, and finally, we can see an orbit of 2T-periodic.

In Figs. 2 and 3, the time and number of traffic lights are shown on the horizontal axis simultaneously. In these figures, the following data are shown: the traffic light signal as a thin line, the normalized speed of the car as a dark line, and the module of the distance as a dashed line.

In Fig. 2, the traffic light normalized cycle is $T_s = 1.5$. If we observe the bifurcation diagram Fig. 1 when the parameter $T_s = 1.5$, the orbit of period three can be found. Therefore, it can be seen in Fig. 2 that the values of the normalized speed and distance module are the same in every three traffic lights or three periods as shown by the green circle.

In Figs. 3 and 4, we can see the increase of the period as the traffic light normalized cycle T_s approaches one. Therefore, it can be seen in Fig. 1 due to the accumulated

Table 3 Distance between semaphores

200	200	100	100	100	100	200	100	100	200

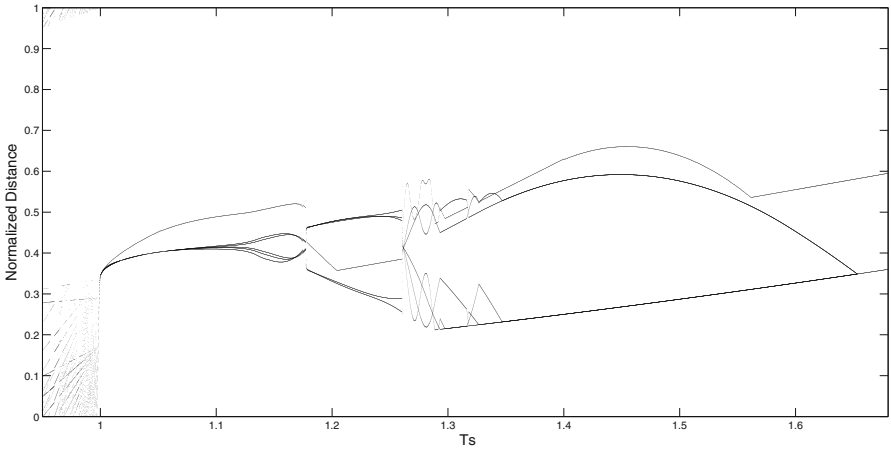


Fig. 1 Bifurcation diagram

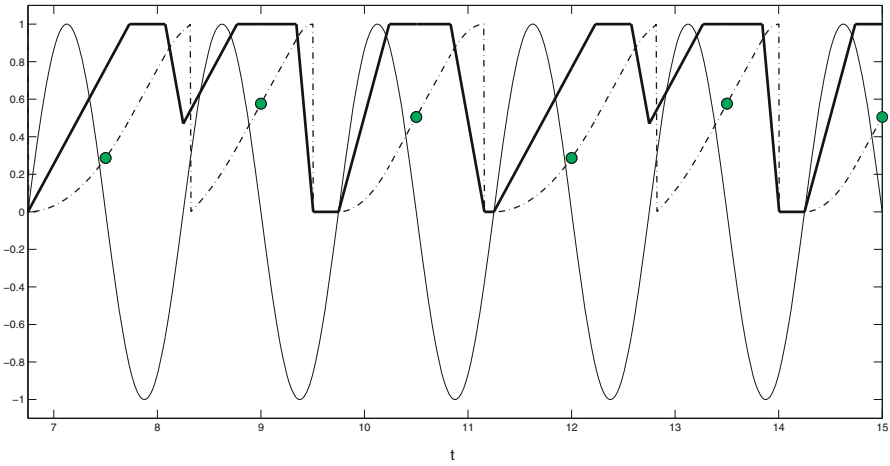


Fig. 2 Diagram for standardized state variables vs. time and $T_s = 1.5$

lines. In addition, we observe that when T_s is the nearest to 1, 0; then the vehicle crosses more green lights with its maximum speed, reducing the travel time through the traffic light sequence is less.

4 Second Model

In this section, we consider a second model characterized by a single vehicle traveling through a sequence of traffic lights. The lights turn on and off where all lights have equal frequency $\omega_n = \omega$ and phase $\varphi_n \neq 0$.

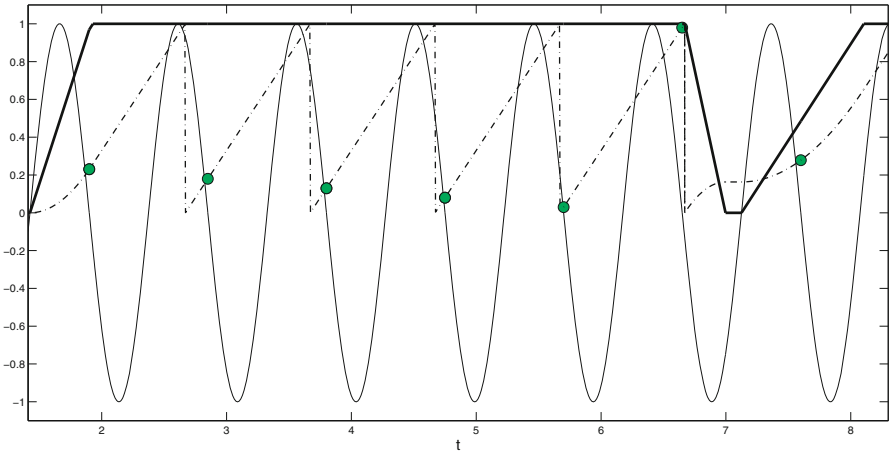


Fig. 3 Diagram for standardized state variables vs. time. $T_s = 0.95$

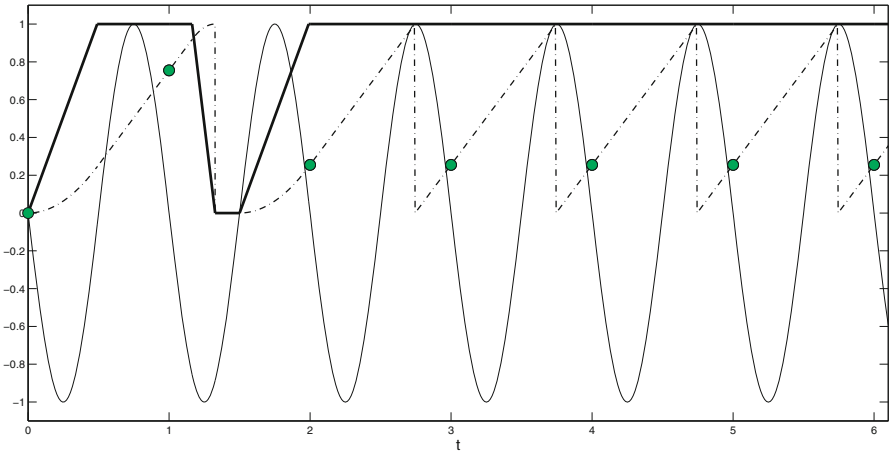


Fig. 4 Diagram for standardized state variables vs. time. $T_s = 1.0$

Toledo proposes to model the offset of traffic light as follows:

$$\varphi_n = -\sum_{m=1}^{n} \frac{L_m \omega_n}{v_{\text{ola}}}, \tag{5}$$

where L_m is the distance between traffic lights, ω_n is the frequency of nth traffic light and v_{ola} is the change velocity at the green light.

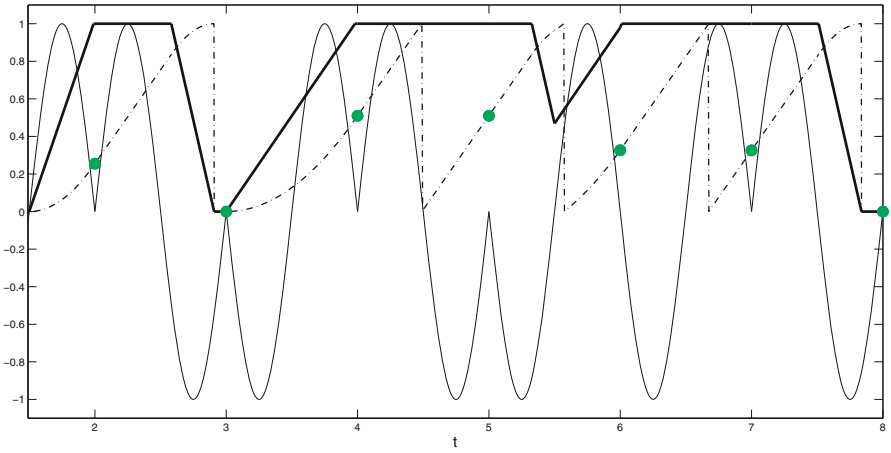


Fig. 5 Diagram for standardized state variables vs. time. Model II, $T_s = 1.0$ and $v_{ola} = 1.0$

4.1 Additional Numerical Simulation Results

Many numeric simulations were made to understand vehicle dynamic behavior comparing The first with the second model. The purpose of comparing these two models is to observe how the phase φ_n influences vehicle behavior.

To make a simulation of model II, we used Tables 2, 3, and Eq. (5).

In Fig. 5 the normalized time and number of traffic lights are shown on the horizontal axis simultaneously. Additionally, in this figure, the following data are shown: the traffic light signal as a thin line, the normalized speed of the car as a dark line, and the module of the distance as a dashed line.

In model I, it is observed that when the traffic light normalized cycle is $T_s = 1.0$, a green wave is obtained. This is guaranteed due to velocity being zero only once, indicating that the driver stopped before passing the second traffic light, see Fig. 4. After that, the driver crossed the entire sequence with the speed limit. In Fig. 5, corresponding to model II, we found that the vehicle is forced to stop on three occasions. In addition, the vehicle travels through the route at varying its speeds, which did not occur in model I.

It was observed that the velocity of the wave v_{ola} and the normalized cycle semaphore T_s have a great influence on vehicle dynamic. In practice, it is useful to synchronize traffic lights to obtain a green wave that could reduce travel time.

The graph shown in Fig. 6 allows us to know the number of times the vehicle stops along the way, due to the configuration of the parameters v_{ola} and T_s. These parameters make the traffic light change to red. Fig. 6 shows the wave speed v_{ola} on the X axis, the normalized cycle T_s on the Y axis, and the number of vehicle stops on Z axis.

The ellipse in Fig. 6 highlights the point $P_1 = (0.9, 1.2, 6)$, which indicates that when $v_{ola} = 0.9$ and $T_s = 1.2$, the vehicle stops six times throughout the way. This

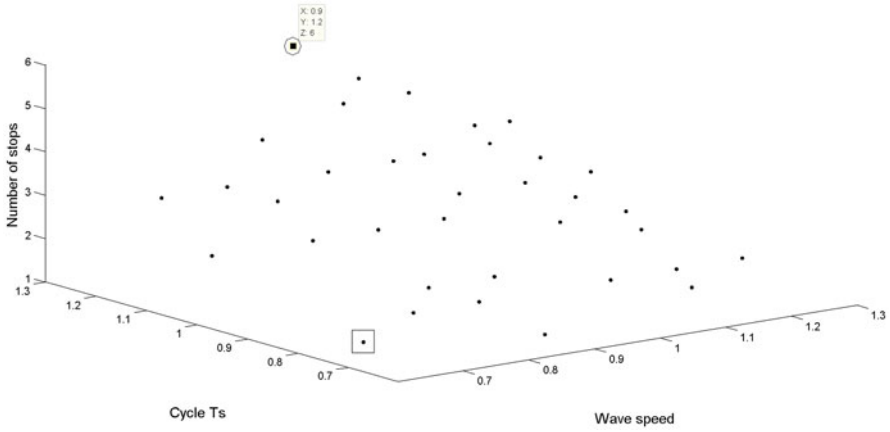


Fig. 6 Number of stops the vehicle. Model II

is not beneficial to the driver because the entire sequence has ten lights. The box on the same figure indicates that if $v_{\text{ola}} = 0.7$ and $T_{\text{s}} = 0.8$, the vehicle is forced to stop only once. This is more beneficial for the driver because he can cross the entire sequence of traffic lights stopping the least number of times, which may reduce the travel time and fuel consumption during the journey.

5 Conclusions

We have presented the numerical scheme to simulate piecewise smooth dynamic systems. Furthermore, we have shown exhaustive numerical simulations for characterizing all phenomena that these kind of systems can exhibit. To simulate these systems it is necessary to know the equations that describe their flow in every state and the conditions in the boundaries to transition between the dynamical states. With this information and the bifurcation diagram, it is possible to simulate a wide range of phenomena that these kind of systems can exhibit. From the bifurcation theory point of view, the results presented in this chapter can be a useful tool to obtain a better model of the systems of interest.

We chose an adequate framework to model and simulate the systems. The cycle of traffic lights is a very important parameter of bifurcation as evident in the bifurcation diagram. Additionally, when the cycle of traffic lights is close one, the green wave appears; then travel time is the smallest.

We show that an appropriate configuration of the traffic lights optimizes the travel time of a vehicle, where a reduction in the number of stops could reduce fuel consumption and thus their environmental impacts.

Additionally, we propose that knowing the value of the semaphore parameters is very useful to minimize the number of stops a vehicle makes while driving. Synchronizing traffic lights by generating a green wave minimizes the number of stops a vehicle makes.

The numerical simulation analysis has allowed us to see the effects of the phase of the traffic lights. This phase is a representative system parameter which approximates the actual dynamic models used in traffic systems in cities. In addition, the proper configuration of this parameter generates or deletes sections of green wave.

Starting from the bifurcation diagrams, a study of the parameter T_s (standard cycle) was performed, showing complex behaviors associated with it. Our work was conducted around the normalized cycle light, which is a control parameter, in addition to Villalobos and Toledo. Robust control theory teaches us that the parameters influencing microscopic systems are the same that determine the dynamics of microscopic systems.

References

1. Elorrieta, D.I., Perlado, S: Libro verde de medio ambiente urbano, Ministerio del medio ambiente España, Marzo 2007
2. Jaramillo, D.: Simulación y control de tráfico vehicular por semaforización. Universidad Pontificia Bolivariana (2004)
3. Lupano, J.A., Sánchez, R.J.: Políticas de movilidad urbana e infraestructura urbana de transporte. Publicación de las Naciones Unidas, Febrero 2009
4. Mesa, M.J., Olivar, G.: Bifurcaciones en un sistema de tráfico. DDays 2012, South America, Noviembre 2012
5. Toledo, B.A., Rogan, J., Muñoz, V., Tenreiro, C., Valdivia, J.A.: Modeling traffic through a sequence of traffic lights. Phys. Rev. E **70**, 016107 (2004)
6. Toledo, B., Cerda, E., Rogan, J., Muñoz, V., Tenreiro, C., Zarama, R., Valdivia, J.: Universal and nonuniversal features in a model of city traffic. Phys. Rev. E **75**(2), 026108 (2007)
7. Valencia, J., Osorio, G.: Bifurcación por impacto con esquina en el sistema leva seguidor. Universidad Nacional de Colombia (2012)
8. Villalobos, J.: Chaos in transit systems. Universidad de los Andes, Departamento de Ingeniería Industrial. Ceiba Complejidad, August 2010