XVII CMA / XXVII CEDYA

PROCEEDINGS BOOK

XXVII
CONGRESO DE ECUACIONES
DIFERENCIALES Y APLICACIONES

XVII
CONGRESO DE
MATEMATICA APLICADA

ZARAGOZA | 18 AL 22 DE JULIO 2022

Zaragoza, July 18th–22nd, 2022

SēMA Sociedad Española de Matemática Aplicada

Instituto Universitario de Investigación de Matemáticas y Aplicaciones
Universidad Zaragoza

Editors: Chelo Ferreira González, José Luis Gracia Lozano, Etelvina Javierre Pérez, Eduardo Martínez Fernández, Pedro J. Miana Sanz, Ester Pérez Sinusía, Luis Rández García, Teresa Sánchez Rúa, Raquel Villacampa Gutiérrez

BY-NC-SA Proceedings XXVII CEDYA / XVII CMA (Zaragoza, 2022)
Chelo Ferreira, José Luis Gracia, Etelvina Javierre, Eduardo Martínez, Pedro J. Miana, Ester Pérez, Luis Rández, Teresa Sánchez y Raquel Villacampa
Edita: Prensas de la Universidad de Zaragoza
ISBN 978-84-18321-66-5

Prensas de la Universidad
Universidad Zaragoza

Servicio de Publicaciones de la Universidad de Zaragoza
Edificio de Ciencias Geológicas
Pedro Cerbuna, 12. 50009 Zaragoza, España

Contents

PREFACE

The Congress of Differential Equations and Applications / Congress of Applied Mathematics (CEDYA / CMA) is the biennial congress of the Spanish Society of Applied Mathematics (S$\vec{e}$MA). The first CEDYA was celebrated in September 1978 in El Escorial (Madrid), and the first joint CEDYA / CMA took place in Málaga in 1989.

The XXVII CEDYA / XVII CMA was held from 18^{th} to 22^{nd} July 2022 at the Facultad de Medicina of the University of Zaragoza and was organized by the Instituto Universitario de Investigación de Matemáticas y Aplicaciones de la Universidad de Zaragoza (IUMA). The congress format was hybrid due to the uncertain pandemic situation caused by COVID-19.

This congress attracted near 350 participants from different universities. They presented 250 lectures, eight of which were invited. The conference was structured in eighteen mini-simposia, proposed by different researchers and groups, eight special sessions and a poster session, both organized by the Local Organizing Committee. The topics of the conference covered Partial Differential Equations, Dynamical Systems and Ordinary Differential Equations, Numerical Analysis and Simulation, Numerical Linear Algebra, Optimal Control and Inverse Problems, Mathematics Applied to Industry, Social Sciences and Biology, Mathematical Education, Scientific Computation, Approximation Theory and Discrete Mathematics.

These Proceedings have been published in the institutional repository of the University of Zaragoza. They contain nineteen selected papers associated with the lectures presented at XXVII CEDYA / XVII CMA. The editors would like to thank the authors for their contributions and cooperation, without them it would have been impossible to produce these proceedings.

Finally, we thank the sponsors of the conference: Sociedad Española de Matemática Aplicada, IUMA, Facultades de Medicina y Ciencias de la Universidad de Zaragoza and the Vicerrectorado de Educación Digital y Formación Permanente de la Universidad de Zaragoza. We also wish to thank the Scientific Committee, the organizes of the mini-symposia, all the conference participants, and the students collaborators, who were hugely helpful in the organization of the conference.

Zaragoza, April 2023

The Local Organizing Commitee CEDYA / CMA 2022

Scientific Committee

- José Carlos Bellido (Universidad de Castilla-La Mancha, Spain)
- R. Bürger (Universidad de Concepción, Chile)
- Luz de Teresa (Universidad Nacional Autónoma de México, México)
- Rosa Donat (Universitat de València, Spain)
- Michael Dumbser (University of Trento, Italy)
- Ernesto Estrada (Instituto de Física Interdisciplinar y Sistemas Complejos, Spain)
- Natalia Kopteva (University of Limerick, Ireland)
- Tere Martínez-Seara (Universitat Politècnica de Catalunya, Spain)
- Juan Ignacio Montijano (Universidad de Zaragoza, Spain)
- Juan Manuel Peña (Universidad de Zaragoza, Spain)

Sponsors

- Sociedad Española de Matemática Aplicada
- Instituto Universitario de Investigación de Matemáticas y Aplicaciones
- Facultad de Medicina de la Universidad de Zaragoza
- Facultad de Ciencias de la Universidad de Zaragoza
- Vicerrectorado de Educación Digital y Formación Permanente de la Universidad de Zaragoza
- Universidad de Zaragoza

Local Organizing Committee

- Chelo Ferreira
- José Luis Gracia
- Etelvina Javierre
- Eduardo Martínez
- Pedro J. Miana
- Ester Pérez
- Luis Rández
- Teresa Sánchez
- Raquel Villacampa

Iteration of rational Hopf-endomorphisms for graphical representation of basins of attracting n-cycles

V. Álvarez-Aparicio[1], J.M. García Calcines[2], L.J. Hernández-Paricio[1], M.T. Rivas-Rodríguez[1]

1. *Universidad de La Rioja, Dpto. de Matemáticas y Computación, Edificio CCT, C/Madre de Dios, 53, Logroño, 26006, Spain*
2. *Universidad de La Laguna, Dpto. de Matemáticas, Estadística e I.O. e Instituto de Matemáticas y Aplicaciones (IMAULL), Avda. Astrofísico Fco. Sánchez, s/n, Facultad de Ciencias. Sección de Matemáticas, La Laguna, 38200, Spain*

Abstract

In this work, we present a new method to compute the basins of attraction of any complex rational map, and to study the discrete-time dynamical behaviour of its attracting n-cycles. This new method, whose development has been influenced by some highly relevant results of Complex Dynamics such as the Ergodic Theorem, is closely related to Lyapunov exponents, a widely used concept on the study of non-linear continuous dynamical systems.

In addition, the implementation of the collection of algorithms that make up this method will be briefly commented, since it helps to solve some of the computational problems that often arise in Numerical Analysis, like overflows or mathematical indeterminations. This implementation is based on the iteration of Hopf-endomorphisms induced by rational maps, and it can be found in the Lyapunov Cycle Detector module, available in the following GitHub repository: github.com/LCD.

Finally, we will address a concrete example of this new method to the study of the basins of attraction induced by Chebyshev's method when applied to a complex cubic polynomial.

1. Introduction

In this work, we present a new collection of algorithms dedicated to compute the basins of attraction of a complex rational map, based on Lyapunov exponents, as well as some of the key theoretical results in which they are based. Computing the basins of attraction of a rational map is a relevant matter when studying some of the most extensively used numerical methods to approximate solutions of non-linear polynomial equations.

For example, when one applies Newton method to a polynomial, it induces a rational map for which the roots of the polynomial are super-attracting 1-cycles (fixed points). In order to compute the basins of attraction, we will define a function that is constant on each basin, and which we will call a Lyapunov function due to its connection with Lyapunov exponents, as we will address later on. This way, we can divide the Riemann sphere $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$, where $\mathbb{C}$ denotes the plane of complex numbers, into the different basins of attraction induced by the rational map and its Julia set. It is also worth noting that the method we propose can be applied without having to compute the list of n-cycles of the rational map beforehand.

From a scientific programming point of view, this new collection of algorithms solves some of the computational problems that often arise in Numerical Analysis, like overflows or mathematical indeterminations. We achieve this by considering the Hopf fibration $S^3 \to S^2 \cong P^1(\mathbb{C})$, where S^n denotes the usual n-sphere and $P^1(\mathbb{C})$ the complex projective line, and computing the Hopf-endomorphism induced by the given rational map. This approach also allows us to easily work with the infinity point ∞. Since this kind of computations are often very heavy, we chose Julia Language in order to implement our algorithms, due to its efficiency, speed and proper syntax for mathematics. Some parallel programming techniques are also applied in order to reduce the execution time of the algorithms.

This article will have the following structure:

- First, we will introduce the mentioned topological-geometrical model, addressing the Hopf fibration and giving a construction of the Hopf-endomorphism induced by a rational map, which are key notions of the presented theory.
- Then, we will define the Lyapunov functions which will be used by our algorithms to compute the basins of attraction, and we will mention some of their properties proved in [2].
- Also, an schematic description of the algorithms will be given, and we will mention how they are capable to sort some computational problems which could appear when computing the basins of attraction.
- Finally, we will apply our theory and algorithms to study a concrete example regarding Chebyshev's method applied to a complex cubic polynomial.

2. Hopf-endomorphisms associated with rational maps

2.1. The Riemann Sphere

We consider three different models of the Riemann sphere: the Alexandroff's compactification of the plane of complex numbers $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$, the usual 2-sphere $S^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}$, and the complex projective line $P^1(\mathbb{C}) = \frac{\mathbb{C}^2 \setminus \{(0,0)\}}{\sim}$, built from the following equivalence relation: $(z, t) \sim (z', t')$ if and only if $\exists \lambda \in \mathbb{C} \setminus \{0\}$ such that $(z, t) = (\lambda z', \lambda t')$.

The equivalence class of a point $(z, t) \in \mathbb{C}^2 \setminus \{(0, 0)\}$ will be denoted by $[z : t] \in P^1(\mathbb{C})$. Moreover, we will say that (z, t) are the *homogeneous coordinates* of the point $[z : t]$, and that its *absolute coordinates* are $\frac{z}{t}$ if $t \neq 0$, and $\frac{t}{z}$ if $t = 0$ and $z \neq 0$.

Given a point $[z : t] \in P^1(\mathbb{C})$, it will be useful for our algorithms to consider its *normalized homogeneous coordinates*, which will be denoted by $[z, t]$, given by $\left(\frac{z}{|z|+|t|}, \frac{t}{|z|+|t|}\right) \in \mathbb{C}^2 \setminus \{(0, 0)\}$. Considering normalized homogeneous coordinates, our algorithms will be able to avoid numerical overflows when we compute the basins of attraction.

It is known that there exists an analytic isomorphism between each pair of the previously mentioned models of the Riemann sphere, so, even though our rational maps $f : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$ are naturally defined over $\hat{\mathbb{C}}$, we can apply the corresponding analytic isomorphism and transfer the study of the iteration of the rational map to $P^1(\mathbb{C})$ (considering rational maps of the form $f : P^1(\mathbb{C}) \to P^1(\mathbb{C})$) in order to use homogeneous coordinates and avoid numerical problems that can arise in a computational environment.

2.2. Homogeneous pairs of bivariate polynomials

One of the key notions needed to construct the Hopf-endomorphism associated with a rational map is that of homogeneous polynomials.

Let $F \in \mathbb{C}[z, t]$ be a bivariate complex polynomial. We say that F is *homogeneous* if $F = 0$ or if $\exists k \in \mathbb{N}$ such that $\forall \lambda \in \mathbb{C} \setminus \{0\}$, $F(\lambda z, \lambda t) = \lambda^k F(z, t), \forall z, t \in \mathbb{C}$.

Let $F \in \mathbb{C}[z, t]$ be a homogeneous bivariate complex polynomial. Then if $F \neq 0$, we say that the *degree of* F is the smallest $k \in \mathbb{N}$ such that $\forall \lambda \in \mathbb{C}\setminus\{0\}$, $F(\lambda z, \lambda t) = \lambda^k F(z, t), \forall z, t \in \mathbb{C}$, and if $F = 0$, we say its degree is $-\infty$.

For every univariate complex polynomial one can define a homogeneization operator that transforms it into a homogeneous bivariate complex polynomial. It can be defined the following way:

Let $\mathbb{C}[z]\underline{\times}\mathbb{N} = \{(A, d) \in \mathbb{C}[z] \times \mathbb{N} \mid deg(A) \leq d\}$. Then we have the following operator $H : \mathbb{C}[z]\underline{\times}\mathbb{N} \to \mathbb{C}[z, t]$ given by $H(A, d) = 0$ if $A = 0$, and

$$H(A, d) = a_0 t^d + a_1 z t^{d-1} + \cdots + a_n z^n t^{d-n}, \tag{2.1}$$

if $A \neq 0, A = a_0 + a_1 z + \cdots + a_n z^n, a_n \neq 0$.

Note that this operator transforms a complex polynomial A and an upper bound of its degree d into a homogeneous bivariate polynomial $H(A, d)$.

Of course, when working with a rational map, one can consider it simply as a pair $(F, G) \in \mathbb{C}[z, t] \times \mathbb{C}[z, t]$ just by applying the homogeneization operator both to its numerator and denominator. Then, we say that a pair $(F, G) \in \mathbb{C}[z, t] \times \mathbb{C}[z, t]$ of homogeneous bivariate complex polynomials is a *homogeneous pair* if $FG = 0$ or if F and G both have the same degree. In addition, if $G \neq 0$, then we say that (F, G) is an *r-homogeneous pair*.

Moreover, we say that a homogeneous pair $(F, G) \in \mathbb{C}[z, t] \times \mathbb{C}[z, t]$ of homogeneous bivariate complex polynomials is *irreducible* if $(F, G) = (0, 0)$ or if $(F, G) \neq (0, 0)$ and, if $(F, G) = (HF_1, HG_1)$ for some $H, F_1, G_1 \in \mathbb{C}[z, t]$ homogeneous bivariate polynomials, then $deg(H) = 0$.

It will also be beneficial for us to consider the normalization of an irreducible r-homogeneous pair of bivariate polynomials, in order to prevent overflows in our computations.

Let $F \in \mathbb{C}[z, t]$ be a bivariate polynomial given by $F(z, t) = a_0 t^d + a_1 z t^{d-1} + \cdots + a_n z^n t^{d-n}$. We say that the *norm of* F, denoted by $\|F\|$, is given by the expression $\|F\| = |a_0| + |a_1| + \cdots + |a_n|$.

Let $(F, G) \in \mathbb{C}[z, t] \times \mathbb{C}[z, t]$ be a homogeneous pair, with $F(z, t) = a_0 t^d + a_1 z t^{d-1} + \cdots + a_n z^n t^{d-n}$ and $G(z, t) = b_0 t^d + b_1 z t^{d-1} + \cdots + b_m z^m t^{d-m}$. We say that the *norm of the homogeneous pair* (F, G), denoted by $\|(F, G)\|$, is given by $\|(F, G)\| = \|F\| + \|G\|$.

Thus, if we have a homogeneous paif of bivariate polynomials (F, G), we define its *normalization*, denoted by $[F, G]$, as

$$[F,G] = \begin{cases} (0,0) & \text{si} \quad (F,G) = (0,0) \\ \left(\frac{F}{\|(F,G)\|}, \frac{G}{\|(F,G)\|}\right) & \text{si} \quad (F,G) \neq (0,0) \end{cases} \tag{2.2}$$

We will say that a homogeneous pair (F, G) is *normalized* if it verifies that $(F, G) = [F, G]$.

Attending to the previous notions, if $R(z) = \frac{A(z)}{B(z)}$ is a rational map, we can consider an irreducible representation $R(z) = \frac{A_1(z)}{B_1(z)}$. Then, if we apply the homogeneization operator both to A_1 and B_1 considering $d = max\{deg(A_1), deg(B_1)\}$ as the upper bound of the degrees of the numerator and denominator, we obtain the irreducible homogeneous (r-homogeneous if $G \neq 0$) pair (F, G) of bivariate complex polynomials, representing the rational map R.

This way, when we have a rational map R we have seen that following this process one can always represent R as an irreducible r-homogeneous pair (F, G) of bivariate complex polynomials, which can be very useful when working with the iteration of R in a computational environment, since we avoid the calculations in which indeterminations and overflows can occur.

2.3. Construction of the Hopf-endomorphism associated with a rational map

In order to construct the Hopf-endomorphism associated with a rational map R, we consider the usual 3-sphere $S^3 = \{(z, t) \in \mathbb{C}^2 \mid |z| + |t| = 1\}$, which we consider as a subspace of $\mathbb{C}^2$.

Consider the following surjective quotient maps

$$\mathbb{C}^2 \setminus \{(0,0)\} \xrightarrow{p} S^3 \xrightarrow{q} P^1(\mathbb{C}),$$

where $p : \mathbb{C}^2 \setminus \{(0,0)\} \to S^3$ is given by $p(z,t) = [z,t] = \left(\frac{z}{|z|+|t|}, \frac{t}{|z|+|t|}\right) \in S^3$, and where $q : S^3 \to P^1(\mathbb{C})$ is given by $q(z,t) = [z : t]$, $(z,t) \in S^3$.

Note that its composition transforms each point $(z,t) \in \mathbb{C}^2 \setminus \{(0,0)\}$ into a point $[z' : t'] \in P^1(\mathbb{C})$ whose normalized homogeneous coordinates are $\left(\frac{z}{|z|+|t|}, \frac{t}{|z|+|t|}\right)$.

We will refer to q as the *Hopf fibration* and to the composition qp as the *extended Hopf fibration*.

Finally, if (F, G) is an irreducible r-homogeneous pair of bivariate polynomials representing a rational map R, the following diagram is induced by the extended Hopf fibration:

$$\begin{array}{ccc} \mathbb{C}^2 \setminus \{(0,0)\} & \xrightarrow{R} & \mathbb{C}^2 \setminus \{(0,0)\} \\ \downarrow p & & \downarrow p \\ S^3 & \xrightarrow{R^S} & S^3 \\ \downarrow q & & \downarrow q \\ P^1(\mathbb{C}) & \xrightarrow{R^P} & P^1(\mathbb{C}) \end{array}$$

where the map $q : S^3 \to P^1(\mathbb{C})$ is the Hopf fibration, and where R, R^S and R^P are given by

$$\begin{aligned} R(z,t) &= (F(z,t), G(z,t)), \text{ where } (z,t) \in \mathbb{C}^2 \setminus \{(0,0)\}, \\ R^S(z,t) &= [F(z,t), G(z,t)], \text{ where } (z,t) \in S^3, \\ R^P([z:t]) &= [F(z,t) : G(z,t)], \text{ where } [z:t] \in P^1(\mathbb{C}), \end{aligned}$$

We will refer to the pair (R^S, R^P) as *Hopf-endomorphism induced by the rational map R*.

We have seen how to represent a rational map through its associated Hopf-endomorphism, which can be useful when computing its basins of attraction and avoiding possible computational problems. Now, our main concern will be to provide a fitting method to compute the basins of attraction and to extract some information regarding the dynamics of the considered rational map.

3. Lyapunov functions to compute the basins of attraction

In order to compute and distinguish between the different basins of attraction, we will define a function which is constant in each basin and that has a strong connection with the Lyapunov exponents of the discrete-time dynamical system induced by the iteration of the rational map.

It involves the notion of *spherical derivative* of a rational map $f : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$, which is a map $f^{\#} : \hat{\mathbb{C}} \to \mathbb{R}_+ = \{r \in \mathbb{R} \mid r \geq 0\}$ given by

$$f^{\#}(z_0) = |f'(z_0)| \frac{1 + |z_0|^2}{1 + |f(z_0)|^2}, \tag{3.1}$$

where $f'(z_0)$ denotes the usual derivative of f in $z_0 \in \hat{\mathbb{C}}$, for every $z_0 \in \hat{\mathbb{C}}$.

Note that if f' is not defined in z_0, we can define $f^{\#}(z_0)$ as the limit of the expression above when $z \to z_0$, and it can be proven that this limit always exist, $f^{\#}$ is well defined and it is continuous. Also, since $f^{\#}$ is continuous and $\hat{\mathbb{C}}$ is compact, $f^{\#}$ is bounded.

We will say that $z_0 \in \hat{\mathbb{C}}$ is a *critical point* of f if $f^{\#}(z_0) = 0$, and we will also consider the usual notions of super-attracting, attracting, indifferent and repelling points and cycles, using the spherical derivative $f^{\#}$.

It is important to note that the we have defined the spherical derivative of a rational map defined over $\hat{\mathbb{C}}$ but, thanks to the previously mentioned analytic isomorphism between $\hat{\mathbb{C}}$ and $P^1(\mathbb{C})$ we can transfer this notion to the complex projective line, as it can be seen in [2]. In the reference, one can also find specific details and results on how to compute the spherical derivative of a rational map represented by its associated Hopf-endomorphism avoiding possible numerical problems.

3.1. Definition and properties of the Lyapunov function

The function we use to compute the basins of attraction of a rational map can be defined for continuous maps in more general spaces as follows: Let X be a topological space, and let $f : X \to X$ and $\phi : X \to \mathbb{R}_+$ be continuous maps. We define the function $L_f(\phi) : X \to [0, +\infty]$ given by the expression

$$L_f(\phi)(x) = \lim_{n \to +\infty} \left(\prod_{k=0}^{n-1} \phi(f^k(x)) \right)^{\frac{1}{n}}, \tag{3.2}$$

which we will call the *Lyapunov function of f associated with ϕ*.

Note that the domain of $L_f(\phi)$ is a subset of X not necessarily equal to X.

Also, if we consider the logarithm of $L_f(\phi)$, we obtain the time average of $log(\phi)$. This allows us to establish a connection between the presented method and Birkhoff's Ergodic Theorem [4].

It is also worth noting that Lyapunov functions are used frequently in the context of Dynamical Systems to study local stability. In our case, despite that the function $L_f(\phi)$ is not a Lyapunov function in that context, we will also call it a Lyapunov function, since it will be used to study local stability and dependency on initial conditions.

Since we want to be able to use and compute the Lyapunov function, it is also interesting to consider an approximation. Let X be a topological space, and let $f : X \to X$ and $\phi : X \to \mathbb{R}_+$ be continuous maps. We define $L_f^{[r,s]}(\phi) : X \to \mathbb{R}_+$ given by the expression:

$$L_f^{[r,s]}(\phi)(x) = \left(\prod_{k=r}^{s} \phi(f^k(x)) \right)^{\frac{1}{s-r+1}}, \tag{3.3}$$

which we will call the $[r, s]$-*approximation of* $L_f(\phi)$.

Note that in this case the domain of $L_f^{[r,s]}(\phi)$ is X.

As it has been proven in [2], the previously defined functions, under certain conditions (that are verified in our study), satisfy that they are equal-valued and constant in each basin of attraction of f. We will call each one of these constants *Lyapunov constant associated with the basin of attraction*. It is important to note that these constants might not be different so, in order to distinguish the basins of attraction, we have to take that into account when implementing the algorithms.

In particular, if we consider $X = P^1(\mathbb{C})$, f a rational map represented by its induced Hopf-endomorphism, and $\phi = f^{\#}$ the spherical derivative of f (considered over $P^1(\mathbb{C})$), then the logarithm of a Lyapunov constant is precisely a Lyapunov exponent of the discrete-time dynamical system induced by the iteration of the rational map f, which are studied in [5].

4. Computing the Lyapunov constants and the basins of attraction

In order to compute the basins of attraction of a rational map, we have to consider a finite set of points of $P^1(\mathbb{C})$ (a grid) to which we will apply the procedure described in this section. Since we want to avoid possible overflows in our calculations, we will consider the normalized homogeneous coordinates of each point of the grid.

It is also worth noting that the algorithm exposed in this section does not depend on the previous calculation of the list of n-cycles of the rational map. Instead, thanks to the properties of the considered Lyapunov functions, our procedure detects when does the orbit of a given point converge to a n-cycle. However, due to computational reasons, in order to apply this method one has to choose the maximum length of the q-cycles that the method will detect.

4.1. Description of the algorithm

Let $x = [z : t]$ be a point of the considered grid, let $f : P^1(\mathbb{C}) \to P^1(\mathbb{C})$ be a rational map, and let $\phi = f^{\#}$ be its spherical derivative. We have that, except in a set with zero Lebesgue measure, if x is in the domain of $L_f(\phi)$, then for every $T \in \mathbb{N}$ and for every $r \in \mathbb{N}$, there exists $N_{(T,r)} > r$ such that

$$|L_f^{[r,n-1]}(\phi)(x) - L_f^{[r,n]}(\phi)(x)| \leq 10^{-T},\ \forall n \geq N_{(T,r)}. \tag{4.1}$$

Therefore, given a grid of points of $P^1(\mathbb{C})$, a tolerance 10^{-T} ($T \in \mathbb{N}$), a maximum number of iterations, and the maximum length of the q-cycles we want to compute, we can apply the following procedure:

0. If there are points of the grid we have not yet considered, we choose a new point $x \in P^1(\mathbb{C})$. If we have already applied the algorithm to every point of the grid, the algorithm ends.
1. We take a new value of $q \in \mathbb{N}$ (starting at 1), where q determines the length of the q-cycle we are trying to detect.
2. We compute the next two terms (starting at $k = 0$) of $(L_f^{[k,k+q-1]}(\phi)(x))_{k\in\mathbb{N}}$; that is, we compute $L_f^{[k,k+q-1]}(\phi)(x)$ and $L_f^{[k+1,k+q]}(\phi)(x)$.
3. We compare the terms and, if we have that

$$|L_f^{[k,k+q-1]}(\phi)(x) - L_f^{[k+1,k+q]}(\phi)(x)| < 10^{-T}, \tag{4.2}$$

   then we consider that the orbit of the point x has converged to a q-cycle and we go back to step 0 to repeat the procedure with another point. Note that, in this case, the last term we have computed, $(L_f^{[k+1,k+q]}(\phi)(x))$, is an approximation of the Lyapunov constant associated with the basin of attraction in which x lies.

   If the condition 4.2 is not satisfied, we go back to step 2 and compute the next two terms.
4. If we have reached the given maximum number of iterations and the process has not yet converged (up to the given tolerance), then we go back to step 1, choose the next value of q and study whether if the orbit of the point x converges to some $(q+1)$-cycle.
5. If we have reached the given maximum length of the q-cycles considered and the process has not yet converged (up to the given tolerance), then we simply go back to step 0 and consider another point of the grid.

Once this procedure has been applied to every point in the considered grid (which represents a region of $P^1(\mathbb{C})$, and thus a region of $\hat{\mathbb{C}}$), we will have divided a region of the Riemann sphere $\hat{\mathbb{C}}$ into the different basins of attraction induced by the iteration of the rational map f and its Julia set. It is also important to note that we can cover the entire Riemann sphere $S^2 \cong P^1(\mathbb{C})$ with just a neighborhood of the origin and a neighborhood of the infinity point ∞. This implies that, if we apply the previous process to two sufficiently large grids of points (each one representing a neighborhood of the origin and of the infinity point, respectively), it will suffice to divide the whole Riemann sphere into the different basins and the Julia set.

Moreover, thanks to the properties of the Lyapunov function involved in the process, one can extract from this method useful information about the dynamics of f, such as its Lyapunov spectrum.

4.2. Implementation in Julia Language

We have implemented the described algorithm, along with some additional functionalities that could be of use, in Julia Language. We have chosen this particular language for several reasons, including its speed for numerical calculations, its proper syntax for mathematical expressions and its designed centered in parallelism (see [3]). In order to reduce the execution time of the algorithms, our implementation benefit from some parallel programming techniques such as *multi-threading*.

As stated earlier, in our implementation we represent a rational map f as its associated normalized irreducible r-homogeneous pair given by the homogeneization operator 2.2 applied to both the numerator and denominator of f. We have seen that this pair induces, through the extended Hopf fibration, the Hopf-endomorphism associated with the rational map f. Also, for each point of the grid (which contains points of the Riemann sphere $P^1(\mathbb{C})$), we consider the normalized homogeneous coordinates of each point. This way, and since the spherical derivative $f^\#$ of f is bounded, our code prevents overflows and indeterminations that could appear in our calculations.

The code containing the exposed algorithm can be found in the following GitHub repository: github.com/LCD [1]. It also provides a user guide and some examples applied to Newton's method, so that the code it is accessible for everyone.

5. Basins of attraction for Chebyshev's method applied to a cubic polynomial

In this last section we will explore and apply the exposed method to the particular case of the rational map induced by Chebyshev's method over the cubic polynomial $p(z) = (z^2 - 1)(z - 1.2866i)$. Of course, given any non linear equation of the form $f(z) = 0$, and given a point $z_0 \in \mathbb{C}$, recall that Chebyshev's method is given by the following recurrent formula $z_{n+1} = C_f(z_n)$, where

$$C_f(z) = z - \left(1 + \frac{1}{2}L_f(z)\right)\frac{f(z)}{f'(z)}, \tag{5.1}$$

where

$$L_f(z) = \frac{f(z)f''(z)}{(f'(z))^2}. \tag{5.2}$$

Note that when we apply Chebyshev's method to the polynomial p, the rational map

$$C_p(z) = \frac{15z^7 - 26\alpha z^6 + 15\alpha^2 z^5 - 6z^5 - 3\alpha^3 z^4 - 9\alpha z^4 + 18\alpha^2 z^3 - z^3 - 6\alpha^3 z^2 + 12\alpha z^2 - 9\alpha^2 z + \alpha^3 - \alpha}{(3z^2 - 2\alpha z - 1)^3}, \tag{5.3}$$

where $\alpha = 1.2866i$, is induced. It is clear that the three roots of p are super-attracting fixed points of C_p, and thus, the Lyapunov constant associated with each of its basins of attraction will be zero. In addition, the induced rational map C_p also presents the attracting 2-cycle $(0.6009i, -0.5640i)$ whose basin of attraction has an associated Lyapunov constant of 0.5093.

As part of the functionalities of the code, we can visualize the basins of attraction in a neighborhood both of the origin and of the infinity point ∞ (on the left and the right of Figure 1, respectively).

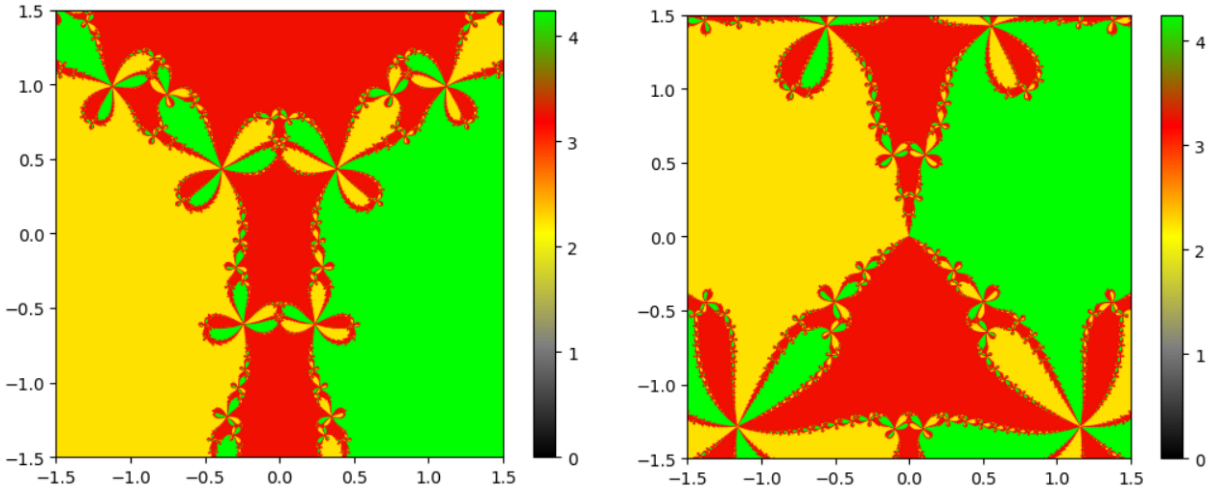


Fig. 1 Basins of attraction induced by Chebyshev's method applied to the polynomial $(z^2 - 1)(z - 1.2866i)$

In these graphics, attending to the color legend, observe that yellow (2), red (3) and green (4) correspond to the basins of attraction of the fixed points of C_p 1, -1, and $1.2866i$ respectively. The color gray (1) corresponds to the basin of attraction of the infinity point ∞, which in this case is a repulsive fixed point of C_p with associated Lyapunov constant 1.8, and the color black (0), corresponds to the basin of the attracting 2-cycle $(0.6009i, -0.5640i)$.

This attracting 2-cycle has been indicated on the Figure 2 as a blue line joining its two elements, in order to better appreciate its presence.

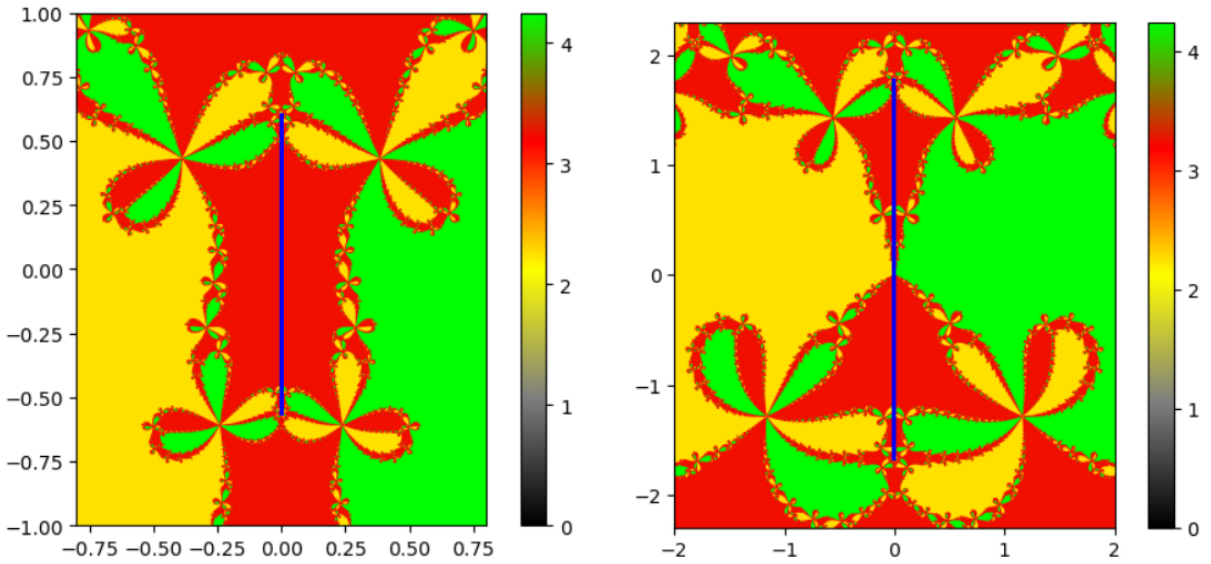


Fig. 2 Attracting 2-cycle appearing in the basins of attraction

We can also visualize the number of iterations that the method required to converge in each point of the considered grid, which is particularly interesting if we focus on a region where part of the basin of the 2-cycle is clearly visible. For that matter, we assign a different color to each number of iterations, so that we can appreciate how such number distributes in the considered region. This graphical information, along with the values of each Lyapunov constant, might be useful when comparing how attractive is each one of the basins.

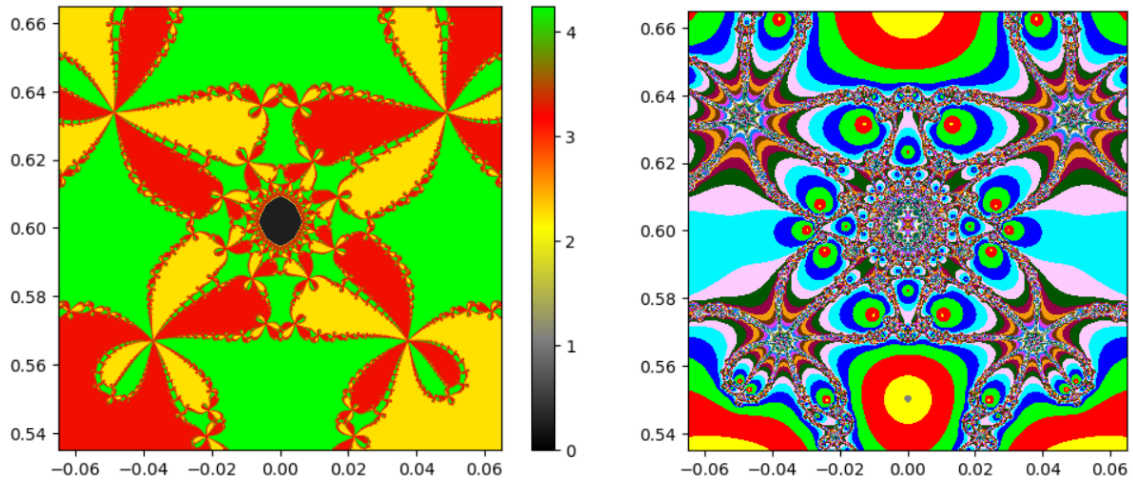


Fig. 3 Number of iterations required to converge

We can see that, in the basin of attraction of the 2-cycle, the number of iterations that were required for the method to end presents clearly chaotic behaviour; that is, it is unstable and extremely dependent on the initial conditions.

Note that the choice of proper values for the maximum number of iterations and the tolerance is an important matter. As the maximum number of iterations allowed increase, greater detail we obtain on the boundary of the

basins, in spite of increasing the execution time of the algorithm. If the considered tolerance is high enough, we might get low precision in the calculations, and if the tolerance is low enough, the algorithm might determine that there are more basins of attraction than they actually are, due to slight numerical errors caused by the precision of the calculations made by the computer. A decrease of the value of the tolerance should always be considered along a significant increase of the maximum number of iterations in order to obtain the expected results. In this case, we have considered 300 as the maximum number of iterations and a tolerance of 10^{-8} to generate the graphics exposed in this section, except for the graphic in which the number of iterations is represented, for which the maximum number of iterations considered was 1000.

A recent study of the super-attracting extraneous fixed points and n-cycles that might appear when applying Chebyshev's method to a cubic polynomial can be found in [6].

Acknowledgements

This research has been funded by project PID2020-118753GB-I00 of the Spanish Ministerio de Ciencia e Innovación.

References

[1] V. Álvarez-Aparicio. Lyapunov Cycle Detector (LCD.jl). 3, 2022.

[2] V. Álvarez-Aparicio, J.M. García-Calcines, L.J. Hernández-Paricio, and M.T. Rivas-Rodr íguez. Algorithms for computing basins of attraction associated with a rational self-map of the Hopf fibration based on Lyapunov exponents. Preprint, 2022.

[3] J. Bezanson, A. Edelman, S. Karpinski, and V.B. Shah. Julia: A fresh approach to numerical computing. *ArXiv*, abs/1411.1607, 2017.

[4] G.D. Birkhoff. Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences of the United States of America*, 17:656-60, 01 1932.

[5] L. DeMarco. Dynamics of rational maps: Lyapunov exponents, bifurcations, and capacity. *Mathematische Annalen*, 326:4373, 04 2003.

[6] J.M. Gutiérrez, and J.L. Varona. Superattracting extraneous fixed points and n-cycles for Chebyshevs method on cubic polynomials. *Qualitative Theory of Dynamical Systems*, 19:123, 2020.

Mathematical modeling and numerical simulation in SisAl project, an innovative pilot for silicon production

Alfredo Bermúdez[1,2], Jorge Albella[3], Óscar Crego[1], José Luis Ferrín[1,2], Branca García[1], Dolores Gómez[1,2,4], Pilar Salgado[1,2]

1. CITMAga, 15782 Santiago de Compostela, Spain

2. Departamento Matemática Aplicada, Universidad de Santiago de Compostela, Spain

3. Departamento Didáctica de la Matemática, Universidad de Santiago de Compostela, Spain

4. mdolores.gomez@usc.es

Abstract

SisAl Pilot is a Horizon 2020-funded project coordinated by the Norwegian University of Science and Technology (NTNU) which comprises 22 partners from 9 countries. The main objective of this project is to demonstrate a patented novel industrial process to produce silicon. The actual carbothermic Submerged Arc Furnace (SAF) process is replaced by a far more environmentally and economically sustainable alternative: the aluminothermic reduction of quartz, which allows using secondary raw materials such as aluminium (Al) EoL scrap and dross, instead of carbon reductant used today. To attain this goal, different types of furnaces are being analysed. Depending on the furnace, the simulations require to study several physical processes strongly coupled: heat transfer, multiphase fluid dynamics, electromagnetism, melting processes or chemical reactions. Thus, the challenge is to carry out numerical simulations based on these models that can support the experimental trials in the plant of the industrial partners. In this talk, we will focus on modelling and simulation of induction furnaces, more precisely in the stirring conditions.

1. Introduction

The main objective of the SisAl project is to demonstrate a novel industrial process to produce silicon, replacing the old carbothermic reduction of quartz with an aluminothermic reduction of quartz in slag using secondary raw materials.

In a simplified description, the quartz (SiO_2) and lime (CaO) mixture is heated until 1500°C to form a slag (this mixture allows to decrease the melting point). Then aluminium is added to the furnace and reacts by reducing the quartz:

$$SiO_2(CaO - SiO_2\, slag) + \frac{4}{3}Al \rightarrow Si + \frac{2}{3}Al_2O_3(CaO - Al_2O_3\, slag) \tag{1.1}$$

Finally, the CaO and Al_2O_3 can be separated from the resulting slag to give a new use to this alumina.

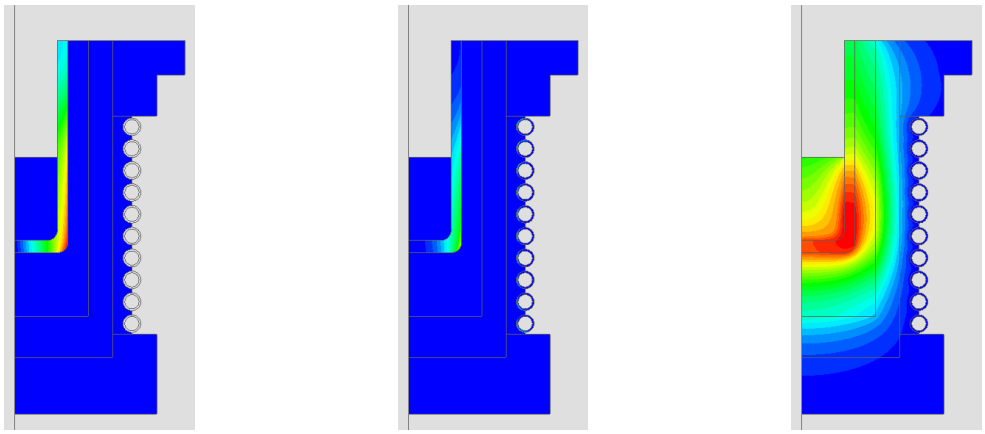
The main advantages of this process are less energy consumption and the avoidance of direct CO_2 emissions. The energy consumption decreases since the slag has a lower melting point than the quartz and the aluminothermic reduction is an exothermic reaction. Thus, SisAl represents an environmentally and economically sustainable alternative to today's carbothermic reduction process in the Submerged Arc Furnace (SAF), allowing Si production in an increasingly carbon-lean Europe.

Different furnaces are involved in the SisAl process: induction furnaces, submerged arc furnaces and rotary furnaces. In this work, we are focused on the induction furnace. Since the mixing of the materials inside the furnace can be crucial during the industrial process, two different standard stirring tools were analysed. In particular, the goal here is to determine the best stirring condition by means of the numerical solution of a hydrodynamic problem. Notice that, for confidentiality reasons, the numerical scales have been removed from all figures included in this document.

2. Induction furnace

An induction heating furnace essentially consists of one helical coil surrounding a crucible containing the load to melt. When the charge pretended to melt is a non-conducting material, a conductor crucible (made of graphite and insulated by a refractory) is used to heat the load by conduction. Thus, alternating current passing through the coil (which is water-cooled to avoid overheating) induces a rapidly oscillating magnetic field which generates eddy currents in the conductor crucible (Fig. 1a). Then, the crucible is heated due to the Joule effect (the dissipated power is shown in Fig. 1b). Thus, the temperature of the system increases (Fig. 1c), the slag is heated by conduction and finally, it melts.

The standard procedure to model an induction heating furnace with a cylindrical crucible is to consider it in an axisymmetric setting (the fields do not depend on the azimuthal component and the current produced only has



(a) Current density **(b)** Dissipated power **(c)** Temperature

Fig. 1 Numerical results on a cross-section of a fully loaded induction furnace after 5 hours of operation.

azimuthal component). This axisymmetric model does not include the detailed geometry of the coil, being the helical coil approximated by toroidal rings. Since it is a standard modelling in the field, the description of these models is not included in this document (for a detailed explanation the reader can consult [1], [2], [3]).

Regarding the numerical results shown in Fig. 1, they were obtained using the commercial software Altair Flux2D which implements a numerical resolution of a thermo-electromagnetic problem using FEM. In this first stage, the main goal is to analyse the melting point of the slag under different operation conditions. Notice that at this level the motion of the slag is negligible as it is a very bad electrical conductor. However, one important condition in this industrial process is the homogenisation of the mixture once the aluminium is added to the mixture, which motivates the following section.

3. Stirring conditions

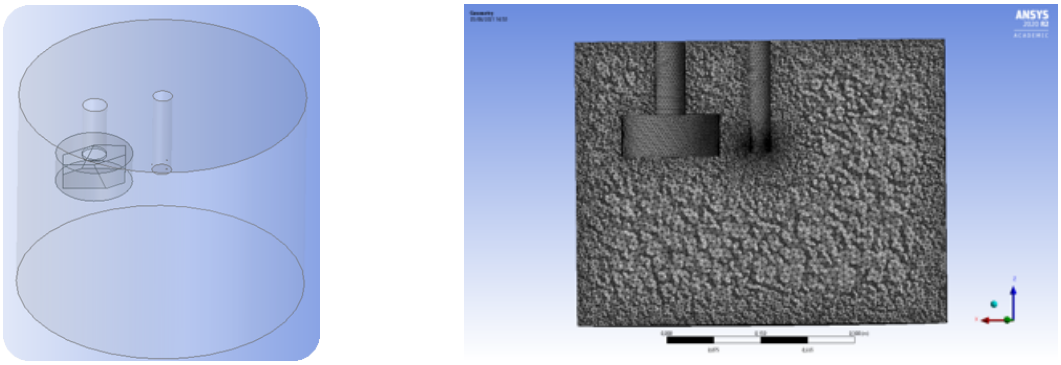
In this section, a model is presented to analyse the Slag-Al mixing (once melted) using two different stirring conditions: a mechanical rotor and a nitrogen injection lance. Thus, we simulated the mixing inside the furnace between two layers of melted materials. The chemical reaction will not be considered but only the fluid dynamic simulation of the stirring process. Fig. 2a shows the shape of the mechanical rotor and the lance considered.

Let us describe the model used to simulate the mixing between two layers of melted materials. We considered the incompressible Reynolds-Averaged Navier Stokes equations with the $k-\omega$ SST turbulence model to capture the movement of the melted materials:

$$\begin{cases} \operatorname{div}\mathbf{v} = 0, \\ \dfrac{\partial}{\partial t}(\rho\mathbf{v}) + \operatorname{div}(\rho\mathbf{v}\otimes\mathbf{v}) = -\operatorname{grad}\pi + \operatorname{div}\left(\mu_{eff}\left(\operatorname{grad}\mathbf{v} + \operatorname{grad}\mathbf{v}^T\right)\right) + \mathbf{f} - \operatorname{div}(\rho k I), \\ \dfrac{\partial}{\partial t}(\rho k) + \operatorname{div}(\rho k\mathbf{v}) = \operatorname{div}(\Gamma_k \operatorname{grad} k) + f_k, \\ \dfrac{\partial}{\partial t}(\rho\omega) + \operatorname{div}(\rho\omega\mathbf{v}) = \operatorname{div}(\Gamma_\omega \operatorname{grad}\omega) + f_\omega, \end{cases} \tag{3.1}$$

where $\mathbf{v}$ denotes the velocity field, ρ the mass density, π the pressure and Γ_k, Γ_ω the effective dissipation of the turbulent variables. These models are coupled via the effective turbulent viscosity μ_{eff} which depends on the turbulent variables, namely, the turbulent kinetic energy, k, and the specific dissipation rate, ω (see [4] for more information). Besides, the different fluids have to be identified and a Volume of Fluid (VOF) model was chosen as the multiphase model to this end. This strategy relies on the fact that two or more fluids are not interpenetrating. Thus, the volume fraction of the slag, α_{slag}, is introduced. The value of α_{slag} is interpreted as the fraction of the cell occupied by the slag. Then, the volume fraction of the aluminium is readily obtained as $\alpha_{Al} = 1 - \alpha_{slag}$. Values other than 0 or 1 for the volume fraction represent a mixture of the fluids and indicate that the free surface is located inside the corresponding cell. The volume fraction verifies the following continuity equation:

$$\frac{\partial}{\partial t}\left(\alpha_{slag}\rho_{slag}\right) + \operatorname{div}\left(\alpha_{slag}\rho_{slag}\mathbf{v}\right) = 0,$$



(a) Three-dimensional sketch

(b) Cross-cut of the tetrahedral mesh

Fig. 2 Geometry (on the left) and mesh (on the right) of a generic induction furnace.

since the mass transfer between fluids is not considered. The models are coupled via the material properties through the volume fraction:

$$\rho = \rho_{slag}\alpha_{slag} + \rho_{Al}(1 - \alpha_{slag}),$$
$$\mu = \mu_{slag}\alpha_{slag} + \mu_{Al}(1 - \alpha_{slag}).$$

Once a model capable of simulating the motion and interaction of both materials has been written, the stirring conditions can be considered.

On the one hand, the gas introduced through the lance is supposed to break into bubbles. Thus, the nitrogen injection is studied using a discrete phase model where the trajectory of the particles (these bubbles) is predicted by integrating the balance force, i.e., the particle inertia with the forces acting on the particle:

$$m_p \frac{d\mathbf{v}_p}{dt} = m_p \frac{\mathbf{v} - \mathbf{v}_p}{\tau_r} + m_p g \frac{\rho_p - \rho}{\rho_p}\mathbf{e}_3,$$

where the subscript p denotes the properties and fields relative to the particle; m denotes the mass and τ_r is the droplet or particle relaxation time that depends on their diameter.

On the other hand, the rotor stirring is simulated by considering the rotational movement of the mechanical rotor. Notice that the domain of the problem studied depends on time. The direct approach of creating a new geometry (and its corresponding mesh) per time step is prohibitive from a computational point of view. Thus, the stirring by a mechanical rotor is simulated by taking a subdomain around the rotor and imposing a rotation of this subdomain. In this way, instead of creating a new geometry (and mesh) the subdomain is rotated per time step and the contact between the subdomain around the rotor and the rest of the geometry is identified using a sliding mesh technique.

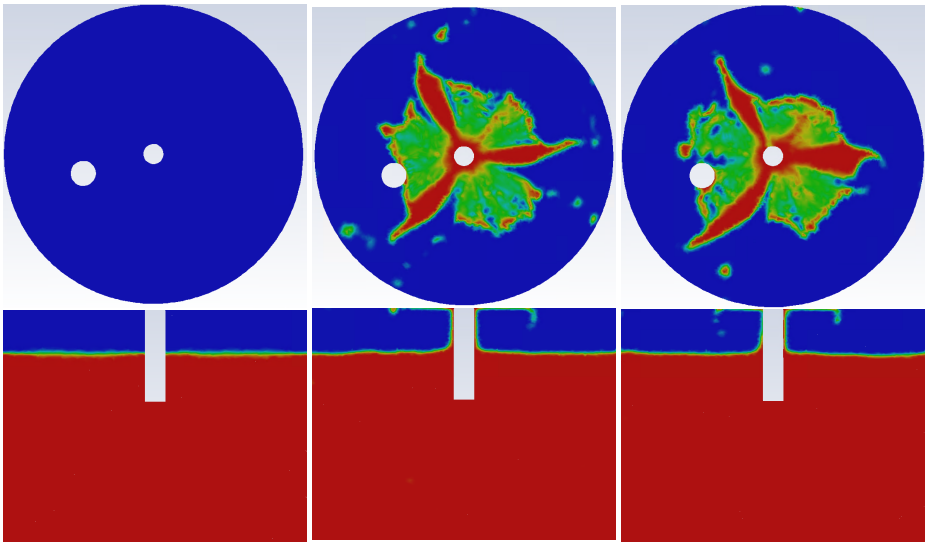
Regarding the numerical results of the stirring analysis that can be found in the next section, they were obtained using the commercial software ANSYS Fluent which uses a cell-centered finite volume method to solve the above equations.

3.1. Numerical results

The two different techniques discussed were analysed at specific operating conditions: the stirring produced by a mechanical rotor whose axis of rotation is off-centred, see Fig. 2a, with a rotation speed of 80 rpm and the one produced by a gas injection using a lance in the centre of the furnace with a nitrogen flow of 10 Nl/min. In this way, to obtain the velocity field in the mixture, some fixed conditions of temperature and pressure are imposed: $T = 1650°\text{C}$ and $\pi = 1\text{atm}$. Thus, the motion produced by these stirring techniques and their corresponding combination in a mixture of two immiscible liquids, Al and SiO_2-CaO slag, are simulated; see Fig. 3 which shows the volume fraction in cross sections (the top of the mixture and a transversal cut) giving an idea of the mixing between the aluminium and the SiO_2-CaO slag.

The nitrogen injection assumes breaking into bubbles which requires certain information related to the diameter of the bubbles. To investigate the importance of this data, we have performed different tests where we consider different distributions of the diameter of the bubbles:

- Test 1: uniform distribution of the diameter of the bubbles, fixed at is 2 mm which is the diameter of the nitrogen inlet in the lance; see Fig. 3.



(a) Only rotor **(b)** Only injection **(c)** Rotor and injection

Fig. 3 Volume fraction in cross sections using different stirring conditions: only rotor, only injection and both. In the nitrogen injection, the diameter distribution of bubbles is considered uniform.

- Test 2: the diameter of the bubbles follows a Rosin-Ramler distribution with a maximum of 2 mm, a minimum of 0.05 mm and a mean value of 0.5 mm and a spread parameter of 2; see Fig. 4.

Based on the results in Figs. 3-4, the injection of nitrogen is the more effective of the two techniques. Furthermore, their combination shows that the rotor influences the procedure. Besides, the analysis of the bubble distribution reveals the importance of the bubble size, which motivates the simulation of the formation of bubbles that is presented below.

4. Bubble formation

In this sense, we have analysed the formation of bubbles for 2D and 3D reduced problems. Fig. 5 shows the flux direction of the nitrogen with green arrows in the original geometry and the different approximations: the lateral simplification, Fig. 5b, assumes a continuous lateral inlet instead of three holes, the axial simplification, Fig. 5c, assumes a unique hole in the centre of the lance and the three-dimensional approximation, Fig. 5d, is a wedge of 120 degrees corresponding with one of the original holes assuming symmetry with the other two parts. Following this strategy, the lateral and axial simplifications can be approximated by bidimensional problems under symmetric assumptions. Finally, to simplify the resolution the aluminium was discarded in this study. Thus, the model presented in Sect. 3 can be used by simply removing the discrete phase model and changing the aluminium considered as an incompressible fluid for the nitrogen considered as an ideal gas. So the incompressibility condition in (3.1) is changed by the mass conservation equation:

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{v}) = 0,$$

and the corresponding equation of state for the nitrogen is introduced:

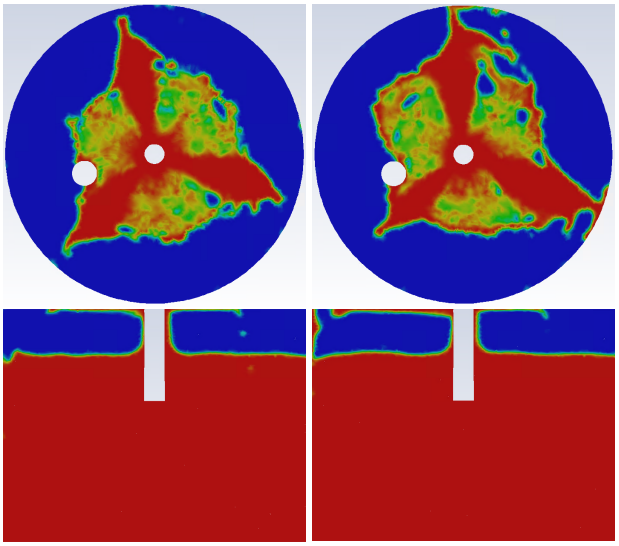
$$\pi V = nRT$$

where V denotes volume, n amount of substance of the gas and R the universal gas constant.

Figs. 6-7 show the volume fraction for the different approaches. Notice that the bidimensional approximations are not realistic. But anyway, the result of all the cases is very similar and the formation of small bubbles is not observed, just one big bubble.

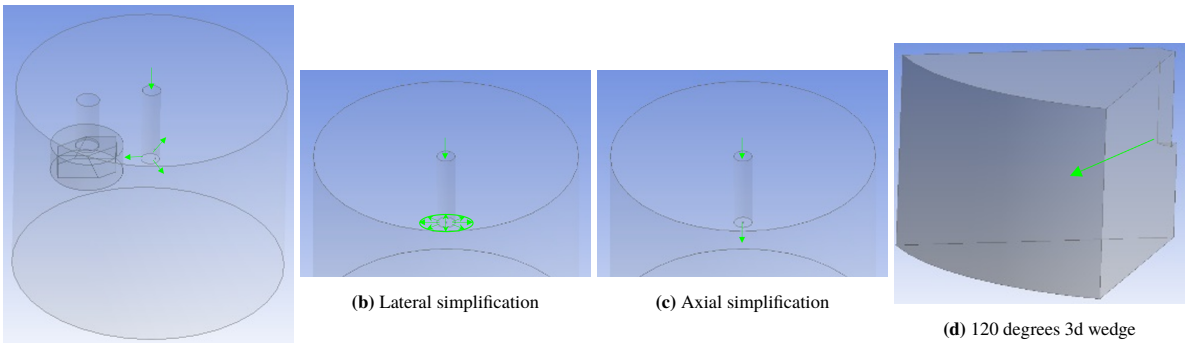
5. Conclusions

We have analysed two different stirring techniques that can be found in industrial furnaces: mechanical rotor and gas injection. First, we have simulated the mixing considering only the rotor and only the injection.The results suggest that gas injection is more efficient for stirring. However, the simulation of the injection requires information concerning the sizes of nitrogen bubbles. Thus, in the second approach, we have performed a sensitivity analysis that shows that a better knowledge of nitrogen bubbles is crucial.



(a) Only injection **(b)** Rotor and injection

Fig. 4 Volume fraction in cross sections using different stirring conditions: only injection and both (rotor and injection). In the nitrogen injection, the diameter distribution of bubbles is considered with a Rosin-Ramler distribution.



(a) Original geometry **(b)** Lateral simplification **(c)** Axial simplification **(d)** 120 degrees 3d wedge

Fig. 5 Different geometrical approximations to study the formation of bubbles. The rotor is not considered in any of the cases.

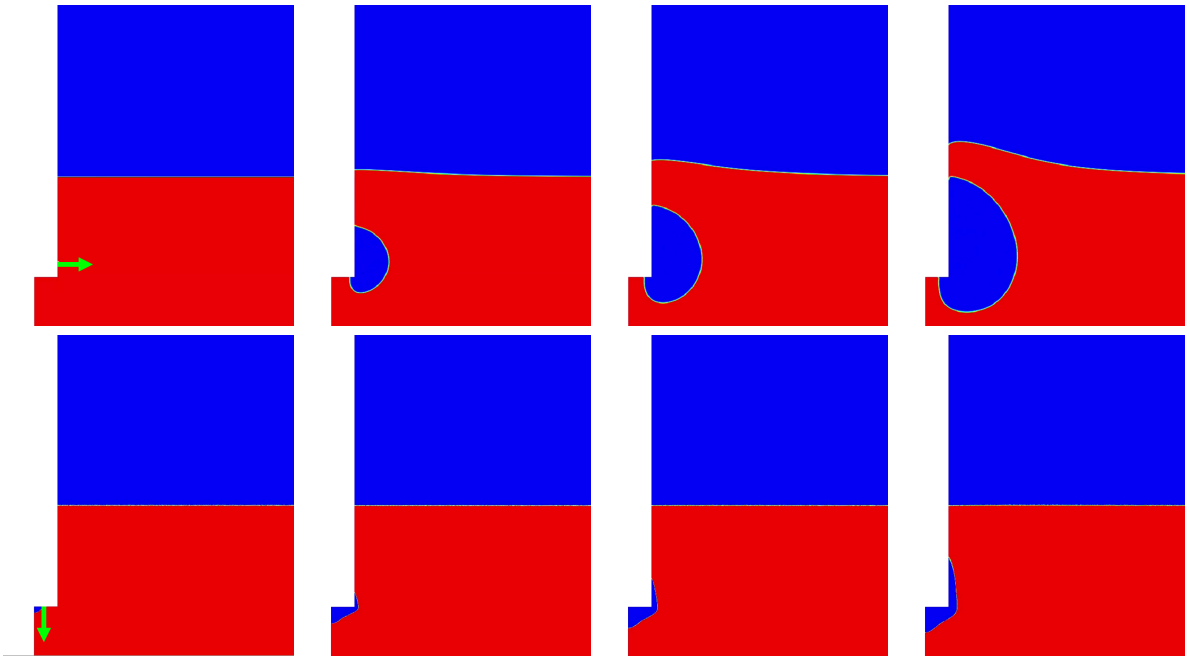


Fig. 6 Volume fraction at different time steps in the bidimensional approximations: top lateral approximation and bottom axial approximation. In the initial time step the green arrow marks the direction of injection.

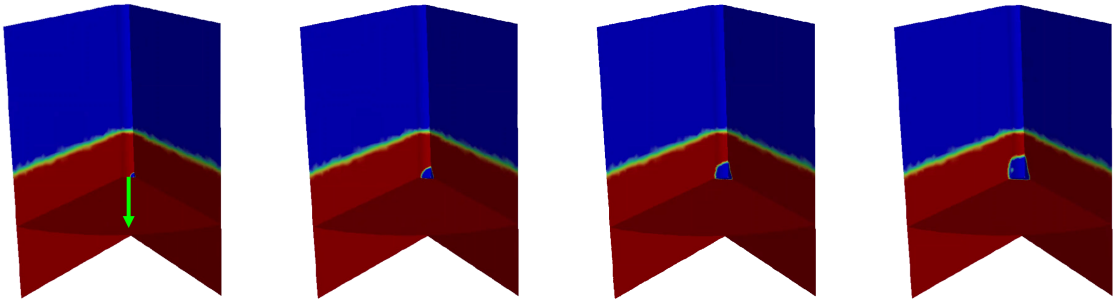


Fig. 7 Volume fraction at different time steps in the tridimensional wedge approximation. In the initial time step the green arrow marks the direction of injection.

In this sense, in the third approach, we have analysed the formation of bubbles for 2D and 3D reduced problems. All simulations are coherent with each other and the formation of small bubbles is not found. What we observe is the formation of a unique bubble that grows until it reaches the nitrogen on the top pushing the slag and forming a wave.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869268. This work has also received financial support from the Xunta de Galicia (2021 GRC GI-1563 - ED431C 2021/15).

References

[1] Alfredo Bermúdez, Dolores Gómez and Pilar Salgado. Mathematical models and numerical simulation in electromagnetism. *Springer*, 2013.

[2] Alfredo Bermúdez, Dolores Gómez, María del Carmen Muñíz and Pilar Salgado. Transient numerical simulation of a thermoelectrical problem in cylindrial induction heating furnaces. *Adv. Comput. Math.*, 26: 39-62, 2007.

[3] Alfredo Bermúdez, Dolores Gómez, María del Carmen Muñiz, Pilar Salgado and Rafael Vázquez. Numerical simulation of a thermo-electromagneto-hydrodynamic problem in an induction heating furnace. *Appl. Numer. Math.*, 59: 2082-2104, 2009.

[4] ANSYS, Inc. ANSYS Fluent, Release 21.0, Theory Guide. ANSYS, Inc.

Physically-Based Reduced-Order Battery Models Including Degradation for Real-Time Control Applications

David Aller Giráldez[3], Alfredo Bermúdez[1,2], David Casasnovas González[3], Manuel Cremades-Buján[1,2], Juan Nicolás Aguado[3], Jerónimo Rodríguez[1,2]

1. Departamento de Matemática Aplicada, Universidade de Santiago de Compostela. Spain
2. CITMAga. Centro de Investigación e Tecnoloxía Matemática de Galicia, Spain
3. Repsol Techonlogy Center, Spain

Abstract

Lithium-ion batteries represent one of the most widely used energy storage devices in both mobile and stationary applications. Its correct operation depends largely on the so-called Battery Management Systems (BMS) that use mathematical models to be able to predict in real time the state of charge, the state of health and the state of function of the battery in order to make a decision. Although equivalent circuit models [4] are often used for this purpose, nowadays advanced BMS incorporate more complex electrochemical models [3] but handling battery degradation is still a challenge. Since these models need to be solved faster than real time, it is neccesary to apply specific order reduction techniques [6]. In this document, a physic-based reduced order model compatible with real-time control applications is developed, which correctly captures the space-localized lithium plating [1] phenomena and porosity decreasing [7] with ageing.

1. Introduction

A lithium ion battery cell is made of a stack of several electrodes where lithium ions can be stored, each pair of them separated by an electrical insulator domain, the so-called separator, which prevents short circuit.Both the electrodes and the separator are porous materials filled with an ionic liquid, the so-called electrolyte, so lithium ions can travel from one electrode to the opposite throught the electrolyte. Each of the electrodes is made up of inactive materials devoted to increase the electrode conductivity and stability and active materials where lithium ion can be intercalated (lithiation) and deintercalated (delithiation). Undesired side reactions can take place reducing the expected life of the battery cell, being solid-electrolyte interface formation and lithium plating the most prominent ones.

In order to develop an advanced battery management system (BMS) able to avoid or reduce this degradation mechanisms a computationally efficient while accurate battery model is required. While equivalent circuit models are the state-of-the-art for BMS applications, their empirical nature make degradation mechanisms difficult to implement. Another alternative, the single-particle models, as they are based on averaging the model equations in space suffer the disadvantage of not being able to capture with accuracy the lithium plating, as it is an space localized phenomena. Hence, the Doyle-Fuller-Newman porous electrode model is the simplest battery model that fulfill our requirements. As those kind of models are computationally expensive, a reduced order version is needed in order to use it on real-time control algorithms and battery management systems applications.

2. Full order model

The Doyle-Fuller-Newman porous electrode battery model (a comprehensible derivation from electrochemical principles can be found on [4]) is comprised by a coupled set of parabolic partial differential equations, elliptic partial differential equations and non-linear algebraic equations defined on two different scales:

- A macroscale, where the ionic potential ϕ_e and the lithium ion concentration within the electrolyte c_e are defined on the electrolyte domain and where the electric potential ϕ_s is defined in the electrodes.
- A microscale, where lithium concentration (assuming spherical symmetry) c_s is defined in the active material particles.

Both scales are coupled by the lithium intercalation reactions j_{int} taking place on the surface of the electrode particles in contact with the electrolyte. In Figure 1 a schema of the domains involved on the model can be consulted. Prior to the spatial discretization we perform an domain decomposition strategy so each continuous variable is split in its negative electrode (a), separator (s) and positive electrode (c) part, including continuity and flux continuity as additional equations.

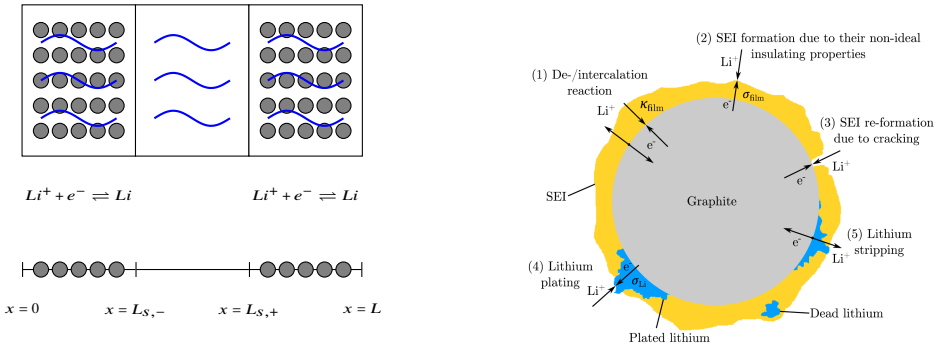


Fig. 1 Left: Schema of a Li-ion cell and computational domain of the porous electrode model. Right: Illustration of the electrochemical reactions on the graphite particle surface (picture extracted from J. Keil et al. [1])

2.1. Porous electrode model without degradation

Lithium ion concentration in the electrolyte is modelled by the parabolic partial differential equations:

$$\frac{\partial(\epsilon_e^l c_e^l)}{\partial t} = \frac{1}{L_l^2}\frac{\partial}{\partial x}\left(D_{e,eff}^l \frac{\partial c_e^l}{\partial x}\right) + \frac{1-t_+}{F} j_{tot}^l, \quad x \in (0,1), \quad t > 0, \quad l \in \{a,s,c\}, \tag{2.1}$$

with initial condition $c_e^l\big|_{t=0} = c_{e,ini}$ and where we are assuming that the total reaction flux (which will be introduced later) in the separator is null, $j_{tot}^s \equiv 0$ and that no ions are allowed to leave the cell by the current collectors, i.e:

$$\left.\frac{\partial c_e^a}{\partial x}\right|_{x=0} = 0, \quad \left.\frac{\partial c_e^c}{\partial x}\right|_{x=1} = 0, \quad t > 0. \tag{2.2}$$

Additionally, the equations are completed with transmission conditions on the electrodes / separator interface:

$$c_e^a\big|_{x=1} = c_e^s\big|_{x=0}, \quad \frac{1}{L_a} D_{e,eff}^a \left.\frac{\partial c_e^a}{\partial x}\right|_{x=1} = \frac{1}{L_s} D_{e,eff}^s \left.\frac{\partial c_e^s}{\partial x}\right|_{x=0}, \quad t > 0, \tag{2.3}$$

$$c_e^s\big|_{x=1} = c_e^c\big|_{x=0}, \quad \frac{1}{L_s} D_{e,eff}^s \left.\frac{\partial c_e^s}{\partial x}\right|_{x=1} = \frac{1}{L_c} D_{e,eff}^c \left.\frac{\partial c_e^c}{\partial x}\right|_{x=0}, \quad t > 0. \tag{2.4}$$

Lithium concentration within the particles of both electrodes is modelled by Fick's law, so it take the form of a parabolic partial differential equation:

$$R_{p,l}^2 \frac{\partial c_s^l}{\partial t} = \frac{1}{r^2}\frac{\partial}{\partial r}\left(D_s^l r^2 \frac{\partial c_s^l}{\partial r}\right), \quad r \in (0,1), \quad x \in (0,1), \quad t > 0, \quad l \in \{a,c\}, \tag{2.5}$$

with initial condition $c_s^l\big|_{t=0} = c_{s,ini}^l$ and boundary conditions

$$\left.\frac{\partial c_s^l}{\partial r}\right|_{r=0} = 0, \quad -\frac{D_s^l}{R_{p,l}} \left.\frac{\partial c_s^l}{\partial r}\right|_{r=1} = \frac{j_{int}^l}{a_s^l F}, \quad x \in (0,1), \quad t > 0, \quad l \in \{a,c\}, \tag{2.6}$$

that come from the spherical symmetry and the lithium intercalation reaction taking place on the particles surface. The ionic potential in the electrolyte is modelled by the elliptic partial differential equations:

$$-\frac{1}{L_l^2}\frac{\partial}{\partial x}\left(\kappa_{eff}^l \frac{\partial \phi_e^l}{\partial x}\right) - \frac{1}{L_l^2}\frac{\partial}{\partial x}\left(\kappa_{D,eff}^l \frac{1}{c_e^l}\frac{\partial c_e^l}{\partial x}\right) = j_{tot}^l, \quad x \in (0,1), \quad t \geq 0, \quad l \in \{a,s,c\} \tag{2.7}$$

with homogeneous boundary conditions as the ions are not allowed to leave the cell by the current collectors, i.e:

$$\left.\frac{\partial \phi_e^a}{\partial x}\right|_{x=0} = 0, \quad \left.\frac{\partial \phi_e^c}{\partial x}\right|_{x=1} = 0, \quad t \geq 0. \tag{2.8}$$

Transmission conditions between the electrodes and separator reads as follows:

$$\phi_e^a\big|_{x=1} = \phi_e^s\big|_{x=0}, \; \frac{1}{L_a}\left(\kappa_{eff}^a \frac{\partial \phi_e^a}{\partial x} + \kappa_{D,eff}^a \frac{\partial \ln c_e^a}{\partial x}\right)\Big|_{x=1} = \frac{1}{L_s}\left(\kappa_{eff}^s \frac{\partial \phi_e^s}{\partial x} + \kappa_{D,eff}^s \frac{\partial \ln c_e^s}{\partial x}\right)\Big|_{x=0}, \quad t \geq 0, \tag{2.9}$$

$$\phi_e^s\big|_{x=1} = \phi_e^c\big|_{x=0}, \; \frac{1}{L_s}(\kappa^s_{eff}\frac{\partial \phi_e^s}{\partial x} + \kappa^s_{D,eff}\frac{\partial \ln c_e^s}{\partial x})\Big|_{x=1} = \frac{1}{L_c}(\kappa^c_{eff}\frac{\partial \phi_e^c}{\partial x} + \kappa^c_{D,eff}\frac{\partial \ln c_e^c}{\partial x})\Big|_{x=0}, \quad t \geq 0. \tag{2.10}$$

The electric potential in the electrodes is modelled by the elliptic partial differential equations:

$$\frac{\partial}{\partial x}\left(\sigma^l_{eff}\frac{\partial \phi_s^l}{\partial x}\right) = j^l_{tot}, \quad x \in (0,1), \quad t \geq 0, \quad l \in \{a, c\}, \tag{2.11}$$

with the following boundary conditions:

$$-\sigma_a^{eff}\frac{\partial \phi_s^a}{\partial x}\Big|_{x=0} = \frac{I_{app}}{A}, \quad \frac{\partial \phi_s^a}{\partial x}\Big|_{x=1} = 0, \quad \sigma_c^{eff}\frac{\partial \phi_s^c}{\partial x}\Big|_{x=0} = \frac{I_{app}}{A}, \quad \frac{\partial \phi_s^c}{\partial x}\Big|_{x=1} = 0, \quad t \geq 0. \tag{2.12}$$

The intercalation reactions taking place on particle surfaces are modelled by the algebraic equations

$$j^l_{int} = 2a^l_s i^l_{0,int}\sinh\left(\frac{1}{2}\frac{F}{RT}\eta^l_{int}\right), \quad x \in (0,1), \quad t \geq 0, \quad l \in \{a, c\}, \tag{2.13}$$

where

$$\eta^l_{int} = \phi^l_s - \phi^l_e - R^l_{i,film}\frac{j^l_{int}}{a^l_s} - U^l(c^l_{s,sur}/c^l_{s,max}), \quad x \in (0,1), \quad t \geq 0, \quad l \in \{a, c\}, \tag{2.14}$$

are the intercalation overpotentials and $R^l_{i,film} = \frac{\delta^l_{film}}{\kappa_{film}}$ the ionic resistances due to the degradation layer.

2.2. Degradation model

The degradation model is based on the one presented on J. Keil et al. [1] but some modifications and simplifications have been made. In this work, solid -electrolyte interface re-formation due to cracking and lithium stripping are omitted. In Figure 1 right one can see where this degradation mechanisms are located on the particle/film surface and their associated conductivities.

Plating flux in the negative electrode follows the (distributed) algebraic equation

$$j_{lpl} = \begin{cases} 2a^a_s i_{0,lpl}\sinh(\alpha_{lpl}\frac{F}{RT}\eta_{lpl}) & \eta_{lpl} \leq 0 \\ 0 & \eta_{lpl} > 0 \end{cases} \quad x \in (0,1), \quad t \geq 0, \tag{2.15}$$

where

$$\eta_{lpl} = \phi^a_s - \phi^a_e - U_{lpl}, \quad x \in (0,1), \quad t \geq 0 \tag{2.16}$$

is the lithium plating overpotential.

Solid-electrolyte interface flux in the negative electrode follows the (distributed) algebraic equation

$$j_{sei} = a^a_s i_{0,sei}\exp(-\alpha_{sei}\frac{F}{RT}\eta_{sei}), \quad x \in (0,1), \quad t \geq 0, \tag{2.17}$$

where

$$\eta_{sei} = \phi^a_s - \phi^a_e - R^a_{e,film}\frac{j^a_{int}}{a^a_s} - U_{sei}, \quad x \in (0,1), \quad t \geq 0 \tag{2.18}$$

is the solid-electrolyte interface formation overpotential and $R^a_{e,film} = \frac{\delta^a_{film}}{\sigma_{film}}$ is the electronic resistance due to the degradation layer.

Remark 2.1 Notice that there are two big differences between the growths of the solid-electrolyte interface and the lithium plating. The first one is that the overpotential of the solid-electrolyte interface has a resistive term that involves its associated flux, and the second one is that lithium plating reaction is strongly dependent on charge/discharge behaviour as opposed to the solid-electrolyte interface which will grow not only in charge, but also (slowly) in discharge and relaxation.

The model is closed by introducing the total reaction flux which is a sum of the intercalation and degradation fluxes:

$$j^a_{tot} = j^a_{int} + j_{sei} + j_{lpl}, \quad j^c_{tot} = j^c_{int}, \quad x \in (0,1), \quad t \geq 0. \tag{2.19}$$

Lastly, it remains to model how the degradation film and the anode porosity evolves with ageing.

Porosity and film thickness evolution

Film thickness of the negative electrode follows the (distributed) ordinary differential equation

$$\begin{aligned} \frac{\partial \delta^a_{film}}{\partial t} &= \frac{M_{sei}}{F\rho_{sei}} \frac{j_{sei}}{a^a_s} + \frac{M_{Li}}{F\rho_{Li}} \frac{j_{lpl}}{a^a_s}, \quad x \in (0,1),\ t > 0, \\ \delta^a_{film}\Big|_{t=0} &= \delta^a_{film,0}, \end{aligned} \tag{2.20}$$

and anode porosity evolution is related to this film thickness by

$$\begin{aligned} -\frac{\partial \epsilon^a_e}{\partial t} &= \frac{M_{sei}}{F\rho_{sei}} j_{sei} + \frac{M_{Li}}{F\rho_{Li}} j_{lpl}, \quad x \in (0,1),\ t > 0, \\ \epsilon^a_e\big|_{t=0} &= \epsilon^a_{e,0}. \end{aligned} \tag{2.21}$$

This last relationship is extracted from X-G. Yang et al. [7] as in J. Keil et al. [1] porosity evolution is not modelled.

2.3. Numerical implementation

We approximate the microscale using Legendre polynomials

$$c^l_s(x,r,t) \approx \sum_{i=0}^{N_l} c^l_{s,i}(x,t)\phi^l_{2i}(r), \quad l \in \{a,c\}$$

where only even degrees are considered due to the spherical symmetry assumption. Then, we build the weak formulation for the coefficients through a standard Galerkin projection:

$$\sum_{i,j=0}^{N_l} M^{c^l_s}_{ij} \int_0^1 \frac{\partial c^l_{s,i}}{\partial t} \bar{c}^l_{s,j} dx + \sum_{i,j=0}^{N_l} \frac{D^l_s}{R^2_{p,l}} K^{c^l_s}_{ij} \int_0^1 c^l_{s,i} \bar{c}^l_{s,j} dx + \sum_{j=0}^{N_l} \frac{1}{R_{p,l} a^l_s F} \phi^l_{2j}(1) \int_0^1 j^l_{int} \bar{c}^l_{s,j} dx = 0, \quad l \in \{a,c\},$$

where

$$M^{c^l_s}_{ij} = \frac{1}{2} \int_{-1}^1 r^2 \phi^l_{2i}(r)\phi^l_{2j}(r) dr, \quad K^{c^l_s}_{ij} = \frac{1}{2} \int_{-1}^1 r^2 \frac{d\phi^l_{2i}}{dr}(r) \frac{d\phi^l_{2j}}{dr}(r) dr.$$

The weak formulation for the complete mixed system of equations is derived and discretized in space using the finite element method. P1 finite elements are used and the implementation is carried on the FEniCS library [2]. After the space discretization we arrive to a differential-algebraic system of the form:

$$\begin{aligned} M\frac{dx}{dt}(t) &= f(t, x(t)), \quad t_0 < t \le t_f, \\ x(t_0) &= x_0, \end{aligned} \tag{2.22}$$

where M is a singular matrix due to the presence of elliptic and algebraic equations. The system is discretized in time using the implicit Euler method and the resulting non-linear systems are solved by Newton's method.

3. Reduced order model

In order to meet the requirements of real-time control applications the previous model is not suitable, not only due to its computational cost but also to non-linearities which complicates the application of filtering techniques. In this section we will deduce a reduced-order model in *state-space* form, i.e:

$$\begin{aligned} x_{k+1} &= Ax_k \quad + Bu_k \quad , \quad k = 0, 1, \dots \\ y_{k+1} &= Cx_{k+1} + Du_{k+1} \end{aligned} \tag{3.1}$$

where $x(\cdot) \in \mathbb{R}^n$ represents the model state, $y(\cdot) \in \mathbb{R}^q$ the model output and $u(\cdot) \in \mathbb{R}^p$ the model input. The matrix $A \in \mathbb{R}^{n\times n}$ represents the state matrix, $B \in \mathbb{R}^{n\times p}$ the input matrix, $C \in \mathbb{R}^{q\times n}$ the ouptut matrix and $D \in \mathbb{R}^{q\times p}$ the feedforward matrix. To do this, we will use the discrete realization time algorithm (DRA) methodology firstly introduced in G. Plett book [4]. This methodology can be summarized in five main steps:

1. Linearize the equations which involves some variable transformations.

2. Compute a function, the so-called *transfer function* that maps the Laplace transform of the input into the Laplace transform of the output

$$H(s) = \frac{Y(s)}{U(s)} = \frac{\mathcal{L}(y(t))}{\mathcal{L}(u(t))}, \quad s \in \mathbb{C},$$

where by capital letters variables we denote the Laplace transform of lowercase letters variables, i.e:

$$Y(s) := \mathcal{L}(y(t))(s) = \int_0^\infty y(t)e^{-st}dt.$$

3. Compute the Markov parameters (response of the system to an unit pulse) for that transfer function.
4. Apply the Ho-Kalman algorithm to compute the matrices defining the state-space model.
5. Revert the variable transformations to recover the variables of interest, cf. Table 1.

Remark 3.1 In [5] the authors apply the Ho-Kalman algorithm for all the variables at once, while here we will apply the realization algorithm for each variable alone and then concatenate the matrices to build the state-space system.

3.1. Transfer functions for the porous electrode model without degradation

In this section we will give an idea about how the transfer function for the electrolyte concentration is derived. We refer to A. Rodríguez et al. [5] for the full derivation of all the involved transfer functions.

Output		Nonlinear correction
$\widetilde{C}_e^a$	Li^+ concentration in the liquid phase	$c_e^a(z,t) = \widetilde{c}_e^a(z,t) + c_{e,0}$
$\widetilde{C}_e^s$	Li^+ concentration in the liquid phase	$c_e^s(z,t) = \widetilde{c}_e^s(z,t) + c_{e,0}$
$\widetilde{C}_e^c$	Li^+ concentration in the liquid phase	$c_e^c(z,t) = \widetilde{c}_e^c(z,t) + c_{e,0}$
$\widetilde{C}_{s,sur}^{a,\star}$	Li surf. concentration in the solid phase	$c_{s,sur}^a(z,t) = \widetilde{c}_{s,sur}^{a,\star}(z,t) + c_{s,0}^a$
$\widetilde{C}_{s,sur}^{c,\star}$	Li surf. concentration in the solid phase	$c_{s,sur}^c(z,t) = \widetilde{c}_{s,sur}^{c,\star}(z,t) + c_{s,0}^c$
J_{int}^a	Intercalation flux	$j_{int}^a(z,t) = \widetilde{j}_{int}^a(z,t)$
J_{int}^c	Intercalation flux	$j_{int}^c(z,t) = \widetilde{j}_{int}^c(z,t)$

Output		Nonlinear correction
$\widetilde{\Phi}_{s-e}^{a,\star}$	Potential diff. between solid and liquid phase	$\phi_{s-e}^a(z,t) = \widetilde{\phi}_{s-e}^{a,\star}(z,t) + U^a(\frac{c_{s,avg}^a(t)}{c_{s,max}^a})$
$\widetilde{\Phi}_{s-e}^{c,\star}$	Potential diff. between solid and liquid phase	$\phi_{s-e}^c(z,t) = \widetilde{\phi}_{s-e}^{c,\star}(z,t) + U^a(\frac{c_{s,avg}^a(t)}{c_{s,max}^a})$
$\widetilde{\Phi}_e^a$	Liquid phase potential	$\phi_e^a(z,t) = \widetilde{\phi}_{e,a}(z,t) - \widetilde{\phi}_{s-e}^a(0,t) - U^a(\frac{c_{s,avg}^a(t)}{c_{s,max}^a})$
$\widetilde{\Phi}_e^s$	Liquid phase potential	$\phi_e^s(z,t) = \widetilde{\phi}_{e,s}(z,t) - \widetilde{\phi}_{s-e}^a(0,t) - U^a(\frac{c_{s,avg}^a(t)}{c_{s,max}^a})$
$\widetilde{\Phi}_e^c$	Liquid phase potential	$\phi_e^c(z,t) = \widetilde{\phi}_{e,c}(z,t) - \widetilde{\phi}_{s-e}^a(0,t) - U^a(\frac{c_{s,avg}^a(t)}{c_{s,max}^a})$
$\widetilde{\Phi}_s^a$	Solid phase potential	$\phi_s^a(z,t) = \widetilde{\phi}_s^a(z,t)$
$\widetilde{\Phi}_s^c$	Solid phase potential	$\phi_s^c(z,t) = \widetilde{\phi}_s^a(z,t) + v_{cell}(t)$

Tab. 1 Outputs of the transfer functions and corrections to recover their respective physically meaningful variables.

After linearization of the model equations around an equilibrium point and after transforming the variables with nonlinear corrections we arrive at the homogeneous partial differential equation in the frequency domain for the electrolyte concentration:

$$\begin{aligned} \frac{\partial^4}{\partial z^4} H_{\widetilde{C}_e^l}(z,s) - \tau_1^l(s)\frac{\partial^2}{\partial z^2} H_{\widetilde{C}_e^l}(z,s) + \tau_2^l(s) H_{\widetilde{C}_e^l}(z,s) = 0, \quad l \in \{a,c\} \\ \frac{\partial^2}{\partial z^2} H_{\widetilde{C}_e^s}(z,s) + \tau_1^s(s) H_{\widetilde{C}_e^s}(z,s) = 0. \end{aligned} \tag{3.2}$$

The generic solution of Equation (3.2) takes the form

$$H_{\widetilde{C}_e^l}(z,s) = \mathfrak{c}_1^l e^{\Lambda_1^l z} + \mathfrak{c}_2^l e^{-\Lambda_1^a z} + \mathfrak{c}_3^l e^{\Lambda_2^l z} + \mathfrak{c}_4^l e^{-\Lambda_2^l z}, \quad l \in \{a,c\} \tag{3.3}$$

$$H_{\widetilde{C}_e^s}(z,s) = \mathfrak{c}_1^s e^{\Lambda_1^s z} + \mathfrak{c}_2^s e^{-\Lambda_1^s z}, \tag{3.4}$$

where

$$\Lambda_1^l(s) = \sqrt{\frac{\tau_1^l - \sqrt{(\tau_1^l)^2 - 4\tau_2^l}}{2}}, \quad \Lambda_2^l(s) = \sqrt{\frac{\tau_1^l + \sqrt{(\tau_1^l)^2 - 4\tau_2^l}}{2}}, \quad \Lambda_1^s(s) = \sqrt{\tau_1^s}.$$

The coefficients $\mathfrak{c}_i^a,\ i \in \{1,\dots,4\}$, $\mathfrak{c}_i^s,\ i \in \{1,2\}$ and $\mathfrak{c}_i^c,\ i \in \{1,\dots,4\}$ are obtained from the boundary and transmission conditions.

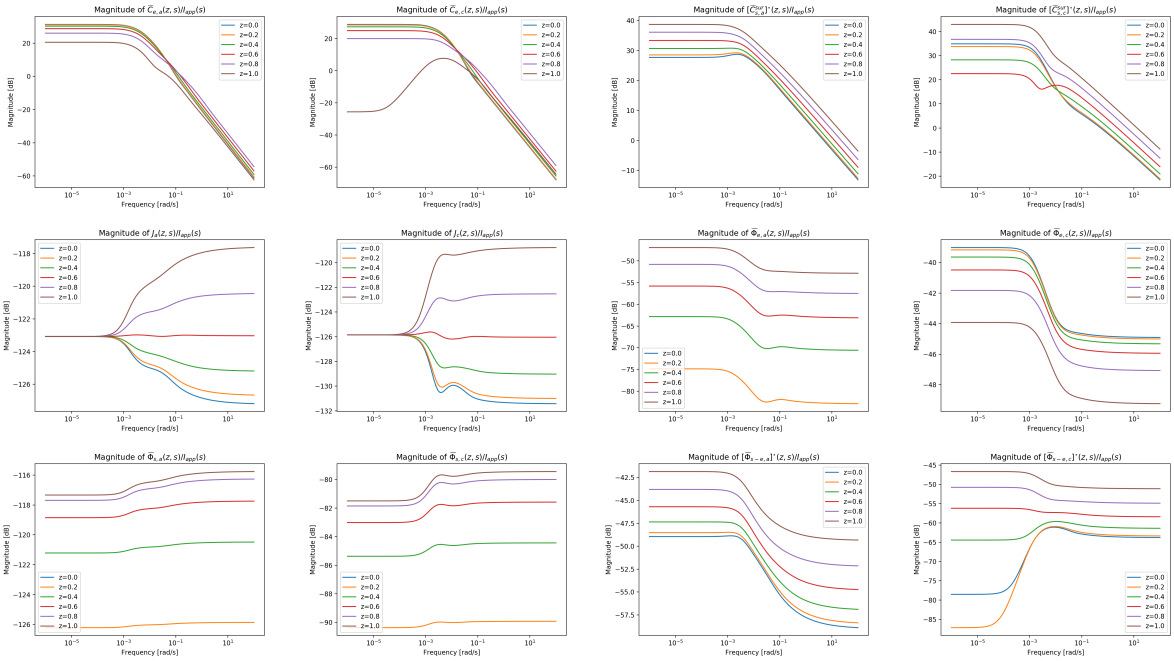


Fig. 2 Bodé plots for the transfer functions. The transfer functions described on this section have been implemented and sampled at different space locations from low (10^{-6}) to high (10^{2}) rad/s frequency

Remark 3.2 Some transfer functions have a pole at $s = 0$ so in the realization algorithm we will use the transfer function obtained after removing the integrator pole, that is,

$$H^{\star}(z,s) := H(z,s) - \frac{1}{s}\lim_{s\to 0} sH(z,s).$$

In this case, we will need to extend the state-space model to include the removed integrator pole which typically has a physical meaning, for example, the average concentration within the particles.

Now that we have deduced all the transfer functions, we can move to the first step of the discrete time realization algorithm.

3.2. Discrete time realization algorithm. Computation of Markov parameters

Assume that we have a system represented by a given transfer function $H(s)$ and we set the time step for the model to be T_s. The associated Markov parameters $g_k,\ k \in \{0, 1, \dots\}$ are defined by the response of the system to an unit pulse

$$u(t) = \begin{cases} 1, & \text{if } t \in [0, T_s], \\ \\ 0, & \text{otherwise,} \end{cases} \tag{3.5}$$

evaluated at times $kT_s,\ k \in \{0, 1, \dots\}$. By considering $h(t)$, the continuous-time impulse response related to $H(s)$ by the Laplace inversion formula $h(t) = \frac{1}{2\pi i}\oint_{\Gamma} H(s)e^{st}\,\mathrm{d}s$, with Γ a suitable contour. The Markov parameters can be computed with the convolution of $h(\cdot)$ and $u(\cdot)$. Indeed, we have

$$g_k = \int_0^{kT_s} h(\bar{t})u(kT_s - \bar{t})\mathrm{d}\bar{t}. \tag{3.6}$$

These integrals are approximated with a convolution quadrature method:

$$\int_0^{t_n} h(\bar{t})\, u(t_n - \bar{t})\mathrm{d}\bar{t} \approx \sum_{m=0}^{n} \omega_n^{T_s^{\text{fast}}}\, u(t_{n-m}), \tag{3.7}$$

where the weights are given by Tustin's formula cf. G. Plett's book [4] for more details.

3.3. Discrete time realization algorithm. Ho-Kalman algorithm

Now that we have the Markov parameters $g_k,\ k = 0, 1, \ldots$ we can compute the state-space model matrices using the Ho-Kalman algorithm. It can be summarized into three steps:

1. Build the Hankel matrix $\mathcal{H}_{i,j} = g_{i+j-1},\ i = 1, \ldots, l_1,\ j = 1, \ldots, l_2$ from the Markov parameters computed in previous section.

2. Compute the singular value decomposition (SVD) of the Hankel matrix $\mathcal{H}$, i.e: $\mathcal{H} = U\Sigma V^T$ and compute the shifted Hankel matrix $\mathcal{H}^{\uparrow}_{i,j} = g_{i+j}$.

3. Compute the observability matrix $O := U\Sigma^{1/2}$ and the controlability matrix $C := \Sigma^{1/2}V^T$. Then, the state space model matrices can be obtained as (using pseudo-inverses)

$$A := O^{-1}\mathcal{H}^{\uparrow}C^{-1}, \quad B := C_{:,0}, \quad C := O_{0,:}, \quad D := \lim_{s\to\infty} H_\varphi(z, s).$$

   If the transfer function $H_\varphi(z, s)$ has a pole at $s = 0$, the system needs to be augmented to take care of the integrator state so

$$A \equiv \begin{pmatrix} A & 0 \\ 0 & 1 \end{pmatrix}, \qquad B \equiv \begin{pmatrix} B \\ 1 \end{pmatrix},$$
$$C \equiv \begin{pmatrix} C \\ \mathrm{res}_{H^\star_\varphi} \end{pmatrix}^T, \quad D := \lim_{s\to\infty} H^\star_\varphi(z, s).$$

The derived state-space model matrices with this procedure are not unique and this is a problem when model blending is considered. Some transformations can be performed to arrive at a standard form, as explained in G. Plett's book [4].

3.4. State-space model validation

In Table 2, the set of parameters used for the discrete-time realization algorithm are shown.

Variable	z	Order	Time [s]	Variable	z	Order	Time [s]
j^a	0	4	1000	j^c	0	2	2000
	1	4	1000		1	2	2000
ϕ^a_e	0	4	0	ϕ^c_e	0	4	2000
	1	4	2000		1	4	2000
ϕ^a_{se}	0	4	1000				
c^a_e	0	3	4000	c^c_e	0	3	4000
	1	3	2000		1	3	2000
c^a_{se}	0	4	1000	c^c_{se}	0	4	2000
	1	4	1000		1	4	2000

	Lower bound	Upper bound	Setpoints
State-of-charge S	0.00	1.00	10
Anode porosity ϵ_a	0.15	0.35	5

	P2D	SSM	SSM+Blending
Total simulation time	18.74s	0.05	2.72
Time to be simulated / Total simulation time	16.00	6000.00	110.29

Tab. 2 Left: Maximum order and minimum time requested by impulse response to decay for each variable and location. Right top: Blending parameters. Right bottom: Performance comparison for the simulation of a 180s discharge followed by 120s relaxation with a time-step of one second

In Figure 3, a comparison between the solutions given by the reduced order model and the full order model from the previous section can be found. A performance comparison of the two models in terms of computational cost is shown in Table 2. We can see in two bottom figures that the weak coupling of degradation described in the previous section gives accurate results and that the reduced order model captures well the space-localized lithium plating. Reduced-order model accuracy can be improved by including model blending in the state-of-charge and the negative electrode porosity. Suppose we have a collection of state-space model matrices computed offline at different states-of-charge and negative electrode porosities. Then, in real time we can construct the state-space model matrices by interpolation between the closest offline computed matrices:

$$\begin{aligned} A_k = (1-\xi_S)[(1-\xi_\epsilon)A_{l,l} + \xi_\epsilon A_{l,r}] \\ + \xi_S\ [(1-\xi_\epsilon)A_{r,l} + \xi_\epsilon A_{r,r}] \end{aligned} \tag{3.8}$$

being $\xi_S = \frac{S_k - S_l}{S_r - S_l}$ and $\xi_T = \frac{\epsilon_k - \epsilon_l}{\epsilon_r - \epsilon_l}$ such that $S_k \in [S_l, S_r]$ and $\epsilon_k \in [\epsilon_l, \epsilon_r]$. The proccedure is analogous for the rest of state-space model matrices. A compromise between computational cost and accuracy is to perform blending only with a certain frequency instead of at every time step.

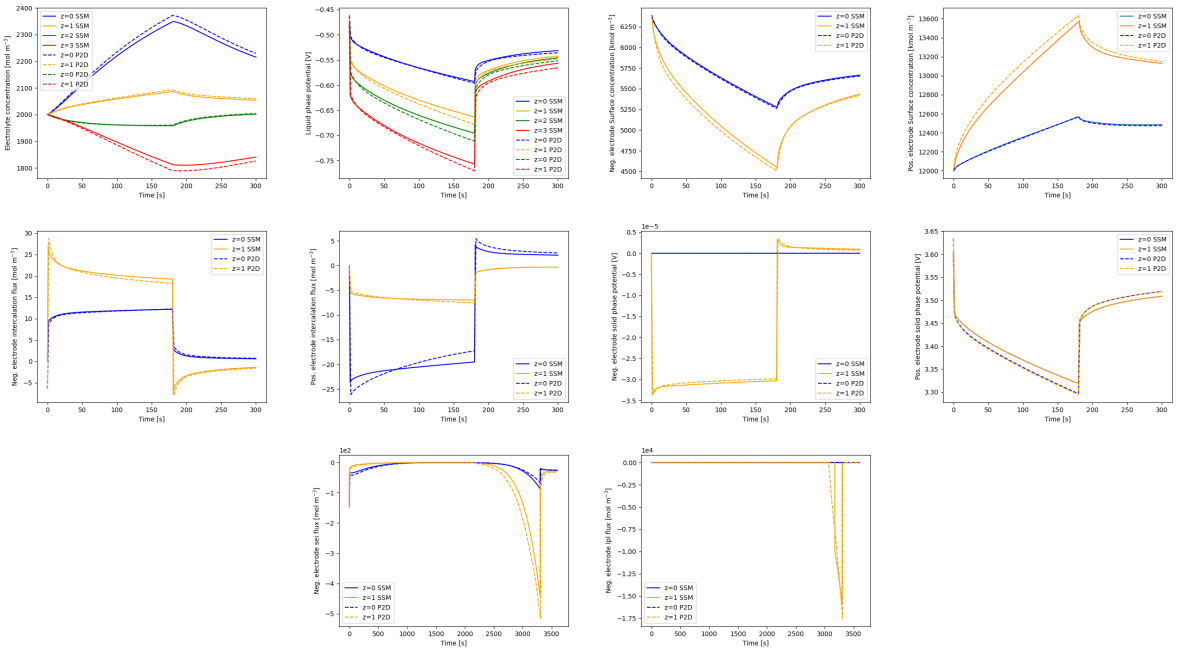


Fig. 3 Comparison between full order model and reduced order model simulations for a 180s discharge followed by 120s relaxation with a time-step of one second

4. Real-time control application. Reference governor

In order to use the electrochemically-derived state-space model in real applications, where the battery user is going to demand (resp. apply) some current, it is useful to include a reference governor which, using the state-space model, will compute the maximum current that the model can supply (resp. receive) while guaranteeing safety in form of user-defined constraints on state-space model states and outputs. On one hand we have constraints defined by the battery manufacturer on voltage, temperature or maximum current. This kind of constraints can be handled without a physic-based model, for example with an empirical equivalent circuit model. However, using these simple models, we could not enforce relevant constraints on lithium plating overpotential, electrode concentrations or even electrolyte depletion. More precisely, with our methodology constraints such as:

$$\theta^a_{min} \leq \frac{c^a_s(x,r,t)}{c^a_{s,max}} \leq \theta^a_{max}, \quad \theta^c_{min} \leq \frac{c^c_s(x,r,t)}{c^c_{s,max}} \leq \theta^c_{max}, \quad c_{e,min} \leq c_e(x,t), \quad \eta_{lpl,min} \leq \eta_{lpl}(x,t), \tag{4.1}$$

can be hadled. In order to show the performance of this methodology, we have implemented a reference governor similar to the one in S. Moura et. al. [3]. It requires to solve two constrained optimization problem at every time step in order to find the maximum allowable current that the battery can provide or receive, and this should be done in real-time.
On Figure 4 it can be seen how the reference governor modifies the charging protocol to keep the side reaction overpotential positive, avoiding lithium plating formation and premature ageing.

Conclusions

In this paper a physically-based reduced order battery model is briefly derived from the Doyle-Fuller-Newman porous electrode model. It includes two of the most relevant degradation mechanisms, that is, solid-electrolyte interface formation and lithium plating. It is shown to be compatible with real time control applications by means of a reference governor.

Acknowledgements

This work has been partially supported by Xunta de Galicia through grant 2021 GRC GI-1563 - ED431C 2021/15 and by the Spanish Ministry of Science and Innovation (MCIN), the Spanish National Research Agency (AEI) and the European Union (FEDER, EU) through grant PID2021-122625OB-I00.

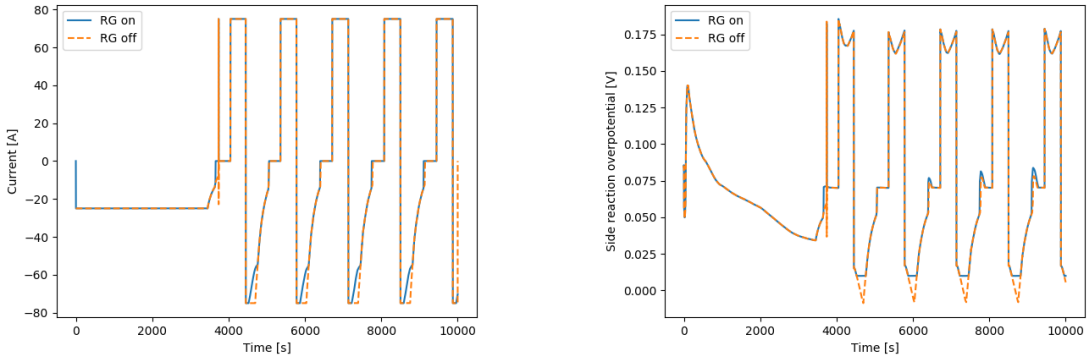


Fig. 4 Applied current and lithium plating overpotential with and without the modified reference governor.

References

[1] Jonas Keil and Andreas Jossen. Electrochemical modeling of linear and nonlinear aging of lithium-ion cells. *Journal of The Electrochemical Society*, 167(11):110535, jul 2020.

[2] J. Hake A. Johansson B. Kehlet A. Logg C. Richardson J. Ring M. E. Rognes M. S. Alnaes, J. Blechta and G. N. Wells. The FEniCS project version 1.5. *Archive of Numerical Software*, 3, 2015.

[3] Hector Perez, Niloofar Shahmohammadhamedani, and Scott Moura. Enhanced performance of li-ion batteries via modified reference governors and electrochemical models. *IEEE/ASME Transactions on Mechatronics*, 20(4):1511–1520, 2015.

[4] Gregory Plett. *Battery Management Systems, Volume I: Battery Modeling*. Artech, 2015.

[5] Albert Rodríguez, Gregory L. Plett, and M. Scott Trimboli. Improved transfer functions modeling linearized lithium-ion battery-cell internal electrochemical variables. *Journal of Energy Storage*, 20:560–575, 2018.

[6] Albert Rodríguez, Gregory L. Plett, and M. Scott Trimboli. Comparing four model-order reduction techniques, applied to lithium-ion battery-cell internal electrochemical transfer functions. *eTransportation*, 1:100009, 2019.

[7] Xiao-Guang Yang, Yongjun Leng, Guangsheng Zhang, Shanhai Ge, and Chao-Yang Wang. Modeling of lithium plating induced aging of lithium-ion batteries: Transition from linear to nonlinear aging. *Journal of Power Sources*, 360:28–40, 2017.

Large saturation effects provoke multiplicity in spatially heterogeneous predator-prey models

Julián López-Gómez and Eduardo Muñoz-Hernández
Universidad Complutense de Madrid, Spain
Instituto de Matemática Interdisciplinar (IMI)

Abstract

This communication analyzes the diffusive spatially heterogeneous predator-prey model introduced by the authors in [16], which takes into account the saturation effects of the predator in the abundance of preys through a saturation coefficient $\gamma m(x) \gneq 0$ with $\|m\|_\infty = 1$. The main result establishes the existence of, at least, two coexistence states for sufficiently large $\gamma > 0$ in a region of the parameters where the Lotka–Volterra counterpart cannot admit any coexistence state, regardless the size and shapes of the logistic and interactions coefficients of the model.

1. Introduction

This communication analyzes the existence and multiplicity of coexistence states for the generalized spatially heterogeneous predator-prey model

$$\begin{cases} \mathfrak{L}_1 u = \lambda u - a(x)u^2 - b(x)\dfrac{uv}{1+\gamma m(x)u} & \text{in } \Omega, \\ \mathfrak{L}_2 v = \mu v + c(x)\dfrac{uv}{1+\gamma m(x)u} - d(x)v^2 & \text{in } \Omega, \\ \mathfrak{B}_1 u = \mathfrak{B}_2 v = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

where Ω is a bounded domain of $\mathbb{R}^N$ whose boundary, $\partial\Omega$, is a $N-1$ dimensional manifold of class C^2, and $\mathfrak{L}_1$ and $\mathfrak{L}_2$ are second order uniformly elliptic operators in Ω of the form

$$\mathfrak{L}_\kappa := -\mathrm{div}\,(A_\kappa(x)\nabla) + \sum_{j=1}^N b_{j,\kappa}(x)\partial_j + c_\kappa(x), \qquad \kappa = 1, 2,$$

where, for every $k = 1, 2$, $A_\kappa(x) := \big(a_{ij,\kappa}(x)\big)_{1\le i,j\le N}$ is a symmetric matrix of order N such that

$$a_{ij,\kappa} = a_{ji,\kappa} \in W^{1,\infty}(\Omega) \quad \text{and} \quad b_{j,\kappa}, c_\kappa \in L^\infty(\Omega) \quad \text{for all} \quad 1 \le i, j \le N.$$

In this model, $\mathfrak{B}_1$ and $\mathfrak{B}_2$ are general boundary operators of mixed type such that, for every $\kappa = 1, 2$ and $\xi \in C(\bar\Omega) \cap C^1(\Omega \cup \Gamma_{1,\kappa})$,

$$\mathfrak{B}_\kappa \xi := \begin{cases} \xi & \text{on } \Gamma_{0,\kappa}, \\ \partial_{\nu_\kappa}\xi + \beta_\kappa(x)\xi & \text{on } \Gamma_{1,\kappa}, \end{cases} \tag{1.2}$$

where $\Gamma_{0,\kappa}$ and $\Gamma_{1,\kappa}$ are two closed and open disjoint subsets of $\partial\Omega$ such that

$$\Gamma_{0,\kappa} \cup \Gamma_{1,\kappa} = \partial\Omega.$$

In (1.2), $\beta_\kappa \in C(\Gamma_{1,\kappa}; \mathbb{R})$, and $\nu_\kappa \in C^1(\Gamma_{1,\kappa}; \mathbb{R}^N)$ is an outward pointing nowhere tangent vector field. Moreover, the functions coefficients $a(x)$, $b(x)$, $c(x)$, $d(x)$ and $m(x)$ are continuous in $\bar\Omega$ and satisfy $b \gneq 0$, $c \gneq 0$, $m \ge 0$, and

$$a(x) > 0, \quad d(x) > 0 \qquad \text{for all } x \in \bar\Omega,$$

while $\gamma > 0$ and $\lambda, \mu \in \mathbb{R}$ are regarded as bifurcation parameters.

From an ecological point of view, (1.1) models the interaction between a prey with density u and a predator with density v in the inhabiting territory Ω, where both species are assumed to have a logistic growth, or decay, in the absence of each other. In the special case when $m = 0$, (1.1) provides us with a rather generalized diffusive counterpart of the classical Lotka–Volterra predator-prey model, while if $m(x)$ is a positive constant, it is a generalized heterogeneous counterpart of the diffusive Holling–Tanner model introduced by Casal et al. [4]. The kinetics in [4] took into account the saturation effects of the predator in the presence of a high population of preys;

the constant $m > 0$ measuring the predator saturation level. In (1.1), the function $\gamma m(x)$ measures the level of saturation of the predator at any particular location $x \in \Omega$ where $m(x) > 0$, while saturation effects do not play any role if $m(x) = 0$. Throughout this note, we assume that

$$\|m\|_\infty \equiv \max_{\bar\Omega} m = 1.$$

Thus, γ can be viewed as the maximal amplitude of the saturation effects of the predator. Under these general assumptions, (1.1) combines, within the same territory Ω, the classical interactions of Lotka–Volterra type in the region $m^{-1}(0)$ with the Holling-Tanner functional responses in $\{x \in \Omega : m(x) > 0\}$. In its greatest generality, (1.1) includes most of the existing models of this type in the literature. In applications, $\lambda - c_1(x)$ and $\mu - c_2(x)$ stand for the neat growth, or decay, rates of the prey and the predator in the absence of each other.

The main goal of this note is analyzing the dynamics of (1.1) when γ grows to infinity. Thus, it is natural to perform the change of variables

$$w := \gamma\, u, \qquad \varepsilon = \frac{1}{\gamma}.$$

According to it, the model (1.1) can be expressed as

$$\begin{cases} \mathfrak{L}_1 w = \lambda w - \varepsilon a(x) w^2 - b(x)\dfrac{wv}{1+m(x)w} & \text{in } \Omega, \\ \mathfrak{L}_2 v = \mu v - d(x)v^2 + \varepsilon c(x)\dfrac{wv}{1+m(x)w} & \text{in } \Omega, \\ \mathfrak{B}_1 w = \mathfrak{B}_2 v = 0 & \text{on } \partial\Omega. \end{cases} \tag{1.3}$$

Hence, the problem of analyzing the dynamics of (1.1) for sufficiently large $\gamma > 0$ is equivalent to analyze the dynamics of (1.3) for sufficiently small $\varepsilon > 0$. Throughout this paper we will focus attention into (1.3) as a sort of shadow system perturbing from

$$\begin{cases} \mathfrak{L}_1 w = \lambda w - b(x)\dfrac{wv}{1+m(x)w} & \text{in } \Omega, \\ \mathfrak{L}_2 v = \mu v - d(x)v^2 & \text{in } \Omega, \\ \mathfrak{B}_1 w = \mathfrak{B}_2 v = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.4}$$

which is an uncoupled problem.

The plan of this note is the following. Section 2 collects some preliminaries. Section 3 gives some necessary and sufficient conditions for the existence of coexistence states of (1.1), as well as a local bifurcation result valid for all $\varepsilon \geq 0$. Section 4 ascertains the fine structure of the component of coexistence states of (1.4) bifurcating from the semitrivial positive solution of the form $(w, v) = (0, v)$ with $v > 0$. Finally, based on these results, in Section 5 we deliver our main multiplicity result for (1.3), with sufficiently small $\varepsilon > 0$. Essentially, as ε moves away from 0, a *metasolution* of (1.4) perturbs into a second coexistence state of (1.3) (see [13], if necessary, for the concept of metasolution).

2. Preliminaries

As a direct consequence of the elliptic L^p-theory, it is apparent that any non-negative weak solution of (1.3), (w, v), satisfies

$$u \in \mathscr{W}_1 := \bigcap_{p=N}^{\infty} W^{2,p}_{\mathfrak{B}_1}(\Omega), \qquad v \in \mathscr{W}_2 := \bigcap_{p=N}^{\infty} W^{2,p}_{\mathfrak{B}_2}(\Omega),$$

where, for every $\kappa = 1, 2$ and $p > N$, $W^{2,p}_{\mathfrak{B}_\kappa}(\Omega)$ stands for the Sobolev space of the functions $w \in W^{2,p}(\Omega)$ such that $\mathfrak{B}_\kappa w = 0$ on $\partial\Omega$. According to the Sobolev imbeddings, there is enough regularity on $\partial\Omega$ as to consider $\mathfrak{B}_\kappa$ in the classical sense, and (u, v) must be a strong solution of (1.1) (see, e.g., [12, Th. 5.11]).

For any given $V \in L^\infty(\Omega)$ and $\kappa = 1, 2$, we will denote by

$$\sigma_0[\mathfrak{L}_\kappa + V, \mathfrak{B}_\kappa, \Omega]$$

the principal eigenvalue of the linear eigenvalue problem

$$\begin{cases} (\mathfrak{L}_\kappa + V)\varphi = \tau\varphi & \text{in } \Omega, \\ \mathfrak{B}_\kappa \varphi = 0 & \text{on } \partial\Omega, \end{cases}$$

whose existence and uniqueness in our general setting was established in [12, Ch. 7]. The associated principal eigenfunction, unique up to a multiplicative positive constant, can be taken strongly positive in Ω, $\varphi \gg_\kappa 0$, in the sense that

$$\varphi(x) > 0 \ \text{ for all } \ x \in \Omega \cup \Gamma_{1,\kappa} \ \text{ and } \ \frac{\partial \varphi}{\partial n}(x) < 0 \ \text{ for all } \ x \in \Gamma_{0,\kappa},$$

where n stands for the outward unit normal vector field to Ω. The following result, going back to Cano-Casanova and López-Gómez [3] in its present generality, establishes the monotonicity of the principal eigenvalue with respect to the potential.

Theorem 2.1 *Let $V_1, V_2 \in L^\infty(\Omega)$ be such that $V_1 \lneq V_2$. Then, for every $\kappa = 1, 2$,*

$$\sigma_0 \left[\mathfrak{L}_\kappa + V_1, \mathfrak{B}_\kappa, \Omega\right] < \sigma_0 \left[\mathfrak{L}_\kappa + V_2, \mathfrak{B}_\kappa, \Omega\right].$$

Thus, the map $V \mapsto \sigma_0 \left[\mathfrak{L}_\kappa + V, \mathfrak{B}_\kappa, \Omega\right]$ is continuous in $L^\infty(\Omega)$ and increasing.

The next characterization is pivotal for analyzing elliptic equations or systems, as it is a key ingredient to infer most of our results. It goes back to López-Gómez and Molina-Meyer [14] for cooperative systems under Dirichlet boundary conditions, and to Amann and López-Gómez [2] and [11] for general boundary conditions of mixed type (see also [12, Th. 7.10] for further details).

Theorem 2.2 *For every $V \in L^\infty(\Omega)$ and $\kappa = 1, 2$, the next conditions are equivalent*:

(a) $\sigma_0 \left[\mathfrak{L}_\kappa + V, \mathfrak{B}_\kappa, \Omega\right] > 0$.

(b) *The tern $(\mathfrak{L}_\kappa + V, \mathfrak{B}_\kappa, \Omega)$ admits a positive strict supersolution, $h \in \mathscr{W}_\kappa$, i.e., for some $h \in \mathscr{W}_k$ such that $h \gneq 0$, the next estimates hold*

$$\begin{cases} (\mathfrak{L}_\kappa + V)h \geq 0 & \text{in } \ \Omega, \\ \mathfrak{B}_\kappa h \geq 0 & \text{on } \ \partial\Omega, \end{cases}$$

with some of these inequalities strict.

(c) *The tern $(\mathfrak{L}_\kappa + V, \mathfrak{B}_\kappa, \Omega)$ satisfies the strong maximum principle, i.e., $w \gg_\kappa 0$ for every function $w \in \mathscr{W}_\kappa$ such that*

$$\begin{cases} (\mathfrak{L}_\kappa + V)w \geq 0 & \text{in } \ \Omega, \\ \mathfrak{B}_\kappa w \geq 0 & \text{on } \ \partial\Omega, \end{cases}$$

with some of these inequalities strict.

The next result is invoked when analyzing the logistic equation in our abstract setting here. For a detailed proof of Theorem 2.3 in the classical case when $\beta_\kappa \geq 0$ the reader is sent to Fraile et al. [8, Th. 3.5]. The general case when β_k changes of sign can be reduced to the classical case through the exponential change of variable of Fernández-Rincón and López-Gómez [7, Sect. 3]. Alternatively, see Theorem 1.1 of Daners and López-Gómez [6], though this change of variable goes back to [12, Ch. 2] in a linear context.

Theorem 2.3 *Suppose $\rho \in \mathbb{R}$ and $\xi \in C(\bar{\Omega}; (0, \infty))$. Then, for every $\kappa = 1, 2$ and $V \in L^\infty(\Omega)$, the semilinear boundary value problem*

$$\begin{cases} (\mathfrak{L}_\kappa + V)w = \rho w - \xi(x)w^2 & \text{in } \ \Omega, \\ \mathfrak{B}_\kappa w = 0 & \text{on } \ \partial\Omega, \end{cases} \tag{2.1}$$

admits a positive solution if, and only if,

$$\rho > \sigma_0 \left[\mathfrak{L}_\kappa + V, \mathfrak{B}_\kappa, \Omega\right],$$

which is unique if it exists. Moreover, if we denote it by

$$w_{\rho,\kappa} \equiv \theta_{[\mathfrak{L}_\kappa + V, \rho, \xi]} \in \mathscr{W}_\kappa,$$

then $w_{\rho,k} \gg_\kappa 0$, the map $\rho \to w_{\rho,\kappa}$ is point-wise increasing if

$$\rho > \sigma_0 \left[\mathfrak{L}_\kappa + V, \mathfrak{B}_\kappa, \Omega\right],$$

and $w_{\rho,\kappa}$ bifurcates from $w = 0$ at $\rho = \sigma_0 \left[\mathfrak{L}_\kappa + V, \mathfrak{B}_\kappa, \Omega\right]$. Furthermore, if $\bar{u}$ is a positive strict supersolution of (2.1)*, then $\bar{u} \gg_\kappa w_{\rho,\kappa}$. Similarly, if $\underline{u}$ is a positive strict subsolution of* (2.1)*, then $\underline{u} \ll_\kappa w_{\rho,\kappa}$.*

More precisely, through this note, we denote by $\theta_{[\mathfrak{L}_\kappa+V,\rho,\xi]}$ the maximal non-negative solution of (2.1). Hence, by Theorem 2.3,

$$\theta_{[\mathfrak{L}_\kappa+V,\rho,\xi]} := \begin{cases} 0 & \text{if } \rho \le \sigma_0\left[\mathfrak{L}_\kappa + V, \mathfrak{B}_\kappa, \Omega\right], \\ \gg_\kappa 0 & \text{if } \rho > \sigma_0\left[\mathfrak{L}_\kappa + V, \mathfrak{B}_\kappa, \Omega\right]. \end{cases}$$

Moreover, as a byproduct of Theorem 2.3, (1.1) has a semitrivial positive solution of the form $(u, 0)$ if, and only if,

$$\lambda > \sigma_{0,1} \equiv \sigma_0[\mathfrak{L}_1, \mathfrak{B}_1, \Omega]$$

and, in such case, $(u, 0) = \left(\theta_{[\mathfrak{L}_1,\lambda,a]}, 0\right)$. Similarly, (1.1) has a semitrivial positive solution of the form $(0, v)$ if, and only if,

$$\mu > \sigma_{0,2} \equiv \sigma_0[\mathfrak{L}_2, \mathfrak{B}_2, \Omega]$$

and, in such case, $(0, v) = \left(0, \theta_{[\mathfrak{L}_2,\mu,d]}\right)$.

3. Coexistence regions and bifurcation of coexistence states from $(0, \theta_{[\mathfrak{L}_2,\mu,d]})$

In this section we are going to estimate the regions of the (λ, μ)–plane where the problem (1.1), or, equivalently, (1.3) has some coexistence state. Then, regardless the values of $\varepsilon > 0$ and $\mu > \sigma_{0,2}$, it is established the existence of a component of coexistence states bifurcating from the semitrivial curve $(0, \theta_{[\mathfrak{L}_2,\mu,d]})$ at a certain (unique) value of λ.

Next result collects some (optimal) necessary and sufficient conditions for the existence of coexistence states and, hence, it determines the coexistence regions of (1.3). It is a direct consequence of [16, Th. 4.1 & 5.1].

Theorem 3.1 *Suppose that, for some $\varepsilon > 0$,* (1.3) *has a coexistence state, (w, v). Then,*

$$\lambda > \varphi_\varepsilon(\mu) \equiv \sigma_0\left[\mathfrak{L}_1 + b\frac{\theta_{[\mathfrak{L}_2,\mu,d]}}{1+m\theta_{[\mathfrak{L}_1,\lambda,\varepsilon a]}}, \mathfrak{B}_1, \Omega\right] \quad \textit{and} \quad \mu > \Psi_\varepsilon(\lambda) \equiv \sigma_0\left[\mathfrak{L}_2 - \varepsilon c\frac{\theta_{[\mathfrak{L}_1,\lambda,\varepsilon a]}}{1+m\theta_{[\mathfrak{L}_1,\lambda,\varepsilon a]}}, \mathfrak{B}_2, \Omega\right]. \tag{3.1}$$

Conversely, under the following condition

$$\lambda > \Phi(\mu) \equiv \sigma_0\left[\mathfrak{L}_1 + b\theta_{[\mathfrak{L}_2,\mu,d]}, \mathfrak{B}_1, \Omega\right] \quad \textit{and} \quad \mu > \Psi_\varepsilon(\lambda), \tag{3.2}$$

the problem (1.3) *has, at least, a coexistence state.*

Figure 1 sketches the construction of the wedges (3.1) and (3.2) given by Theorem 3.1. By Theorem 2.1,

$$\varphi_\varepsilon(\mu) \equiv \sigma_0\left[\mathfrak{L}_1 + b\frac{\theta_{[\mathfrak{L}_2,\mu,d]}}{1+m\theta_{[\mathfrak{L}_1,\lambda,\varepsilon a]}}, \mathfrak{B}_1, \Omega\right] < \sigma_0\left[\mathfrak{L}_1 + b\theta_{[\mathfrak{L}_2,\mu,d]}, \mathfrak{B}_1, \Omega\right] \equiv \Phi(\mu), \qquad \text{for all } \mu > \sigma_{0,2}.$$

According to Theorem 3.1, (1.3) has a coexistence state in the solid area of Figure 1, whereas outside the union of the solid and dashed wedges of Figure 1, it cannot admit any coexistence state. Thus, the dashed wedge must contain the edge of the coexistence region. By the analysis already done in [16, Sec. 3], the global structure of the curve $\mu = \Psi_\varepsilon(\lambda)$ can change according to the nature of $m(x)$, as illustrated in Figure 1 and explained in its caption.

Since $\theta_{[\mathfrak{L}_1,\lambda,\varepsilon a]} = \varepsilon^{-1}\theta_{[\mathfrak{L}_1,\lambda,a]}$, it is apparent that

$$\lim_{\varepsilon\downarrow 0}\varphi_\varepsilon(\mu) = \lim_{\varepsilon\downarrow 0}\sigma_0\left[\mathfrak{L}_1 + b\frac{\theta_{[\mathfrak{L}_2,\mu,d]}}{1+\frac{m}{\varepsilon}\theta_{[\mathfrak{L}_1,\lambda,a]}}, \mathfrak{B}_1, \Omega\right] = \sigma_0\left[\mathfrak{L}_1 + \left(1 - \chi_{\text{int supp}\, m}\right) b(x)\theta_{[\mathfrak{L}_2,\mu,d]}, \mathfrak{B}_1, \Omega\right],$$

where, for any subset $A \subset \mathbb{R}^N$, χ_A stands for the characteristic function of the set A, i.e., $\chi_A(x) = 1$ if $x \in A$, and $\chi_A(x) = 0$ if $x \in \mathbb{R}^N \setminus A$. In the next section, it will become apparent that the function

$$\varphi_0(\mu) := \sigma_0\left[\mathfrak{L}_1 + \left(1 - \chi_{\text{int supp}\, m}\right) b(x)\theta_{[\mathfrak{L}_2,\mu,d]}, \mathfrak{B}_1, \Omega\right], \qquad \mu > \sigma_{0,2}, \tag{3.3}$$

provides us with the left limiting curve to the region where the uncoupled model (1.4) possesses a coexistence state. The curve $\lambda = \varphi_0(\mu)$ has been also plotted in Figure 1 and, again by Theorem 2.1, $\varphi_0(\mu) < \varphi_\varepsilon(\mu)$ if $bm \gneq 0$.

According to Theorem 2.2, for every real number $e > \max\{-\sigma_{0,1}, -\sigma_{0,2}\}$ and $\kappa = 1, 2$, $(\mathfrak{L}_\kappa + e, \mathfrak{B}_\kappa, \Omega)$ is an invertible operator with strongly positive inverse. Thus, the solutions of the problem (1.3) are the zeroes of the operator

$$\mathfrak{F} : \mathbb{R}\times\mathbb{R}\times\mathbb{R}\times C^1_{\mathfrak{B}_1}(\bar\Omega)\times C^1_{\mathfrak{B}_2}(\bar\Omega) \to \mathscr{W}_1\times\mathscr{W}_2,$$

defined, for every $\lambda, \mu, \varepsilon \in \mathbb{R}$, $w \in C^1_{\mathfrak{B}_1}(\bar\Omega)$ and $v \in C^1_{\mathfrak{B}_2}(\bar\Omega)$, by

$$\mathfrak{F}(\lambda, \mu, \varepsilon, w, v) := \begin{pmatrix} w - (\mathfrak{L}_1 + e)^{-1}\left[(\lambda + e)w - \varepsilon a w^2 - b\frac{wv}{1+mw}\right] \\ v - (\mathfrak{L}_2 + e)^{-1}\left[(\mu + e)v - dv^2 + \varepsilon c\frac{wv}{1+mw}\right] \end{pmatrix}.$$

The next result shows the bifurcation to coexistence states from the semitrivial positive solution $(0, \theta_{[\mathfrak{L}_2,\mu,d]})$ along the curve $\lambda = \Phi(\mu)$. It is a direct consequence of the theorem of bifurcation from simple eigenvalues of Crandall and Rabinowitz [5]. It provides us with the local structure of the set of bifurcating coexistence states.

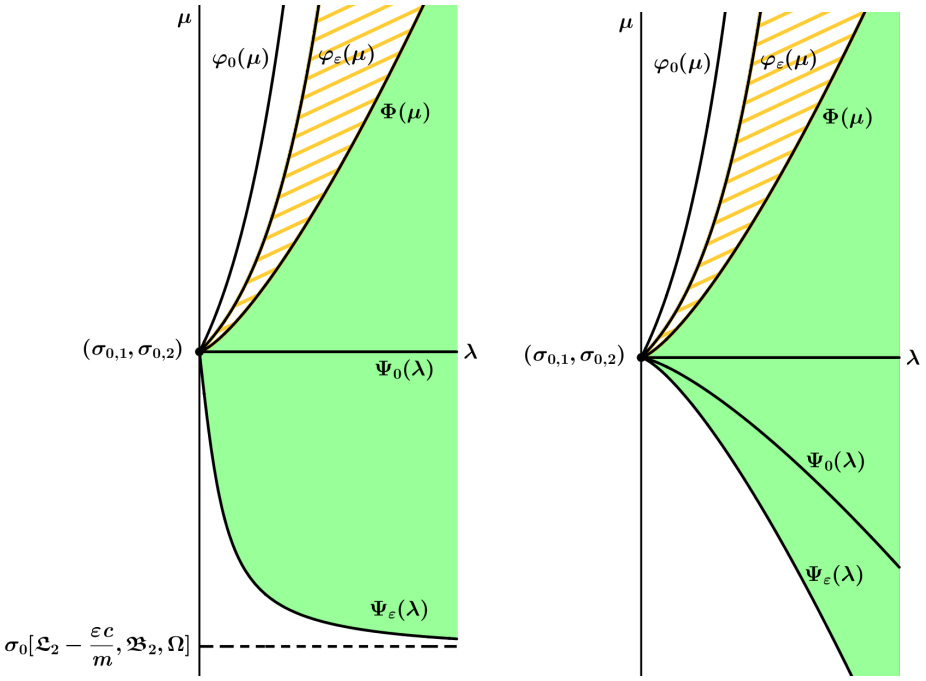


Fig. 1 The coexistence regions of (1.3) according to Theorem 3.1 if $m(x) > 0$ for all $x \in \bar{\Omega}$ (left picture) or $\operatorname{int} m^{-1}(0) \neq \emptyset$ (right picture). When $m(x) > 0$ for all $x \in \bar{\Omega}$ the curve $\mu = \Psi_\varepsilon(\lambda)$, $\lambda > \sigma_{0,1}$, inherits the same asymptotic behavior as in the classical Holling–Tanner case when $m(x) \equiv m > 0$, whereas in case $\operatorname{int} m^{-1}(0) \neq \emptyset$ it possesses the same asymptotic features as in the classical Lotka–Volterra model with $m \equiv 0$.

Theorem 3.2 *For every $\mu > \sigma_{0,2}$ and $\varepsilon \in \mathbb{R}$, there exist $\delta = \delta(\mu, \varepsilon) > 0$ and an analytic map $(\lambda, w, v) : (-\delta, \delta) \to \mathbb{R} \times \mathscr{W}_1 \times \mathscr{W}_2$ such that*:

(i) $(\lambda(0), w(0), v(0)) = \big(\Phi(\mu), 0, \theta_{[\mathfrak{L}_2,\mu,d]}\big)$.

(ii) $\mathfrak{F}(\lambda(s), \mu, \varepsilon, w(s), v(s)) = 0$ *for all* $s \in (-\delta, \delta)$.

(iii) $v(s) \gg_2 0$ *if* $s \in (-\delta, \delta)$, $w(s) \gg_1 0$ *if* $s \in (0, \delta)$ *and* $w(s) \ll_1 0$ *if* $s \in (-\delta, 0)$.

(iv) *The set of solutions of* (1.3) *in a neighborhood of* $(\lambda, w, v) = \big(\Phi(\mu), 0, \theta_{[\mathfrak{L}_2,\mu,d]}\big)$ *consists of the curves* $\big(\lambda, 0, \theta_{[\mathfrak{L}_2,\mu,d]}\big)$, $\lambda \sim \Phi(\mu)$, *and* $(\lambda(s), w(s), v(s))$, $s \in (-\delta, \delta)$.

Moreover, there are two functions $w_1, w_1^* \gg_1 0$ such that

$$\lambda'(0) = \int_\Omega \big(\varepsilon a - b\theta_{[\mathfrak{L}_2,\mu,d]}\big)\, w_1^2 w_1^* + \int_\Omega b\, \big(\mathfrak{L}_2 + 2d\theta_{[\mathfrak{L}_2,\mu,d]} - \mu\big)^{-1} \big(\varepsilon c \theta_{[\mathfrak{L}_2,\mu,d]} w_1\big)\, w_1 w_1^*. \tag{3.4}$$

Remark 3.3 As the dependence of $\mathfrak{F}$ on $\varepsilon \in \mathbb{R}$ is also analytic, by the implicit function theorem used in the proof of the theorem of Crandall and Rabinowitz [5], it becomes apparent that the bifurcated curve

$$(\lambda(s), w(s), v(s)) \equiv (\lambda(s, \varepsilon), w(s, \varepsilon), v(s, \varepsilon))$$

also is analytic with respect to the parameter ε.

4. The coexistence states of the uncoupled problem (1.4)

This section determines the set of coexistence states of the limiting shadow problem (1.4). As v satisfies

$$\begin{cases} \mathfrak{L}_2 v = \mu v - d(x) v^2 & \text{in } \Omega, \\ \mathfrak{B}_2 v = 0 & \text{on } \partial\Omega, \end{cases}$$

the condition $\mu > \sigma_{0,2} \equiv \sigma_0[\mathfrak{L}_2, \mathfrak{B}_2, \Omega]$ is imperative so that (1.4) can have a coexistence state. Otherwise, $v = 0$ for any component-wise nonnegative solution, (w, v), of (1.4). Thus, throughout this section, we assume that

$\mu > \sigma_{0,2}$. In such case, by Theorem 2.3, for every coexistence state (w, v) of (1.4), necessarily $v = \theta_{[\mathfrak{L}_2,\mu,d]} \gg_2 0$, and $w \gg_1 0$ is a positive solution of the associated problem

$$\begin{cases} \mathfrak{L}_1 w = \lambda w - b(x)\theta_{[\mathfrak{L}_2,\mu,d]} \dfrac{w}{1+m(x)w} & \text{in } \Omega, \\ \mathfrak{B}_1 w = 0 & \text{on } \partial\Omega. \end{cases} \tag{4.1}$$

The next result ascertains the range of λ's where (4.1) has a coexistence state.

Lemma 4.1 *Suppose $bm \gneq 0$ and $w \neq 0$ is a positive solution of* (4.1)*. Then, $w \gg_1 0$ and*

$$\sigma_{0,1} \leq \varphi_0(\mu) < \lambda = \sigma_0\Big[\mathfrak{L}_1 + \tfrac{b(x)\theta_{[\mathfrak{L}_2,\mu,d]}}{1+m(x)w}, \mathfrak{B}_1, \Omega\Big] < \Phi(\mu), \tag{4.2}$$

where $\varphi_0(\mu)$ and $\Phi(\mu)$ are the functions defined in (3.3) *and* (3.2)*, respectively.*

According to Theorem 3.2, there is a bifurcation to positive solutions of (4.1) from $(w, v) = (0, \theta_{[\mathfrak{L}_2,\mu,d]})$ at $\lambda = \Phi(\mu)$, which is subcritical, because

$$\lambda'(0) = -\int_\Omega b\theta_{[\mathfrak{L}_2,\mu,d]} w_1^2 w_1^* < 0. \tag{4.3}$$

Set $\mathfrak{F}_0(\lambda, \mu, w, v) \equiv \mathfrak{F}(\lambda, \mu, 0, w, v)$, and let denote by $\mathcal{S}_0$ the set of nontrivial solutions of (4.1) defined by

$$\mathcal{S}_0 := \{(\lambda, \mu, w, \theta_{[\mathfrak{L}_2,\mu,d]}) \in \mathfrak{F}_0^{-1}(0) \,:\, w \neq 0\} \cup \{(\lambda, \mu, 0, \theta_{[\mathfrak{L}_2,\mu,d]}) \,:\, \lambda \in \Sigma(\mathscr{L}(\lambda))\},$$

where $\Sigma(\mathscr{L}(\lambda))$ stands for the generalized spectrum of the Fredholm curve

$$\mathscr{L}(\lambda) := D_{(w,v)}\mathfrak{F}_0(\lambda, \mu, 0, \theta_{[\mathfrak{L}_2,\mu,d]}).$$

The next result establishes that the component $\mathscr{C}_0^+$ of positive solutions of $\mathcal{S}_0$ with $(\Phi(\mu), \mu, 0, \theta_{[\mathfrak{L}_2,\mu,d]}) \in \bar{\mathscr{C}}_0^+$ satisfies

$$\mathcal{P}_\lambda(\mathscr{C}_0^+) = (\varphi_0(\mu), \Phi(\mu)), \tag{4.4}$$

where $\mathcal{P}_\lambda$ stands for the λ-projection operator, $\mathcal{P}_\lambda(\lambda, \mu, w, \theta_{[\mathfrak{L}_2,\mu,d]}) \equiv \lambda$. Moreover, it shows that $\mathscr{C}_0^+$ is unbounded at $\lambda = \varphi_0(\mu)$ and it provides us with its fine structure nearby $\lambda = \Phi(\mu)$ and $\lambda = \varphi_0(\mu)$. This is a crucial information to obtain the main multiplicity result of this note for (1.3) with sufficiently small $\varepsilon > 0$.

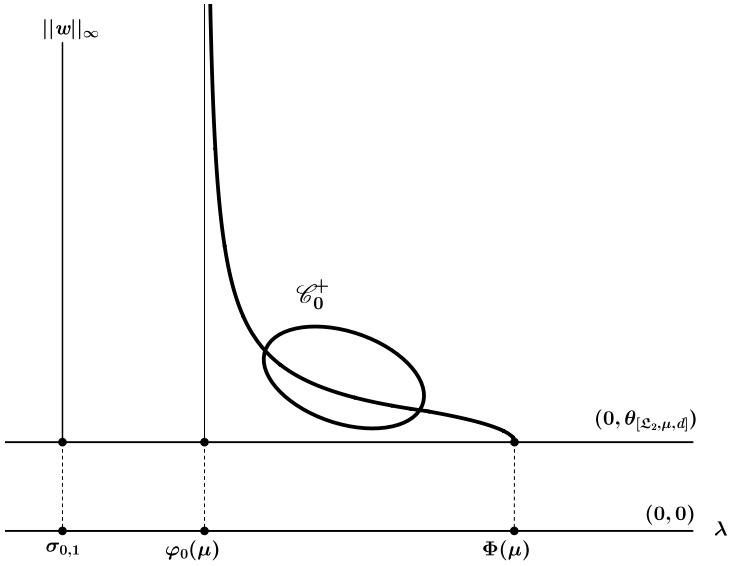


Fig. 2 An admissible component $\mathscr{C}_0^+$ in case $bm \gneq 0$.

Theorem 4.2 *The component $\mathscr{C}_0^+$ satisfies* (4.4)*. Moreover, for every sequence of positive solutions, in $\mathscr{C}_0^+$, $\{(\lambda_n, \mu, w_n, \theta_{[\mathfrak{L}_2,\mu,d]})\}_{n\geq 1}$, such that $\lim_{n\to\infty} \lambda_n = \varphi_0(\mu)$, necessarily*

$$\lim_{n\to\infty} \|w_n\|_\infty = +\infty. \tag{4.5}$$

On the other hand, in a neighborhood of $(\lambda, \mu, w, \theta_{[\mathfrak{L}_2,\mu,d]}) = (\Phi(\mu), \mu, 0, \theta_{[\mathfrak{L}_2,\mu,d]})$ *in* $\mathbb{R}\times\mathbb{R}\times\mathscr{W}_1\times\{\theta_{[\mathfrak{L}_2,\mu,d]}\}$, $\mathscr{C}_0^+$ *consists of the analytic curve* $(\lambda(s), \mu, w(s), \theta_{[\mathfrak{L}_2,\mu,d]})$ *given by Theorem* 3.2. *Actually, there exists* $r > 0$ *such that, for every* $\lambda \in [\Phi(\mu) - r, \Phi(\mu))$, (4.1) *has a unique positive solution. Moreover, for sufficiently small* $r > 0$, *this positive solution is linearly unstable with one-dimensional unstable manifold.*

Furthermore, there exists $r > 0$ *such that, for every* $\lambda \in (\varphi_0(\mu), \varphi_0(\mu) + r]$, (4.1) *has a unique positive solution,* $(\lambda, \mu, w_\lambda, \theta_{[\mathfrak{L}_2,\mu,d]})$, *which is non-degenerate. Thus, for these values of* λ, $\mathscr{C}_0^+$ *consists of an analytic curve of positive solutions bifurcating from* $+\infty$ *at* $\lambda = \varphi_0(\mu)$.

Figure 2 shows an admissible component $\mathscr{C}_0^+$ of positive solutions of (4.1) adjusted to the patterns of Theorem 4.2. Although (4.1) has a unique positive solution for λ sufficiently close to either $\Phi(\mu)$, or $\varphi_0(\mu)$, the problem might possess an arbitrarily large number of positive solutions for some intermediate range of values of the parameter λ, as illustrated in Figure 2.

5. An optimal multiplicity result for the original model

The next multiplicity result is the main theorem of this communication. Recall that, owing to Theorem 3.1, for every $\mu > \sigma_{0,2}$, (1.3) has a coexistence state if $\lambda > \Phi(\mu)$. Moreover, in such case, $\lambda > \varphi_\varepsilon(\mu)$.

Theorem 5.1 *Fix* $\lambda^* \in (\varphi_0(\mu), \Phi(\mu))$. *Then, there exists* $\varepsilon_0 \equiv \varepsilon_0(\lambda^*) > 0$ *such that, for every* $\varepsilon \in (0, \varepsilon_0)$, (1.3) *possesses a component* $\mathscr{C}_\varepsilon^+$ *of coexistence states satisfying the following properties*:

(a) $\mathcal{P}_\lambda\left(\mathscr{C}_\varepsilon^+\right) = [\lambda_T, +\infty)$ *for some* $\lambda_T \equiv \lambda_T(\varepsilon) \in (\varphi_\varepsilon(\mu), \lambda^*]$.

(b) *For every* $\lambda \in [\lambda^*, \Phi(\mu))$, (1.3) *has, at least, two different coexistence states.*

(c) $\mathscr{C}_\varepsilon^+$ *is an analytic curve, with respect to the parameter* λ, *in a neighborhood of*

$$(\lambda, \mu, w, v) = (\Phi(\mu), \mu, 0, \theta_{[\mathfrak{L}_2,\mu,d]}).$$

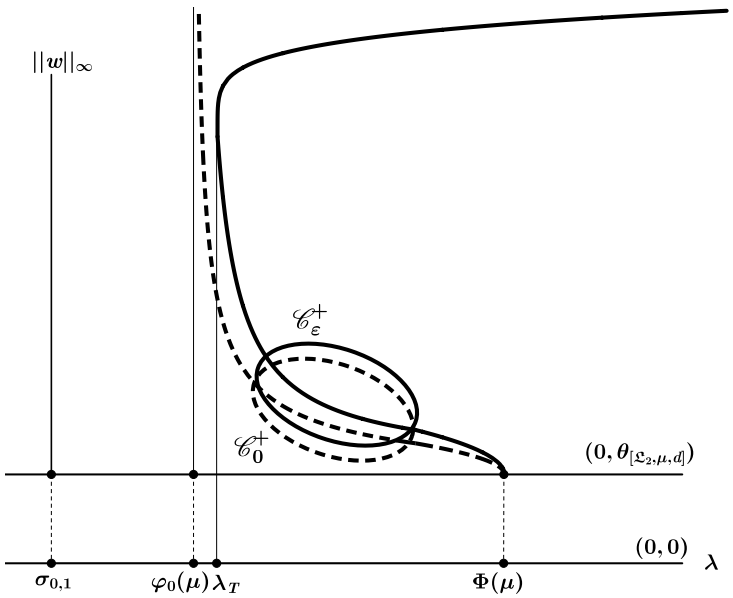


Fig. 3 The components $\mathscr{C}_0^+$ (dashed line) and $\mathscr{C}_\varepsilon^+$ (solid line) for small $\varepsilon > 0$

Naturally, $\mathscr{C}_\varepsilon^+$ is the perturbation of the component $\mathscr{C}_0^+$ constructed in Section 4 as $\varepsilon > 0$ leaves $\varepsilon = 0$. Figure 3 shows an admissible component $\mathscr{C}_\varepsilon^+$ (solid line) perturbing from $\mathscr{C}_0^+$ (dashed line) and satisfying Theorem 5.1. Roughly spoken, the proof of Theorem 5.1 relies on the following features:

- The existence of a priori bounds for the coexistence states of the problem (1.4). These bounds can be derived as an application of Theorems 2.2 and 2.3. The existence of a priori bounds together with [10, Th. 7.2.2] guarantee that the component $\mathscr{C}_\varepsilon^+$ is unbounded in λ, i.e., $\mathcal{P}_\lambda\left(\mathscr{C}_\varepsilon^+\right)$ should contain an interval of the form $[\hat{\lambda}, +\infty)$ for some $\hat{\lambda} > \varphi_0(\mu)$.

- The use of the implicit function to make sure that $\mathscr{C}_\varepsilon^+$ consists of two arcs of analytic λ-curve for $\lambda \sim \varphi_0(\mu)$ and $\lambda \sim \Phi(\mu)$ and sufficiently small $\varepsilon > 0$, and the construction of an open isolating neighborhood, O, for a certain subcomponent of $\mathscr{C}_0^+$ joining these two arcs.
- Showing that, for sufficiently small $\varepsilon > 0$, the isolating neighborhood also packages the components $\mathscr{C}_\varepsilon^+$.
- Using the fixed point index in cones, as axiomatized by Amann [1], to infer the multiplicity result as in [9]. According to it, the existence of a second coexistence state for all $\lambda \geq \lambda^*$ holds.

Remark 5.2 According to Theorem 3.2 and [10, Th. 7.2.2], the existence of a $\lambda^* \in (\varphi_\varepsilon(\mu), \Phi(\mu))$ such that, for every $\lambda \in [\lambda^*, \Phi(\mu))$, (1.3) has, at least, two coexistence states is guaranteed if $\lambda'(0) < 0$. This occurs for sufficiently small ε, which might be larger than the ε_0 given by Theorem 5.1; at least, for $\varepsilon \in (0, \varepsilon^*)$, where ε^* satisfies $\lambda'(0, \varepsilon^*) = 0$.

Acknowledgements

This communication has been written under the auspices of the Ministry of Science and Innovation of Spain, under Research Grant PID2021-123343NB-I00, and of the IMI of Complutense University. The second author, ORCID 0000-0003-1184-6231, has been also supported by contract CT42/18-CT43/18 of Complutense University of Madrid.

References

[1] Herbert Amann. Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces. *SIAM Rev.*, 18:620–709, 1976.

[2] Herbert Amann and Julián López-Gómez. A priori bounds and multiple solutions for superlinear indefinite elliptic problems. *J. Diff. Equations*, 146(2):336–374, 1998.

[3] Santiago Cano-Casanova and Julián López-Gómez. Properties of the principal eigenvalues of a general class of non-classical mixed boundary value problems. *J. Diff. Equations*, 178:123–211, 2002.

[4] Alfonso Casal, Chris Eilbeck and Julián López-Gómez. Existence and uniqueness of coexistence states for a predator-prey model with diffusion. *Diff. Int. Eqns.*, 7:411–439, 1994.

[5] Michael G. Crandall and Paul H. Rabinowitz. Bifurcation from simple eigenvalues. *J. Funct. Anal.*, 8:321–340, 1971.

[6] Daniel Daners and Julián López-Gómez. Global dynamics of generalized logistic equations. *Adv. Nonl. Studies*, 18:217–233, 2018.

[7] Sergio Fernández-Rincón and Julián López-Gómez. The singular perturbation problem for a class of generalized logistic equations under non-classical mixed boundary conditions. *Adv. Nonl. Studies*, 19:1–27, 2019.

[8] José María Fraile, Pablo Koch Medina, Julián López-Gómez, and Sandro Merino. Elliptic eigenvalue problems and unbounded continua of positive solutions of a semilinear elliptic equation. *J. Diff. Equations*, 127:295–319, 1995.

[9] Julián López-Gómez. Positive periodic solutions of Lotka-Volterra reaction-diffusion systems. *Diff. Int. Eqns.*, 5(1):55–72, 1992.

[10] Julián López-Gómez. *Spectral Theory and Nonlinear Functional Analysis*. Research Notes in Mathematics 426, Chapman & Hall/CRC Press, Boca Raton, FL, 2001.

[11] Julián López-Gómez. Classifying smooth supersolutions for a general class of elliptic boundary value problems. *Adv. Diff. Eqns.*, 8:1025–1042, 2003.

[12] Julián López-Gómez. *Linear Second Order Elliptic Operators*. World Scientific, Singapore, 2013.

[13] Julián López-Gómez. *Metasolutions of Parabolic Equations in Population Dynamics*. CRC Press, Boca Raton, 2015.

[14] Julián López-Gómez and Marcela Molina-Meyer. The maximum principle for cooperative weakly elliptic systems and some applications. *Diff. Int. Eqns*, 7:383–398, 1994.

[15] Julián López-Gómez and Eduardo Muñoz-Hernández. A spatially heterogeneous predator-prey model. *DCDS-Series B*, 26(4):2085–2113, 2021.

[16] Julián López-Gómez and Eduardo Muñoz-Hernández. Multiplicity in a spatially heterogeneous predator-prey model. *In preparation*, 2023.

Evaluation of a general car-following model for micro/macro traffic modelling

José Enríquez Gabeiras[1], **Juan Francisco Padial Molina**[2]

1. jose.enriquez@alumnos.upm.es *Department of Matemática Aplicada. Universidad Politécnica de Madrid, Spain*
2. jf.padial@upm.es *Department of Matemática Aplicada. Universidad Politécnica de Madrid, Spain*

Abstract

In transport traffic analysis, road traffic is modelled as a system of interacting particles. In the last decades a number of mathematical models have been proposed to represent the dynamics of this process. Most important among them are the car-following models, which have been used to model both the inter-vehicles dynamics (micro modelling) and the aggregated traffic process (macro modelling).

Even though these models accurately represent phenomena observed in real roads, like congestion creation, metastability and wave propagation, they are sometimes based on narrow assumptions that do not match the behaviour of real drivers and vehicles. The accuracy of these models is specially relevant now that new communication technologies like 5G and V2X (Vehicle to everything) are intended to be a key part of the future road-traffic infrastructure, and will require stringent latency requirements, complicating the dynamics of the traffic process.

This paper presents results on a new, general purpose traffic model to be used in the study of the impact of new V2X technologies on the road traffic phenomena. The results include application to micro (particle dynamics) and macro (fluid-like dynamics) simulation scenarios and numerical evaluation and validation of the mathematical model using an extension to network simulator *ns-3*.

1. Introduction

Transport traffic analysis has gained significant attention in the last decades, joining the techniques developed in modern physics and the capabilities offered by modern computing platforms. Road traffic is modelled as a system of interacting particles (vehicles), and present phenomena similar to those observed in physical systems, like phase transitions, metastability and wave propagation [7, 10].

The models developed have been focused on modelling the behaviour of the driver and its interaction with the environment, specially the dynamics of other vehicles. These models accurately represent some of the situations observed in real roads, but they are usually based on narrow assumptions that do not match the behaviour of real drivers and vehicles. Moreover, the modelling is now further complicated because a significant number of autonomously driven vehicles is expected to appear in the roads in the near future. On the other hand, technologies introduced in the last generation of mobile network standards (5G), called vehicle-to-everything communications (V2X), are now capable to offer latencies low enough to enable delay critical driving assistance over mobile connections. This is considered a fundamental cornerstone for safely deploying self-autonomous driving systems.

Paramount for the successful adoption of these services is the understanding of the challenges and risks derived of the mobility scenarios. This will be especially critical for applications highly dependent on strict timing, which will be affected by both network traffic and road traffic conditions.

This paper presents results on a new car-following model recently introduced [8], based on a delay-difference equation of a sigmoidal class of functions, that provides a realistic alternative to the existing car-following models, overcoming some of their limitations. This model is useful to assess two-vehicle dynamics in which the follower vehicle is autonomously driven, and may be affected by variations on the reaction time to adjust its velocity (for instance, because of congestion on some communication process involved). The accuracy, advantages and range of applications of the model is demonstrated using simulation techniques. Then, the applicability to a scenario involving V2X communications (see [11]) is shown by considering the influence of the variability of the delay inherent to the communications process in the car-following model dynamics.

1.1. Car following models

Most road traffic studies are focused on the determination of the conditions that generate congestion and traffic jams, for which two main techniques are used: the *microscopic model*, where traffic is seen as individual interacting particles, and the *macroscopic model*, where traffic is seen as a compressible fluid.

A typical micro model is the *follow-the-leader* or *car-following* model, where the vehicles move in a single lane and no lane changes are considered, and the evolution of the vehicle n is affected by the vehicle $n-1$ ahead. There exist different approaches to describe the dynamics involved, the most common being those based on modelling the influence of the distance between their position ($X_{n-1} - X_n$) and the difference between their velocities ($X'_{n-1} - X'_n$).

In particular, the General Motors' model (GM) states that the acceleration of the follower is a function of its speed, the speed difference and the distance [4]. Also, it is normally considered that there is a reaction delay τ_n in the application of the acceleration to the vehicle, due to human and mechanic reaction time in such a way that the dependence of the acceleration with the position and velocity can in general be expressed by some delay differential equation (DDE) that includes the reaction delay τ_n. Historically there have been successive proposals for this DDE trying different dependencies on the variables to reflect real world driver behaviour, which can be summarized in the generalized GM model as [3]:

$$X_n''(t+\tau_n) = c\frac{X_n'(t)^m}{(X_{n-1}(t)-X_n(t))^l}(X_{n-1}'(t)-X_n'(t)) \tag{1.1}$$

This formula can model different degrees of influence of the variables by means of the exponents m and l. Nevertheless, despite accurately represent real world situations and explain some important traffic dynamics, these models allow the vehicles to be arbitrarily close when both have the same velocity, which is unrealistic when such velocity is high.

An alternative approach called the Optimal Velocity Model (OVM) (see [2, 12]) has also been introduced to model a behaviour in which the follower adjust its velocity according to an optimal velocity, which is a function V of the distance between the vehicles, i.e.:

$$X_n'(t+\tau_n) = V(X_{n-1}(t)-X_n(t)) = V(\Delta X_n(t)) \tag{1.2}$$

where V is nonlinear, has to be monotonically increasing and has a maximum value (typically, a sigmoidal function).

This approach does modulate the velocity depending on the distance, and presents the advantage of being easily tractable analytically. In fact, one can approximate the acceleration function and remove the implicit dependency on the delay by series expansion, and, assuming that the delay τ is the same for all the vehicles, we get:

$$X_n''(t) \approx \frac{1}{\tau}(V(\Delta X_n(t)) - X_n'(t)) \tag{1.3}$$

This provides an explicit dependence of the acceleration on the distance, the velocity and the delay. This makes this model very convenient for establishing a relationship between the micro and the macro models ([7, 9, 10]). Nevertheless, the assumption behind equation (1.3) is that the vehicle always can change the velocity to the optimal one in time τ, regardless of the present value of the velocity $X_n'(t)$. This allow for extremely high values for the acceleration function in certain circumstances, which is rather unrealistic.

2. The nDDE model

As a way to overcome the limitations of existing models described above, a new car-following model has been proposed in [8]. This model complies with the main requirements used in former models (dependence on the distance and the relative velocity, and inclusion of the delay due to reaction time), while, at the same time, provides more realistic dynamics.

The model similarly focuses on the case of two vehicles (denoted now as 0 for the leading car and 1 for the follower car). It also considers that the leading car has a constant velocity $X_0'(t) = v_0$, and defines new variables for the distance between the cars ($s(t)$) and the relative velocity ($s'(t)$):

$$s(t) = (X_0(t) - X_1(t)) \qquad s'(t) = (v_0 - X_1'(t)) \qquad s''(t) = (-X_1''(t)) \tag{2.1}$$

So, the retarded dependence of the acceleration function (like in equations (1.1) and (1.3)) can be expressed in a general way as:

$$-s''(t+\tau) = X_1''(t+\tau) = g(s(t), s'(t)) \tag{2.2}$$

where $g()$ is some function, to which some additional requirements are added:

i.) We must avoid $s(t) = 0$, so, for any v_0, it is defined a minimum distance between cars m. In equilibrium, $s = m$, and $s'' = 0$.

ii.) There is a maximum acceleration $a > 0$ and a maximim deceleration $b < 0$, so $b < g(s, s') < a$

iii.) The function g is increasing with respect to s.

The model is based on the following sigmoidal function, that meets the previous requirements:

$$g(s, s') = a - \frac{(a+b)}{1 + \frac{b}{a}e^{d(s-m+ks')}}, \forall (s, s') \in \mathbb{R}^2 \tag{2.3}$$

where d is a parameter to model the intensity of the response of the car-driver ensemble, and k is a parameter to model the driver's response according to the safe distance and the perceived relative velocity [8].

This is the ***nDDE*** model (new car-following traffic model), which offers a more realistic behaviour than the previous models, since the acceleration is now bounded to a range of real vehicle values, and the equilibrium distance is an increasing function of the vehicle velocity. Reference [8] characterizes the nDDE equation as a Retarded Functional Differential Equation (RFDE) [6], and provides an analysis of the dynamical characteristics and the change from stability to oscillatory solutions in equilibrium depending on the value of the delay term (Hopf bifurcation).

We next introduces an improvement to the nDDE base model that accounts for real vehicle mechanics and has been used in the simulations presented in next sections of the paper.

2.1. Modelling real vehicle behaviour

In the nDDE model, once the deviation on s or s' is significant, the acceleration and deceleration values easily reach the maximum values. For the deceleration action one can expect sudden variations in its value, if needed (by the action of variably pressing the brakes). But for the acceleration process it is physically impossible to accelerate over the limits imposed by the engine-gear mechanism. Thus, for the model to be more realistic, it should be expected that the acceleration process is somewhat more progressive. Indeed, the maximum acceleration achievable is dependent not only on a given engine power output, but also on the vehicle speed and the gear engaged. The acceleration is typically a decreasing function approaching 0 when the vehicle approaches the maximum velocity (see, for instance, [13]). For the deceleration it is possible to assume that the value b is independent on the velocity, and determined by the brakes, tires and road ensemble [13].

Therefore it is expected that the real acceleration process is smoother than the one in the original $nDDE$ model, with the maximum acceleration a dependent on the velocity, instead of a constant value. For the purpose of this analysis, we consider the following function to model the acceleration process, inspired by the example provided in [13]:

$$a(s') = \frac{2}{3} e^{0.25 - \frac{((v_0 - s') + 10)^2}{400}} ((v_0 - s') + 10) \tag{2.4}$$

This function tries to emulate the progressive decrease of the maximum acceleration when the vehicle gains speed, until it is barely capable of accelerating at higher speeds (in this case, around 40 m/s).

In this way, the $nDDE$ model is transformed on account of this maximum acceleration:

$$-s''(t+\tau) = g(s, s') = a(s') - \frac{(a(s') + b)}{1 + \frac{b}{a(s')} e^{d(s - m + ks')}}, \forall (s, s') \in \mathbb{R}^2 \tag{2.5}$$

For the sake of clarity, in the rest of the paper we will refer to this< modified $nDDE$ model as $nDDE_a$.

3. Numerical evaluation

This section contains a numerical evaluation of the modified $nDDE_a$ model introduced in the preceeding section. We base the analysis on an extension of the *ns-3* simulation platform widely used in communication network analysis.

3.1. *ns-3* simulation platform

ns-3 (Network Simulator version 3, `https://www.nsnam.org/`) is an open source platform for discrete-event (i.e., Monte-Carlo) simulation for communication systems. It is based on C++ and Python, with an extensive model library for Internet and mobile communications systems, and is a key tool for many projects in current network research activity.

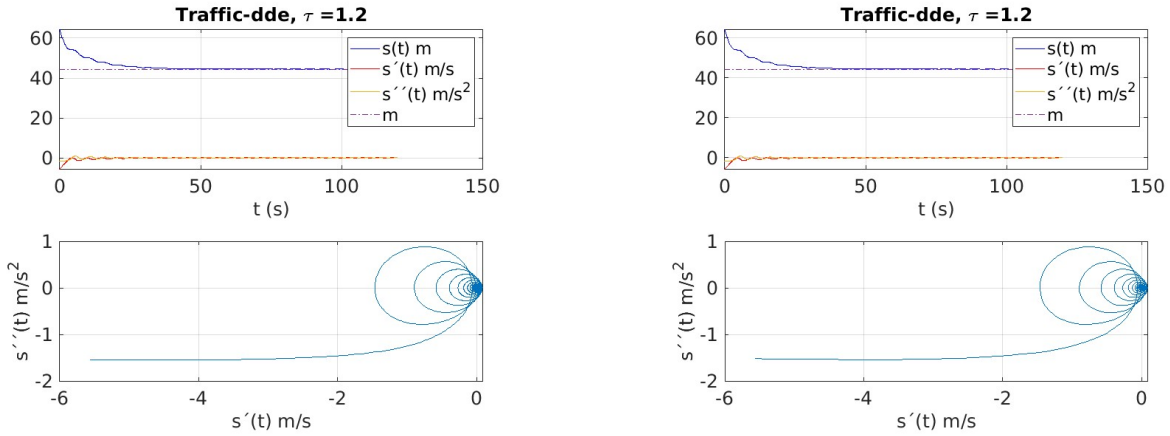
The interesting aspect of using *ns-3* for our analysis is three-fold: first, it provides a library of mobility models that has been extended to model the dynamics of the $nDDE$ models; secondly, it offers models to simulate all modern communication systems, thus allowing a seamless integration of the simulation of the mobility and communications dynamics; finally, it allows us to simulate micro (i.e., two vehicles) and macro (many vehicles) scenarios using the same code base.

It is worth noting that this work introduces complex mobility dynamics and delayed reactions in *ns-3* simulations. The implementation of the $nDDE$ models required to build the simulation objects for the road and vehicles system, the modification of existing *ns-3* objects to enable the delayed update of the acceleration, and the introduction of the nDDE model dynamics in the simulation logic. To perform the validation we have simulated the same scenarios as in [8], which use the following parameters: $a = 2.0576\ m/s^2$, $a(0) = 6.6666\ m/s^2$, $b = 1.5677\ m/s^2$, $v_0 = 22.2222\ m/s$, $m = 44.4444\ m$, $d = 0.1124$ and $k = 11.3890\ s$.

3.2. Micro behaviour

[8] includes a detailed characterization of the original nDDE model, in which numerical simulations of the solutions in the Hopf bifurcation for different delay parameters τ are provided. We follow the same structure to present the results, in which each graphic has two parts: above is the graphic of the s, s' and s'' with respect to time; and below the 2D-curve (s, s').

Figure 1a provides the original results in [8] for a scenario in which the follower vehicle is initially at a higher distance and a lower speed than the equilibrium values, with a delay $\tau = 1.2s$. This has been generated using the code dde23 of Matlab to solve the DDE. Figure 1b shows the results produced by the simulator. We can see that the simulation reproduces with a high degree of accuracy the results produced by the numerical integration of the equations.

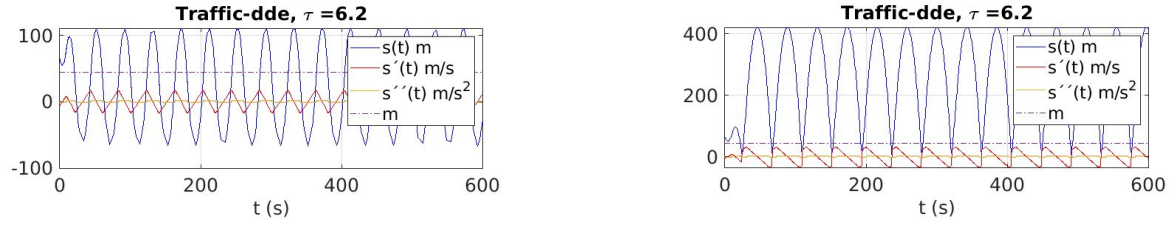


(a) Numerically solved with *dde23* routine of Matlab.

(b) Using the *ns-3* simulation

Fig. 1 Oscillatory and stable solution for $\tau = 1.2s$ in $nDDE$

It is also interesting to check for higher delay values, since for very high values the instability of the system lead to the crash of the vehicles (distance = 0). This can be seen in figure 2a, where after a short time, the solution contains negative values for the inter-vehicle distance. But this is explicitly avoided in the simulation, as can be seen in figure 2b. This allow us to use the simulation platform to study situations with more complex interactions beyond the analysis of the Hopf bifurcation, which is specially important to evaluate the behaviour of the $nDDE$ acceleration model in multi vehicle (macro) scenarios.



(a) Numerically solved with *dde23* routine of Matlab.

(b) Using the *ns-3* simulation

Fig. 2 Oscillatory and unstable solution for $\tau = 6.2s$ in $nDDE$

3.3. Macro behaviour

Macroscopic analysis is focused on the traffic process seen as a fluid motion. A most important objective is to determine the conditions that generate congestion and traffic jams, for which it is usually used the so-called fundamental diagram, which relates the vehicle density ρ with the density current (flow) q (see for instance [7, 10]). The traffic transition from *free traffic* at low densities to *congested traffic* at high densities is a phenomena usually compared to phase transition processes in physics.

Macro models represent the traffic flow in the same way as a compressible fluid. The traffic process in a point in space x and a time t is described in terms of the vehicle density $\rho(x,t)$ and the average velocity $v(x,t)$. The traffic current is then $q(x,t) = \rho(x,t)v(x,t)$

The model is described by the continuity equation for fluids:

$$\frac{\partial \rho}{\partial t} + \frac{\partial q}{\partial x} = 0 \tag{3.1}$$

For some microscopic models, a relationship has been established with its dual macroscopic model by finding the equation (*momentum* equation) that relates the density and the velocity of the flow. For instance, for the Optimal Velocity Model (OVM) ([2, 12]), the following equation has been found to explain the macroscopic behaviour [9]:

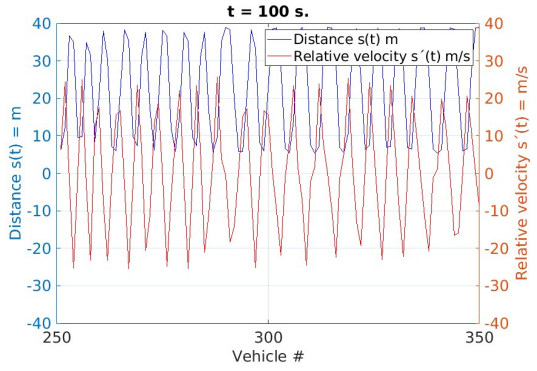
$$\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial x} = \frac{1}{\tau}[V(\rho^{-1}) - v] - \frac{V'}{2\tau\rho^3}\frac{\partial \rho}{\partial x} + \frac{1}{6\tau\rho^2}\frac{\partial^2 v}{\partial x^2} \tag{3.2}$$

where V is the optimal velocity (dependent on the vehicle distance (1.2)) and τ is the delay implicit in the OVM. These equations have been used to explain the formation of density waves in situations of traffic variability. The purpose of this section is to show that the proposed $nDDE$ models are suitable for modelling the macro process too, compared to this well known model

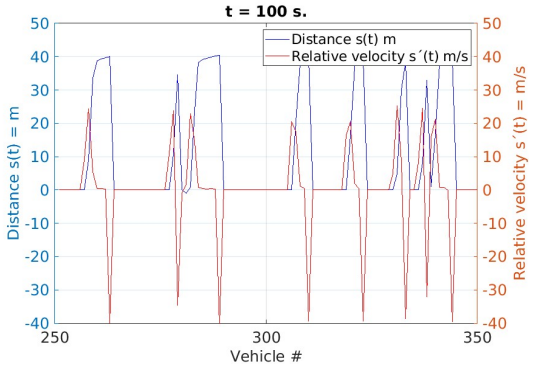
For instance, stability analysis for equation (3.2) indicates that in the OVM the flow will be stable for delay values under the limit $\tau_{lim} < \frac{1}{2V'(s_0)}$, where s_0 is the headway distance in the equilibrium and V the optimal velocity in (1.2). If V is defined as $V(s) = v_{max} \cdot (1 - \frac{1}{1+e^{d(s-m)}})$ with d and m with the same values as in Section 3.1, we get an stability limit for the delay $\tau_{lim} = 0.2s$.

Figure 3a shows the simulation of the evolution in $t = 100s$ of a traffic flow that initially is stable with an equilibrium distance $s_{init} = 50m$ and an initial velocity $v_{init} = 22.2222m/s$, when the delay is $\tau = 0.45s > \tau_{lim}$. The evolution for 100 vehicles in the simulation (index 251 to 350) is shown. The propagating density waves are manifested in groups of vehicles with similar velocities and distances (stop and go traffic), groupings that are backwards propagated in the road to the following vehicles.

The changes of velocity of the OVM model are rather sudden and symmetric in the acceleration and deceleration processes. Vehicles behaves with the same aggressiveness in the two phases, and the density waves show regular, quasi-periodic patterns, all of which is not very realistic.



(a) OVM simulation. $\tau = 0.5s$, $t = 100s$



(b) nDD_a model simulation. $\tau = 3s$, $t = 100s$

Fig. 3 Distance and velocity evolution

To check the adequacy of the nDDE model to study macroscopic traffic evolution, similar scenarios have been simulated with the $nDDE_a$ acceleration model. In figure 3b the same evolution for the $nDDE_a$ model is presented. In this instance instability sets in for higher values of the delay ($\tau = 3s$), and we get asymmetric and aperiodic density waves. We also observe a smoother transition from the stop situation to the higher speeds as a result of the application of the maximum acceleration dependence on the velocity. This also produces groups with a higher number of vehicles, which seems more in agreement with reality.

Thus these results indicate that the $nDDE_a$ model reproduce the behaviour expected at the macroscopic scale (wave formation and propagation), and, when a realistic function for the acceleration process is used, it provides a more realistic traffic representation. The analysis of the macroscopic characteristics and the link with the microscopic dynamics to determine the limiting delay that produces instability is intended as future work.

4. Applicability to V2X communications

V2X technologies [11] have seen an strong development as a result of the advances on the mobile communications technologies in the last two decades, up to the point to be considered a reliable support of communications between vehicles to coordinate driving functions. The application of these technologies are focused mostly in safety applications and autonomous driving (also referred to as ITS, Intelligent Transportation Systems), which inherently

require a very low packet loss probability and communications delay. The required level of performance has only started to be achieved with the fourth generation of mobile networks in the decade 2010-2020, though the range of applications and the amount of real applications have been somewhat limited as of today [11].

One of the most important standars in the first wave of standardization is Dedicated Short Range Communications (DSRC), also known as the 802.11p standard from IEEE, consisting on the setting up of an ah-hoc network among close vehicles to exchange status updates.

We use now a simplified model of the DSRC protocol as a first approach to test the performance of the car-following model under study, taking into account the additional source of complexity introduced by the variability of the delay of the communications process.

4.1. Ad-hoc network modelling

We consider a multi-vehicle system in which each vehicle broadcasts the information about its state (position-speed) to the rest of vehicles using DSRC messages, and each vehicle adapts its own dynamics by adjusting the acceleration according to the $nDDE_a$ model (eq. (2.5)). Each vehicle send its status messages accessing the shared channel, and the message is received by the vehicles in its area of coverage (defined by a radius R). The messages are sent by each vehicle periodically but asynchronously, in such a way that there exist a certain probability that two transmissions *collide* (i.e., two vehicles transmitting at the same time and thus mutually interfering their transmission), which translates to a corresponding delay by the need of retransmitting the message after a random waiting time [5]. The most important source of such collisions is the *"hidden node"* problem [14]. Due to the limited reach of the transmission of a given vehicle, it is possible that two vehicles are so far away that they do not detect each other's transmissions, but some vehicles in between are able to receive both of them, so, if they are simultaneous, neither message is received correctly by this intermediate vehicle.

For this situation the probability P of a successful transmission can be calculated using the classical unslotted ALOHA model [1, 14]:

$$P = e^{-2\cdot\rho\cdot\lambda(D\cdot TS+2T_t(2R-D))} \tag{4.1}$$

where λ is the message generation rate of each vehicle, T_t is the packet transmission time, TS is the length of a timeslot (basic digital transmission time), R is the radius of the coverage area of the vehicle, D is the distance for which a collision is detected ($D > R$) and ρ is the vehicle density.

A receiving vehicle will get the information correctly after a number of trials N (geometrically distributed with parameter P), so the average delay caused by the hidden node problem is:

$$\tau = \frac{\overline{N}}{\lambda} = \frac{1}{\lambda \cdot P} \tag{4.2}$$

4.2. Results for the coupled system

The effect of introducing a variable delay in the $nDDE_a$ model as an effect of communications congestion has been simulated. First we focus on the impact on the microscopic modelling, and later we provide some insights on the macroscopic modelling.

Microscopic behaviour of nDDE model with variable delay

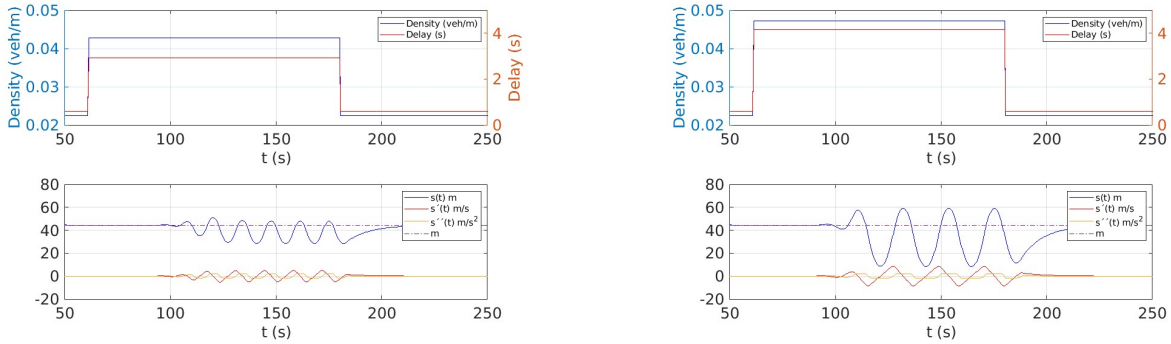
The consideration of the dependence of the delay on the traffic density introduces and additional degree of complexity if we want to consider that the changes in density could influence the velocity of all the vehicles, including the leader. Nevertheless, we can test the influence of variable delay on the microscopic behaviour by considering a simpler two-vehicle system that is part of an autonomously driven platoon (with the leader maintaining constant velocity) that travel along a road with light load, but with the road in opposite direction having variable density, which would induce changes in the experienced delays (in both directions of the road).

In such a case, gathering equations (2.5), (4.1) and (4.2), we now have the following system of equations governing the dynamics of the two-vehicle system:

$$\begin{cases} s''(t+\tau(\rho)) = -g(s,s') = -a(s') + \frac{(a(s')+b)}{1+\frac{b}{a(s')}e^{d(s-m+ks')}} \\ \tau(\rho) = \frac{1}{\lambda}e^{2\cdot\rho\cdot\lambda(D\cdot TS+2T_t(2R-D))} \end{cases} \tag{4.3}$$

where ρ is the density of the road in opposite direction.

As an example of the impact, we first provide the results for the case in which the platoon traverses a road sector with significantly more density in the opposite road. Figures 4a and 4b show the evolution of the dynamics of the follower vehicle when the density it encounters in the opposite road changes from 0.0225 veh./m (inter-vehicle



(a) Density change from 0.0225 veh./m to 0.04275 veh./m

(b) Density change from 0.0225 veh./m to 0.04725 veh./m

Fig. 4 Micro evolution for delay depending on vehicle density.

distance of 44.44 m) to values of 0.04275 veh./m (distance 21.85 m) and 0.04725 veh./m (distance 21.16 m) respectively.

It can be appreciated that small variations of vehicle density produce remarkably different instability oscillations (the highest density value differs only 10% between the two examples), and the instability takes some time to set in after the density changes, and it takes a much shorter time to recover once the density and the delay have returned to lower values.

Despite the complexity introduced by the variable delay, this scenario is relatively simple, since the delay variation is not influenced by the action of the two vehicles of the car-following model. Nevertheless, from the point of view of the global traffic process (beyond the car-following two-car system), it is also interesting to analyze the complete coupling of the vehicles and communication system, including the feedback process that the change of the velocity of the vehicles introduce on the density of vehicles, which in turn modify the delay and therefore again the velocity.

Given the complexity associated with the system defined by equation (4.3) to analyze a multi-vehicle system following the nDDE model, to study the global behaviour of the system requires the usage of the paradigm of the macro model. We will provide some ideas about this in next section.

Macroscopic behaviour of nDDE model with variable delay

We have tested an scenario that is initially stable, where the system is composed by 600 vehicles cruising at a velocity $v = 15,6m/s$ and an inter-vehicle distance of $s = 33.33m$, each one behaving according to the $nDDE_a$ model with constant delay $\tau = 0.5s$. As expected, for such low values for the delay, the traffic remains in equilibrium, maintaining the initial conditions (time $t = 100$ and $t = 150s$ are shown).

The impact of a variable delay induced by the changes in density is shown in figures 5a and 5b. In this case, the value of density at each point x is calculated as the average of the vehicle density $\overline{\rho}(x)$ at the surrounding area $[x - R, x + R]$ $(R = 300m)$. The delay experienced in x is then a function of the density $\tau(\overline{\rho}(x))$ following equation (4.3).

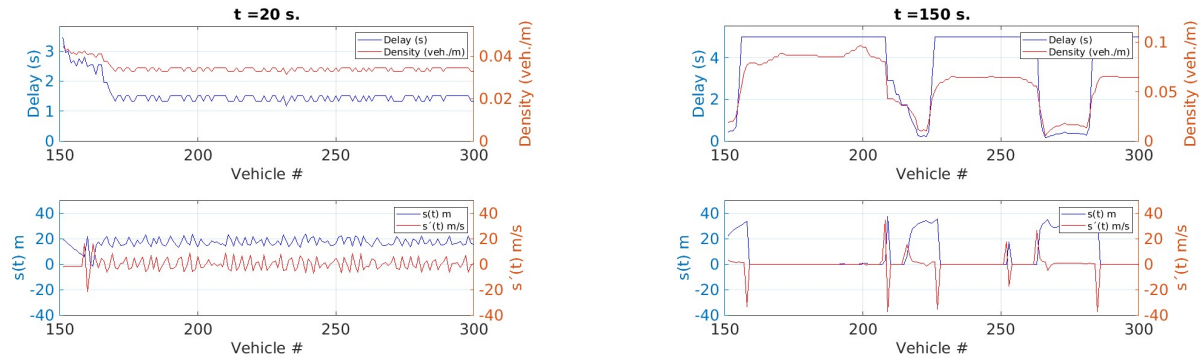
We show in figures 5a and 5b the evolution of the simulation for two points in time, namely 20 s and 150 s. For practical considerations, we have limited the maximum delay in the function $\tau(\overline{\rho}(x))$ to $\tau_{max} = 5s$.

- After 20 s ($t = 20s$) the system is still near the equilibrium state, with values of density and delay starting to change, which introduce variation in the distance and relative velocity
- In $t = 150s$ the density waves grow in size, and the delay fluctuates between the maximum and minimum values depending on the accumulation of vehicles.

These results show that due to the exponential dependence of the delay on the density (equations (4.1), small density variations may produce significant excursions in the delay, which in turn cause the formation of density waves that reinforce the density variations.

5. Conclusion and future work

This paper has presented an evaluation of a new car-following model (the $nDDE$ model introduced in [8]) which present more realistic behaviour with respect to former car-following models. An improvement on the original model (called $nDDE_a$) has been introduced to adapt the model to real vehicle mechanics. The capabilities of the model to represent realistic traffic situations, both at micro (2 vehicles) and macro (N vehicles) scales has been



(a) t = 20 s.

(b) t = 150 s.

Fig. 5 Macro evolution with density 0.03 veh./m and and variable delay dependent on the density.

shown using a new simulation tool developed on the *ns-3* platform. This tool is intended to study the dynamics of the road traffic management that relies on vehicle communications to update state information. In this regard, a first example of the interaction of the two dynamics (road traffic and communication traffic) has been presented.

Future developments of this research will progress along the following lines:

- Evolution of the $nDDE_a$ model to include more precise assumptions, specially on the acceleration and deceleration mechanical models.
- Application of the presented simulation methodology to the study of stability and performance evaluation of other V2X standards, specially the new generation systems (802.11bd and NR-V2X)
- Introduce communication protocol simulation instead of the performance model, taking advantage of the existing models in the *ns-3* library.
- Study of the stability limit of the macro behaviour of the $nDDE_a$ model, and validation with the simulation tool.

References

[1] Norman Abramson. The aloha system: Another alternative for computer communications. In *Proceedings of the November 17-19, 1970, fall joint computer conference*, pages 281–285, 1970.

[2] M. Bando, K. Hasebe, A. Nakayama, A. Shibata, and Y. Sugiyama. Dynamical model of traffic congestion and numerical simulation, 1995.

[3] Leslie C. Edie. Car-following and steady-state theory for noncongested traffic. *Operations research*, 9(1):66–76, 1961.

[4] Denos C. Gazis, Robert Herman, and Richard W. Rothery. Nonlinear follow-the-leader models of traffic flow. *Operations research*, 9(4):545–567, 1961.

[5] Khalid Abdel Hafeez, Lian Zhao, Bobby Ma, and Jon W. Mark. Performance analysis and enhancement of the dsrc for vanet's safety applications. *IEEE Transactions on Vehicular Technology*, 62(7):3069–3083, 2013.

[6] Jack K. Hale. *Theory of functional differential equations*. New York Springer-Verlag, New York, 1977.

[7] Dirk Helbing. Traffic and related self-driven many-particle systems. *Reviews of modern physics*, 73(4):1067–1141, 2001.

[8] A. Casal J. F. Padial. Bifurcation in car-following models with time delays and driver and mechanic sensitivities. *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales.Serie A, Matemáticas*, 116(4), 2022.

[9] H. K. Lee, H-W Lee, and D. Kim. Macroscopic traffic models from microscopic car-following models. *Physical Review E*, 64(5):056126, 2001.

[10] Takashi Nagatani. The physics of traffic jams. *Reports on progress in physics*, 65(9):1331, 2002.

[11] Gaurang Naik, Biplav Choudhury, and Jung-Min Park. Ieee 802.11 bd and 5g nr v2x: Evolution of radio access technologies for v2x communications. *IEEE access*, 7:70169–70184, 2019.

[12] Gordon Frank Newell. Nonlinear effects in the dynamics of car following. *Operations research*, 9(2):209–229, 1961.

[13] Stefania Santini, Alessandro Salvi, Antonio Saverio Valente, Antonio Pescapé, Michele Segata, and Renato Lo Cigno. A consensus-based approach for platooning with intervehicular communications and its validation in realistic scenarios. *IEEE Transactions on Vehicular Technology*, 66(3):1985–1999, 2016.

[14] Fouad Tobagi and Leonard Kleinrock. Packet switching in radio channels: Part ii-the hidden terminal problem in carrier sense multiple-access and the busy-tone solution. *IEEE Transactions on Communications*, 23(12):1417–1433, 1975.

High-order well-balanced finite volume schemes for 1d and 2d shallow-water equations with Coriolis forces

V. González-Tabernero[1], M. J. Castro[2], J. A. García-Rodríguez[3]

1. *Dpto. de Matemáticas, Facultad de Informática, Universidad de A Coruña, A Coruña, Spain,* v.gonzalez.tabernero@udc.es
2. *Dpto. Análisis Matemático, Estadística e Investigación Operativa y Matemática Aplicada, Universidad de Málaga, Bulevar Louis Pasteur, 31, 29010 Málaga, Spain,* mjcastro@uma.es
3. *Dpto. de Matemáticas, Facultad de Informática, Universidad de A Coruña, A Coruña, Spain,* jose.garcia.rodriguez@udc.es

1. Introduction

The design of well-balanced schemes has been a very interesting and challenging task in the last decades , and continues to be a very active field of research. These schemes are able to preserve some (or all) stationary solutions for a given problem, resulting on a large improvement on accuracy in most cases. One of the first articles that treats this problem is [3]. Other important contributions for the shallow-water system of equations can be found on [1–7, 9, 10, 13] and the references therein. In particular, this system is a balance law that can be written as

$$\partial_t U + \partial_x f(U) = s(x,U), \qquad U \in \mathbb{R}\times\mathbb{R}^+ \to \Omega \subset \mathbb{R}^N, \quad f \in C^1(\mathbb{R}^N,\mathbb{R}^N), \quad s\colon \mathbb{R}\times\Omega \mapsto \mathbb{R}^N.$$

This balance law is discretized in a finite volume framework. In particular, we consider semi-discrete high-order finite-volume schemes:

$$\frac{dU_i}{dt} = -\frac{1}{\Delta x}\left(F_{i+\frac{1}{2}}(t) - F_{i-\frac{1}{2}}(t)\right) + \frac{1}{\Delta x}S_i. \tag{1.1}$$

In this work, we present two novel schemes for the shallow-water equations with Coriolis forces. First, a scheme for the one dimensional case, which is able to preserve the geostrophic stationary solutions. Second, a scheme for the two dimensional case which improves the accuracy of the standard finite volume schemes when the system is close to a stationary solution. We follow the well-balancing method proposed in [8] which will be revisited in Section 2. In section 3 we introduce the first order well-balanced scheme. In Section 4 we describe the second order well-balanced scheme. Finally, in Section 5 we provide some numerical experiments for both schemes.

2. Well-balanced schemes

In the finite volume framework for balance laws, we are interested in the preservation of the stationary solutions of the PDE described by:

$$\nabla \mathbf{f}(\mathbf{u}) = \mathbf{S}(\mathbf{u}). \tag{2.1}$$

According to this interest, we introduce the following definition.

Definition 2.1 (Exactly well-balanced scheme) A finite volume scheme for balance laws is said to be exactly well-balanced if it is able to preserve the exact stationary solutions of the hyperbolic PDE given by (2.1).

We will not always be able to compute the stationary solutions of a given balance law. This difficulty motivates us to introduce a weaker definition:

Definition 2.2 (Well-balanced scheme) A finite volume scheme for balance laws is said to be well-balanced if it is able to preserve a high order aproximation of the exact stationary solution of the PDE.

To preserve the stationary solution we follow the procedure described in [7] and [8]. Here we will describe only the discrete well-balanced procedure. First of all, we write the quadrature formula

$$\mathbf{U}_i^n = \sum_{k=0}^{M} \alpha_k^i \mathbf{U}(x_k^i, t^n) \approx \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} \mathbf{u}(x,t^n)dx,$$

where x_k^i are the quadrature points and α_k^i their weights. This quadrature formula is used to compute all the integrals involved in the numerical scheme and its order must be equal or greater than the reconstruction order. Given the quadrature points in each cell, we introduce the following definition.

Definition 2.3 (Points of interest) Given a (exactly) well-balanced finite volume scheme, we define points of interest of a stencil $\mathcal{S}_i$ as the set of points in the domain $x \in \Omega(\mathcal{S}_i)$ where the local stationary solution has to be evaluated.

Now, we can describe the general discrete well-balanced procedure following [8]. The main idea described here is to obtain a reconstruction operator $P_i(x; \{U_j\})$ for every stencil that satisfies

$$P_i(x) = \mathbf{u}^*(x),$$

when $\mathbf{U}_j = \mathbf{U}_j^*$ where $\mathbf{u}^*(x)$ is the exact solution (or a high order approximation) of (2.1). This means that the reconstruction operator is the stationary solution when applied to the cell averages of the stationary solution.

In order to obtain this reconstruction operator, we must follow the next steps for every stencil in our domain (we drop the time dependency for simplicity):

1. Compute, if possible, the local stationary solution $\mathbf{U}_i^*(x)$ on the stencil of cell I_i, $(\cup_{j\in S_i} I_j)$ defined as the solution of the ODE
$$\nabla \mathbf{f}(\mathbf{U}_i^*) = \mathbf{s}(x, \mathbf{U}_i^*), \ \sum_{k=0}^{M} \alpha_k^i \mathbf{U}_i^*(x_k^i) = \mathbf{U}_i. \tag{2.2}$$
Finding this local stationary solutions is most nightmarish step of this procedure. If it is not possible to found this local steady state on the stencil, we set $U_i^* \equiv 0$.

2. Compute the *fluctuations* $\{\mathbf{V}_j\}_{j\in S_i}$ given by
$$\mathbf{V}_j = \mathbf{U}_j - \sum_{k=0}^{M} \alpha_k^j \mathbf{U}_i^*(x_k^j), \quad j \in \mathcal{S}_i,$$
and compute the reconstruction operator:
$$Q_i(x) = Q_i(x; \{\mathbf{V}_j\}_{j\in\mathcal{S}_i}).$$

3. Finally, define
$$P_i(x) = \mathbf{U}_i^*(x) + Q_i(x). \tag{2.3}$$

Once we have built the well-balanced reconstruction operator, we have to modify the semidiscrete finite volume scheme in the following way:

$$\begin{aligned}\frac{d\mathbf{U}_i}{dt} =& \frac{1}{\Delta x}\Big(\mathbf{F}_{i-1/2} - \mathbf{f}(\mathbf{U}_i^{t,*}(x_{i-1/2})) - \mathbf{F}_{i+1/2} + \mathbf{f}(\mathbf{U}_i^{t,*}(x_{i+1/2}))\\ &+ \sum_{k=0}^{M} \alpha_k^i \left(\mathbf{S}(P_i^t(x_k^i)) - \mathbf{S}(\mathbf{U}_i^{t,*}(x_k^i))\right) dx\Big).\end{aligned} \tag{2.4}$$

Which results in following theorem that can be proved (see [8, 11]).

Theorem 2.4 *If the reconstruction operator is well-balanced for a stationary solution $U^*(x)$, then the numerical method* (2.4) *is exactly well-balanced in the sense of the Definition* 2.1.

Remark 2.5 It is important to notice that (2.2) and (2.4) determine the points of interest of our problem (see Definition 2.3). We need to know the stationary solution at the points of quadrature in every cell of the stencil in order to calculate average values of the stationary solution and the source term. Also, we need to know the value of the stationary solution at the two intercells $x_{i\pm1/2}$ in order to compute the fluxes.

3. Well-balanced scheme for the 1d shallow-water equations with Coriolis forces

In the case of the one dimensional shallow-water equations with Coriolis forces, the hyperbolic system reads:

$$\partial_t \begin{bmatrix} h \\ hu \\ hv \end{bmatrix} + \partial_x \begin{bmatrix} hu \\ hu^2 + gh^2/2 \\ huv \end{bmatrix} = \begin{bmatrix} 0 \\ fhv - gh\partial_x z \\ -fhu \end{bmatrix},$$

where h is the fluid depth, hu and hv are the horizontal linear moments, z is the bottom topography, g and f are the gravity and Coriolis constant, respectively.

If $hu \neq 0$, the geostrophic stationary solutions [10] of the system are given by:

$$\partial_x \begin{bmatrix} h^* u^* \\ (u^*)^2/2 + g(h^* + z) \\ v^* \end{bmatrix} = \begin{bmatrix} 0 \\ f v^* \\ -f \end{bmatrix}, \tag{3.1}$$

we recall that our main interest is to determine, locally, this stationary solutions, therefore we can follow the procedure described in the previous section to obtain the well-balanced reconstruction operator.

This ODE system describing the stationary solutions can be solved locally in order to determine the local stationary solution. First equation leads to $(hu)^*(x) = hu_i$. Third equation indicates that the velocity v is linear, in this case

$$v^*(x) = -f(x - x_i) + v_i. \tag{3.2}$$

These two solutions are easily computed in any point of interest, notice that we have neither set a reconstruction order nor quadrature formula. The main difficulty arises from the calculation of h^*, which is described by the second equation. This equation indicates that the stationary value of the water depth $h^*(x_p)$ must be a root of the cubic equation

$$\zeta(h^*, v_i, x_p, E_i) = g(h^*)^3(x_p) + \left(gz(x_p) - fV^*(x_p) - E_i\right)(h^*)^2(x_p) + \frac{(hu)_i}{2}, \tag{3.3}$$

where x_p is any point of interest of our problem, $V_i(x)$ is a primitive function of $v^*(x)$ and E_i is the local energy of the system. Notice that (3.3) has always a negative root and may have one, two or no positive roots. Here, we follow the same procedure that the one described in [6] to choose the appropriate value root. Following this procedure, we have an exactly well-balanced scheme according to Definition 2.1.

3.1. Well-balanced scheme for first and second order reconstructions

In this case, we can use the midpoint quadrature formula. Therefore, we can relate the values of the cell averages with the values of the solution at the cell centre. This means that the velocity v^* is

$$v^*(x) = -fx + \frac{(hv)_i}{h_i} + fx_i,$$

so the potential energy associated to v^* can be calculated by a direct integration:

$$V_i^*(x) = -f\frac{x^2}{2} + (fx_i + v_i)x.$$

And using (3.3) we can calculate the local energy at the cell centre

$$E_i = \frac{(u_i)^2}{2} + g\left(h_i + z(x_i) - \frac{f}{g}V_i^*(x_i)\right),$$

where $u_i = (hu)_i/h_i$. Knowing the local energy, we can find the rest of values of the water depth $h^*(x)$ in the points of interest x_p by solving the cubic equation (3.3).

3.2. Well-balanced scheme for third order reconstruction

In this case, we are assuming a third order CWENO reconstruction. The main issue arises from the quadrature formula, we need at least two points of quadrature in each cell to compute the integrals. So, in this case, we can not calculate point-wise values of the velocity field, and accordingly, we can not calculate the local energy by a direct substitution of values.

First, we multipliy (3.2) and h^*, we derive the following relation

$$v_i + fx_i = \frac{2(hv)_i + fx_i^0 h^*(x_i^0) + fx_i^1 h^*(x_i^1)}{h^*(x_i^0) + h^*(x_i^1)},$$

where x_i^0 and x_i^1 are the two quadrature points on the i-th cell. Then, the primitive is

$$V_i^*(x) = -f\frac{x^2}{2} + \left(\frac{2(hv)_i + fx_0 h^*(x_i^0) + fx_1 h^*(x_i^1)}{h^*(x_i^0) + h^*(x_i^1)}\right)x.$$

Now we introduce the following functions

$$\omega_0(x_i^0, x_i^1, E_i) = 0.5\left(h^*(x_i^0) + h^*(x_i^1)\right) + h_i,$$
$$\omega_1(x_i^0, x_i^1, E_i) = \zeta(h^*(x_i^0), v_i, x_i^0, E_i),\ \omega_2(x_i^0, x_i^1, E_i) = \zeta(h^*(x_i^1), v_i, x_i^1, E_i).$$

We recall the dependency of V^* on both $h^*(x_i^0)$ and $h^*(x_i^1)$. Finally, we solve the non-linear system of equations $(\omega_0, \omega_1, \omega_2) = (0, 0, 0)$ for $(h^*(x_i^0), h^*(x_i^1), E_i)$. Once this solution is computed, the other values of h^* at the points of interest are obtained by solving the cubic equation (3.3) with the local energy E_i obtained from this procedure.

4. Well-balanced scheme for the 2d shallow-water equations with Coriolis forces

The two dimensional shallow-water system reads:

$$\partial_t \begin{bmatrix} h \\ hu \\ hv \end{bmatrix} + \partial_x \begin{bmatrix} hu \\ hu^2 + \frac{gh^2}{2} \\ huv \end{bmatrix} + \partial_y \begin{bmatrix} hv \\ huv \\ hv^2 + \frac{gh^2}{2} \end{bmatrix} = \begin{bmatrix} 0 \\ -gh\partial_x z(x,y) \\ -gh\partial_y z(x,y) \end{bmatrix} + \begin{bmatrix} 0 \\ fhv \\ -fhu \end{bmatrix}.$$

In this case, we deal only with 1st and 2nd order schemes in rectangular meshes. We are interested in the set of geostrophic stationary solutions (see [5]) given by:

$$\partial_x u^* + \partial_y v^* = 0, \tag{4.1}$$
$$\partial_x (h^* + z) = fv^*/g,\ \partial_y (h^* + z) = -fu^*/g. \tag{4.2}$$

The main difficulty on this system is that it is an eliptic problem, while the one dimensional case (3.1) is an EDO. Also, we have a null divergence equation (4.1) which is very restrictive for our discrete system. Also, this system allows infinite functional forms for u and v while we only know hu and hv.

Our approach consists on modifying the continuous equation (4.1) with a discrete version of itself:

$$\delta_x U^*_{i,j} + \delta_y V^*_{i,j} = 0. \tag{4.3}$$

Where

$$\delta_x U_{i,j} = \frac{U_{i+1,j} - U_{i-1,j}}{\Delta x},\ \delta_y U_{i,j} = \frac{U_{i,j+1} - U_{i,j-1}}{\Delta y}.$$

As we are applying first and second order reconstructions, we can compute:

$$(u, v)_{i,j} = \left(\frac{hu}{h}, \frac{hv}{h}\right)_{i,j},$$

with these values, we check equation (4.3). If it is satisfied, then $U^* = U$ and $V^* = V$. In the other case, we calculate the values $U^*_{i\pm1,j}$ and $V^*_{i,j\pm1}$ such that:

$$\delta_x U^*_{i,j} + \delta_y V^*_{i,j} = 0,$$
$$\min_{U^*,V^*}\left[\sum_{k=-1,1}\left((U^*_{i+k,j} - U_{i+k,j})^2 + (V^*_{i,j+k} - V_{i,j+k})^2\right)\right].$$

Once known the point values of the stationary velocity field, we can calculate the functional form and the water depth:

1. In the case of the first order reconstruction, our points of interest are the intercells and the cell centre. In this case, we assume affine velocities:

$$q^*_{i,j}(x) = q_{i,j} + \delta_x q_{i,j} + \delta_y q_{i,j},\ \text{for } q = u, v. \tag{4.4}$$

Then, we can calculate the stationary water depth by a direct integration using (4.2):

$$h^*(x,y) = h_i + \int_{C_{i,j}} z dV + \frac{f}{g}\left(v_{i,j}(x - x_i) + \frac{\delta v_x}{2}(x - x_i)^2 + \delta v_y (y - y_j)(x - x_i)\right.$$
$$\left. -u_{i,j}(y - y_j) - \frac{\delta u_y}{2}(y - y_j)^2\right) - z(x,y).$$

2. For the second order reconstruction we need to know the values of the stationary solution in the adjacent cells. In this case, we assume picewise affine velocities such that the conservation property is satisfied in the whole stencil. This is acquired by imposing that the velocity field is (4.4) in the middle cell and extending this function linearly from the midpoints to the centres of the adjacent cells. After so, one can calculate h^* by a direct integration of this function.

Attending to definitions 2.1 and 2.2, our scheme is exactly well-balanced for an affine velocity field with h^* fullfilling (4.2). While the scheme will be well-balanced for any other geostrophic stationary solution.

5. Numerical experiments

In this section we are going to provide numerical experiments that test both schemes. In both problems (1d and 2d), we use an HLL Riemann solver ([12]) with the wave speed estimates given by the procedure from [2]. Also a CFL condition is applied:

$$\Delta t \leq CFL \min_{\text{all cells}} \left(\frac{\Delta V}{\max_k \lambda_k \}} \right),$$

where λ_k is the wave speed in each intercell and ΔV is the volume of the cell in the selected mesh. For the second order scheme we choose a minmod limiter, in the 1d and 2d cases. For the third order scheme a CWENO reconstruction is applied. The errors are calculated under the $L_1(\Omega)$ norm, $E_N = \|\mathbf{U}(x,T) - \mathbf{u}^*(x)\|_{L_1(\Omega_N)}$, where $\mathbf{u}^*$ is the exact solution when known.

5.1. 1d geostrophic stationary solutions test

For this numerical experiment, we consider the steady-state proposed in [10]:

$$h^*(x) = e^{2x},\ u^*(x) = \frac{e^{-2x}}{2},\ v^*(x) = -fx,\ z(x) = -\frac{f^2}{2}x^2 - e^{2x} - \frac{e^{-4x}}{8}.$$

The numerical experiment will be initizialised under this stationary state. While we consider $f = g = 1$, $CFL = 0.8$, $T = 5s$ and domain $x \in [0,1]$. We consider Neumann boundary conditions. In table 1 we can find the results for $N = 25, 50, 100, 200, 400$. We see how our scheme is exactly well-balanced for this stationary solution as the errors are of the order of the machine error. It is important to notice that a small accumulation error rises in the third order scheme due to the non-linear system that has to be solved.

Variable	h			hu			hv		
Reconstruction / N	$O(1)$	$O(2)$	$O(3)$	$O(1)$	$O(2)$	$O(3)$	$O(1)$	$O(2)$	$O(3)$
25	0e+00	0e+00	5e-14	2e-16	6e-16	1e-14	1e-15	7e-16	5e-14
50	4e-16	0e+00	9e-14	3e-15	3e-16	7e-15	4e-14	6e-16	1e-13
100	1e-15	0e+00	3e-13	6e-15	2e-16	7e-14	9e-14	4e-16	2e-13
200	1e-14	0e+00	8e-13	4e-14	5e-16	2e-13	4e-13	4e-16	5e-13
400	1e-17	1e-17	2e-12	3e-16	3e-16	6e-13	2e-15	4e-16	9e-13

Tab. 1 L_1 errors for the 1st, 2nd and 3rd order schemes.

5.2. 2d affine geostrophic stationary solutions

For this experiment we choose affine velocities and fix h and z to be a geostrophic stationary solution:

$$u^*(x,y) = 0.5 + 0.1x + 0.05y,\ v^*(x,y) = 0.4 + 0.2x - 0.1y,$$
$$z(x,y) = \frac{f}{g}\left(0.1x^2 - 0.025y^2 - 0.1xy\right),\ h^*(x,y) = 3 + \frac{f}{g}\left(0.4x - 0.5y\right).$$

We set $CFL = 0.9$, $f = g = 1$, a domain $\Omega = [0,1] \times [0,1]$ and a final time $T = 1$. The mesh chosen is $N \times N$ with $N = 10, 20, 40, 80, 160$. In Table 2 we show the obtained errors. We observe that the scheme is exactly well-balanced as we expected.

N	h		hu		hv	
	$O(1)$	$O(2)$	$O(1)$	$O(2)$	$O(1)$	$O(2)$
10	1e-15	1e-15	2e-15	7e-16	2e-15	6e-16
20	1e-15	1e-15	3e-15	2e-15	4e-15	2e-15
40	2e-15	2e-15	6e-15	1e-15	5e-15	1e-15
80	5e-15	5e-15	1e-14	1e-15	1e-14	1e-15
160	2e-15	2e-15	1e-14	1e-15	2e-14	1e-15

Tab. 2 L_1 errors for the first and second order schemes.

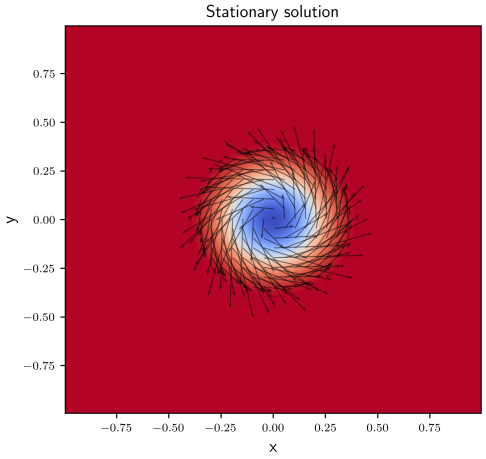


Fig. 1 Exact stationary solution.

5.3. 2d vortex stationary solution

For this last experiment, we choose a stationary vortex proposed in [9], with:

$$h(x,y) = 1 + \varepsilon^2 \Psi(x,y),\ u(x,y) = -\varepsilon y \Theta(x,y),\ v(x,y) = \varepsilon x \Theta(x,y),$$

$$\Psi(x,y) = \begin{cases} 2.5(1+5\varepsilon^2)r^2, & \text{if } r < \frac{1}{5}, \\ \frac{1+5\varepsilon}{10} + 2r - 0.3 - 2.5r^2 + \varepsilon^2\left(4\log(5r) + \frac{7}{2} - 20r + \frac{25}{2}r^2\right) & \text{if } \frac{1}{5} \le r < \frac{2}{5}, \\ \frac{1-10\varepsilon^2+20\varepsilon^2\log(2)}{5}, & \text{if } \frac{2}{5} \le r, \end{cases}$$

$$\Theta(x,y) = \begin{cases} 5 & \text{if } r < \frac{1}{5}, \\ \frac{2}{r} + 5 & \text{if } \frac{1}{5} \le r < \frac{2}{5}, \\ 0 & \text{if } \frac{2}{5} \le r, \end{cases}$$

where $r = \sqrt{x^2 + y^2}$. We consider $(x,y) \in [-1,1] \times [-1,1]$, $T = 10s$, $CFL = 0.5$. We fix $g = 1/\varepsilon^2$ and $f = 1/\varepsilon$, for $\varepsilon = 0.05$. Also we take null flux boundaries and a flat bottom topography. In figures 2-3 we show the comparison between the stationary solution and the the 1st and 2nd order schemes in the WB and non WB cases. The WB schemes preserve much better the shape of the vortex while the non WB schemes break its symmetries. Also, the WB second order scheme performs much better than the WB 1st order one.

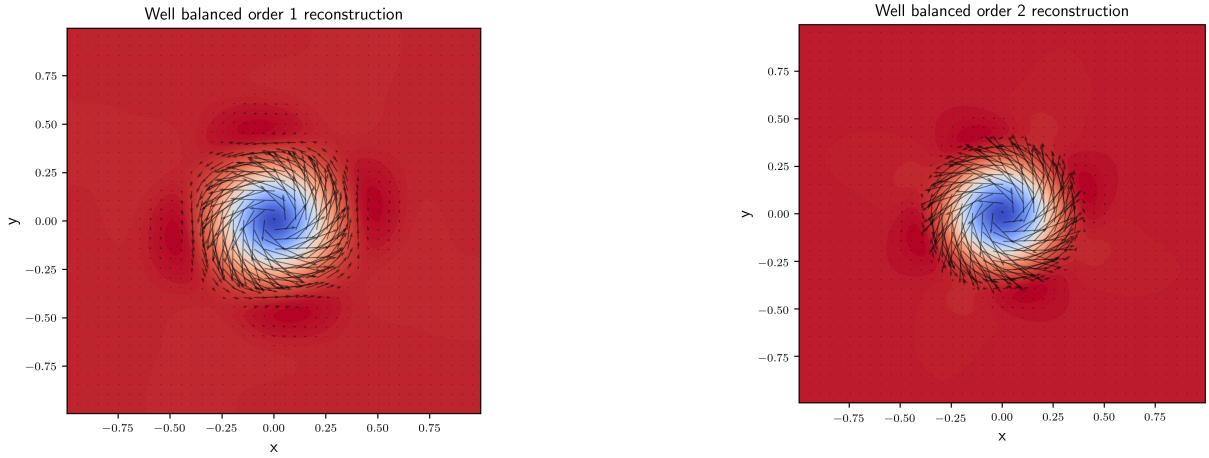


Fig. 2 Stationary solutions. WB first order, left, WB second order, right.

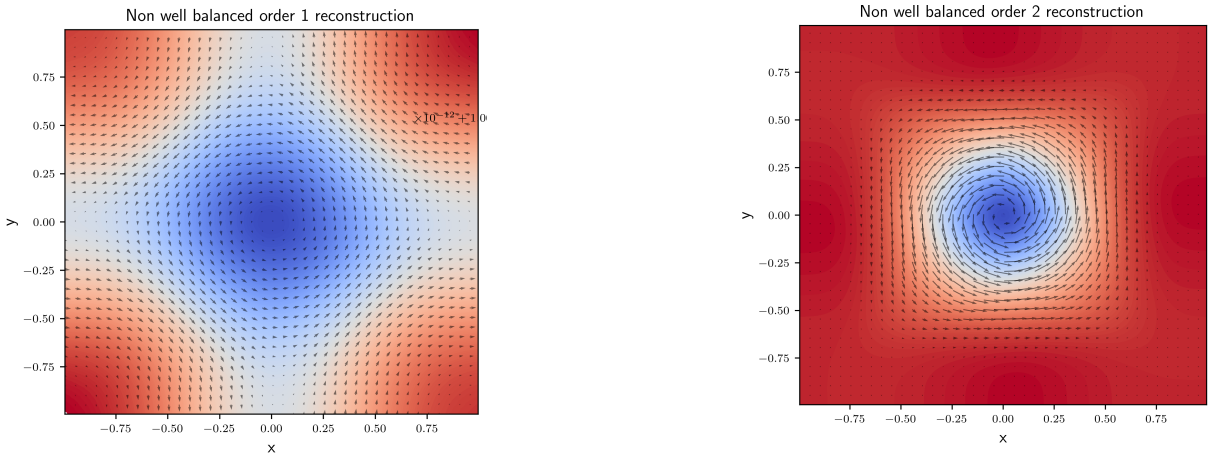


Fig. 3 Stationary solutions. Non WB first order, left, non WB second order, right.

References

[1] E. Audusse, R. Klein, D. D. Nguyen S. and Vater: Preservation of the discrete geostrophic equilibrium in shallow water flows. In Fort, J., Furst, J., Halama, J., Herbin, R., and Hubert, F., editors, Finite Volumes for Complex Applications VI Problems & Perspectives, pages 59–67 (2011). Berlin, Heidelberg. Springer Berlin Heidelberg.

[2] M. de la Asunción, M. J. Castro, E.D. Fernández-Nieto, J. M. Mantas, S. O. Acosta, J. M. González-Vida: Efficient GPU implementation of a two waves TVD-WAF method for the two-dimensional one layer shallow water system on structured meshes, Computers & Fluids, Vol. 80, pp. 441-452 (2013), ISSN: 0045-7930,

[3] A. Bermúdez, A. and M. E. Vázquez: Upwind methods for hy perbolic conservation laws with source terms. Computers & Fluids, 23(8):1049–1071 (1994).

[4] F. Bouchut, J. L. Sommer and V. Zeitlin: Frontal geostrophic adjustment and nonlinear wave phenomena in one-dimensional rotating shallow water. part 2. high-resolution numerical simulations. Journal of Fluid Mechanics, 514:35–63 (2004).

[5] A. Buttinger-Kreuzhuber, Z. Horváth, S. Noelle, G. Blöschl, J. Waser: A fast second-order shallow water scheme on two-dimensional structured grids over abrupt topography. Advances in Water Resources, vol. 127, pp. 89-108 (2019), ISSN: 0309-1708

[6] M. J. Castro Díaz, C. Chalons and T. Morales de Luna: A Fully Well-Balanced Lagrange–Projection-Type Scheme for the Shallow-Water Equations, SIAM Journal on Numerical Analysis, vol. 56, num. 5, pp 3071-3098 (2018), DOI: 10.1137/17M1156101

[7] M. J. Castro, J. A. López and C. Parés: Finite Volume Simulation of the Geostrophic Adjustment in a Rotating Shallow-Water System. SIAM Journal on Scientific Computing, Vol. 31, num. 1, pp. 444-477 (2008), DOI: 10.1137/070707166.

[8] M.J. Castro and C. Parés: Well-Balanced High-Order Finite Volume Methods for Systems of Balance Laws. J. Sci. Comput, Vol 82, num 48, pp. 939–973 (2020)

[9] A. Chertock, M. Dudzinski, A. Kurganov et al.: Well-balanced schemes for the shallow water equations with Coriolis forces, Numer. Math. 138, 939–973 (2018).

[10] V. Desveaux and A. Masset: A fully well-balanced scheme for shallow water equations with Coriolis force, Comm. Math. Sci., vol. 20, num.7 (2022), pp. 1875 – 1900, DOI: https://dx.doi.org/10.4310/CMS.2022.v20.n7.a4

[11] I. Gómez-Bueno, M. J. Castro, C. Parés and G. Russo: Collocation Methods for High-Order Well-Balanced Methods for Systems of Balance Laws, Math. Journal, vol. 9, num. 15 (2021), ISSN: 2227-7390, DOI: 10.3390/math9151799

[12] A. Harten, P. D. Lax and B. van Leer: On Upstream Differencing and Godunov-Type Schemes for Hyperbolic Conservation Laws, Springer Berlin Heidelberg, pp. 53–79 (1997).

[13] Y. Xing and C.-W. Shu: High order finite difference weno schemes with the exact conservation property for the shallow water equations. Journal of Computational Physics, 208(1):206–227 (2005).

Weak solutions to the total variation flow in metric measure spaces

Wojciech Górny[1,2], **José M. Mazón**[3]

1. *Faculty of Mathematics, University of Vienna, Austria,* wojciech.gorny@univie.ac.at
2. *Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland*
3. *Department of Mathematical Analysis, University of València, Spain*

Abstract

In this paper, we consider the total variation flow on bounded domains in metric measure spaces. For Neumann or Dirichlet boundary conditions, we generalise the notion of weak solutions to the metric case using the first-order linear differential structure due to Gigli and a version of the Gauss-Green formula. Moreover, we provide a notion of solutions to the Neumann problem which is valid for L^1 initial data.

1. Introduction

The total variation flow in an open bounded set $\Omega \subset \mathbb{R}^N$ is formally the equation

$$u_t = \operatorname{div}\left(\frac{Du}{|Du|}\right) \qquad \text{in } \Omega \times (0,\infty). \tag{1.1}$$

Since its introduction in the seminal work [18] by Rudin, Osher and Fatemi in order to solve the denoising problem, it has remained one of the most popular tools in image processing. From the mathematical point of view, the natural setting to look for solutions to (1.1) is to require that for a.e. $t \in (0,\infty)$ the function $u(\cdot,t)$ is a function of bounded variation, i.e. its distributional derivative is a Radon measure. The main difficulty is that the operator on the right-hand side of (1.1) is degenerate, so it is not immediately clear how to properly define the right-hand side of equation (1.1), taking into account that the denominator may disappear on a set of positive Lebesgue measure; actually, this phenomenon of formation of facets it a typical property of solutions to the total variation flow. A characterisation of solutions was established in the monograph [4] by replacing $\frac{Du}{|Du|}$ with a vector field with integrable divergence which agrees $|Du|$-a.e. with the Radon-Nikodym derivative $\frac{dDu}{d|Du|}$, using the classical theory of maximal monotone operators due to Brezis [7], the Crandall-Liggett theorem [9], and a Gauss-Green formula for vector fields with integrable divergence and BV functions due to Anzellotti [5].

The study of gradient flows in metric spaces faces some additional difficulties (for a standard reference, see [2]) and requires very different methods from the Euclidean setting. The reason is that, in general, in a metric space directions are not well-defined (even locally), so it is unclear how to define directional derivatives of a function. In general, in place of a derivative of a Lipschitz or Sobolev function, one can use one of several equivalent objects (such as the minimal upper gradient) which in the Euclidean case correspond to the length of the gradient. Therefore, in this way we do not obtain a linear structure, and consequently the definitions of solutions necessarily avoid direct use of the directional derivatives. A classical problem of this type is the heat flow or the p-Laplacian evolution equation: it has been studied by Ambrosio, Gigli and Savaré in a series of papers (see for instance [3]) using the semigroup approach. In a complete and separable metric measure space $(\mathbb{X}, d, \nu)$, under mild assumptions on the measure ν, the authors define it as the gradient flow in $L^2(\mathbb{X},\nu)$ of the Dirichlet-Cheeger energy and study its properties under the assumption that $(\mathbb{X}, d, \nu)$ has Ricci curvature bounded from below. In these papers, the gradient flow in $L^2(\mathbb{X},\nu)$ is defined in the framework of maximal monotone operators in Hilbert spaces and the corresponding p-Laplacian operator is defined through the subdifferential of the p-Cheeger energy (but without giving any direct characterisation); some basic properties of the total variation flow were obtained using the same method by Ambrosio and di Marino in [1].

One definition of solutions known in the Euclidean case which proved particularly difficult to generalise to the metric setting was the definition of weak solutions. In this paper, we summarize the results obtained in [14] (see also [12] for a similar study of the p-Laplacian evolution equation), and show how to introduce weak solutions to the total variation flow in metric measure spaces. Using the first-order differential structure on a metric measure space introduced by Gigli, we characterise the subdifferential in $L^2(\mathbb{X},\nu)$ of the total variation. This leads to a new definition of solutions to the total variation flow in metric measure spaces, in which a crucial role is played by a vector field (defined via Gigli's differential structure) satisfying some compatibility conditions. We provide a characterisation of solutions and prove their existence and uniqueness, using the classical theory of maximal monotone operators when the data is in $L^2(\Omega,\nu)$ and the theory of completely accretive operators for data in $L^1(\Omega,\nu)$, in the following three situations: the evolution with Neumann boundary conditions with initial data in $L^2(\Omega,\nu)$; with Dirichlet boundary condition in $L^1(\partial\Omega, |D\chi_\Omega|_\nu)$ and initial data in $L^2(\Omega,\nu)$; and finally we introduce the notion of entropy solutions to the Neumann problem and allow for initial data in $L^1(\Omega,\nu)$.

2. Preliminaries

2.1. Standing assumptions

Throughout the paper, we assume that the metric space $(\mathbb{X}, d)$ is complete and separable. Furthermore, we require that ν is a doubling measure and $\mathbb{X}$ supports a weak $(1,1)$-Poincaré inequality. We restrict the presentation of the notions from analysis on metric spaces to a minimum; for more details, we refer to [6] and [11].

2.2. Sobolev and BV spaces

Given a metric measure space $(\mathbb{X}, d, \nu)$ and $p \in [1, \infty)$, there are several possible definitions of Sobolev spaces on $\mathbb{X}$; nonetheless, under the assumptions of this paper, all of them agree (see [3]), and to simplify the presentation we choose to use the Newtonian approach. We follow the presentation in [6]; we say that a Borel function g is an *upper gradient* of a Borel function $u : \mathbb{X} \to \mathbb{R}$ if for all curves $\gamma : [0, l_\gamma] \to \mathbb{X}$ we have

$$\left|u(\gamma(l_\gamma)) - u(\gamma(0))\right| \leq \int_\gamma g := \int_0^{l_\gamma} g(\gamma(t))\, |\dot\gamma(t)|\, dt\, ds,$$

where

$$|\dot\gamma(t)| := \lim_{\tau \to 0} \frac{\gamma(t+\tau) - \gamma(t)}{\tau}$$

is the *metric speed* of γ. If this inequality holds for p-almost every curve, i.e. the p-modulus (see for instance [6, Definition 1.33]) of the family of all curves for which it fails equals zero, then we say that g is a *p-weak upper gradient* of u. The Sobolev space $W^{1,p}(\mathbb{X}, d, \nu)$ is defined as

$$W^{1,p}(\mathbb{X}, d, \nu) := \left\{ u \in L^p(\mathbb{X}, \nu) : \text{ there exists an upper gradient } g \in L^p(\mathbb{X}, \nu) \right\}.$$

In the literature, this space is sometimes called the Newton-Sobolev space (or Newtonian space) and is denoted $N^{1,p}(\mathbb{X})$. For every $u \in W^{1,p}(\mathbb{X}, d, \nu)$, there exists a minimal p-weak upper gradient $|Du| \in L^p(\mathbb{X}, \nu)$, i.e. we have $|Du| \leq g$ ν-a.e. for all p-weak upper gradients $g \in L^p(\mathbb{X}, \nu)$ (see [6]). It is unique up to a set of measure zero. The space $W^{1,p}(\mathbb{X}, d, \nu)$ is endowed with the norm

$$\|u\|_{W^{1,p}(\mathbb{X},d,\nu)} = \left(\int_{\mathbb{X}} |u|^p \, d\nu + \int_{\mathbb{X}} |Du|^p \, d\nu \right)^{1/p}.$$

Also for functions of bounded variation there are several different ways to introduce them in metric measure spaces, but again under the assumptions of this paper they are equivalent (see [1]). In this paper, we follow the definition of total variation introduced by Miranda in [16]. For $u \in L^1(\mathbb{X}, \nu)$, we define the total variation of u on an open set $\Omega \subset \mathbb{X}$ by the formula

$$|Du|_\nu(\Omega) := \inf \left\{ \liminf_{n\to\infty} \int_\Omega g_{u_n} \, d\nu \ : \ u_n \in \mathrm{Lip}_{\mathrm{loc}}(\Omega),\ u_n \to u \text{ in } L^1(\Omega, \nu) \right\}, \tag{2.1}$$

where g_{u_n} is a 1-weak upper gradient of u. The total variation $|Du|_\nu(\mathbb{X})$ defined by formula (2.1) is lower semicontinuous with respect to convergence in $L^1(\mathbb{X}, \nu)$. The space of functions of bounded variation $BV(\mathbb{X}, d, \nu)$ consists of all functions $u \in L^1(\mathbb{X}, \nu)$ such that $|Du|_\nu(\mathbb{X}) < \infty$. It is a Banach space with respect to the norm

$$\|u\|_{BV(\mathbb{X},d,\nu)} := \|u\|_{L^1(\mathbb{X},\nu)} + |Du|_\nu(\mathbb{X}).$$

We turn our attention to the definition of the boundary measure of an open set in a metric measure space. A set $E \subset \mathbb{X}$ is said to be of finite perimeter if $\chi_E \in BV(\mathbb{X}, d, \nu)$, and its perimeter is defined as

$$\mathrm{Per}_\nu(E) := |D\chi_E|_\nu(\mathbb{X}).$$

Another common way to define the boundary measure in metric measure spaces in the *codimension one Hausdorff measure*. Given a set $A \subset \mathbb{X}$, it is defined as

$$\mathcal{H}(A) := \liminf_{R\to 0} \left\{ \sum_{i=1}^\infty \frac{\nu(B(x_i, r_i))}{r_i} \ : \ A \subset \bigcup_{i=1}^\infty B(x_i, r_i),\ 0 < r_i \leq R \right\}.$$

If $E \subset \mathbb{X}$ is a set of finite perimeter, then for any Borel set $A \subset \mathbb{X}$ we have $\frac{1}{C}\mathcal{H}(A \cap \partial_* E) \leq |D\chi_E|_\nu(A) \leq C\mathcal{H}(A \cap \partial_* E)$, where $\partial_* E$ is the measure-theoretic boundary of E, i.e. the set of all $x \in \mathbb{X}$ for which simultaneously

$$\limsup_{r\to 0^+} \frac{\nu(B(x,r) \cap E)}{\nu(B(x,r))} > 0 \quad \text{and} \quad \limsup_{r\to 0^+} \frac{\nu(B(x,r) \setminus E)}{\nu(B(x,r))} > 0.$$

In particular, if $\partial_*\Omega = \partial\Omega$, the spaces $L^p(\partial\Omega, \mathcal{H})$ and $L^p(\partial\Omega, |D\chi_\Omega|_\nu)$ coincide as sets for every $p \in [1, \infty]$, and are equipped with equivalent norms. Definition of boundary values of BV functions in a metric measure space is a more delicate issue; we restrict our attention to open sets and adopt the following definition.

Definition 2.1 Let $\Omega \subset \mathbb{X}$ be an open set and let u be a ν-measurable function on Ω. A number $T_\Omega u(x)$ is a *trace* of u at $x \in \partial\Omega$ if

$$\lim_{r\to 0^+} \fint_{\Omega\cap B(x,r)} |u - T_\Omega u(x)|\, d\nu = 0.$$

We say that u has a trace in $\partial\Omega$ if $T_\Omega u(x)$ exists for $\mathcal{H}$-almost every $x \in \partial\Omega$.

Well-posedness of the trace and identifying the trace space of $W^{1,1}(\Omega, d, \nu)$ or $BV(\Omega, d, \nu)$ in the setting of metric measure spaces is not immediate and requires additional structural assumptions on Ω. We summarize the results known in the literature in the following Theorem (see [15] and [17]).

Theorem 2.2 *Let Ω be an open bounded set which supports a weak $(1,1)$-Poincaré inequality. Assume that Ω additionally satisfies the measure density condition, i.e. there is a constant $C > 0$ such that*

$$\nu(B(x,r) \cap \Omega) \geq C\nu(B(x,r))$$

for $\mathcal{H}$-a.e. $x \in \partial\Omega$ and every $r \in (0, \operatorname{diam}(\Omega))$. Moreover, assume that $\partial\Omega$ is Ahlfors codimension 1 regular, i.e. there is a constant $C > 0$ such that

$$C^{-1}\frac{\nu(B(x,r))}{r} \leq \mathcal{H}(B(x,r)\cap\partial\Omega) \leq C\frac{\nu(B(x,r))}{r}$$

for all $x \in \partial\Omega$ and every $r \in (0, diam(\Omega))$.

Under these assumptions, Definition 2.1 defines an operator $T_\Omega : BV(\Omega, d, \nu) \twoheadrightarrow L^1(\partial\Omega, \mathcal{H})$. Moreover, the operator T_Ω is linear, bounded and surjective.

Under the same assumptions there is a (nonlinear) bounded extension operator $\operatorname{Ext} : L^1(\partial\Omega, \mathcal{H}) \to BV(\Omega, d, \nu)$ such that $T_\Omega \circ \operatorname{Ext}$ is the identity operator on $L^1(\partial\Omega, \mathcal{H})$. When $\partial\Omega$ is Ahlfors codimension one regular, we have $\partial_*\Omega = \partial\Omega$, so the spaces $L^p(\partial\Omega, \mathcal{H})$ and $L^p(\partial\Omega, |D\chi_\Omega|_\nu)$ coincide as sets and have equivalent norms.

2.3. The differential structure

We follow Gigli [11] and Buffa-Comi-Miranda [8] in the introduction of a first-order differential structure on a metric measure space $(\mathbb{X}, d, \nu)$.

Definition 2.3 We define the cotangent module to $\mathbb{X}$ as

$$\mathrm{PCM}_p = \left\{ \{(f_i, A_i)\}_{i\in\mathbb{N}} \,:\, (A_i)_{i\in\mathbb{N}} \subset \mathcal{B}(\mathbb{X}),\ f_i \in W^{1,p}(A_i),\ \sum_{i\in\mathbb{N}} \int_{A_i} |Df_i|^p\, d\nu < \infty \right\},$$

where A_i is a partition of $\mathbb{X}$. We define the equivalence relation $\sim$ as

$$\{(A_i, f_i)\}_{i\in\mathbb{N}} \sim \{(B_j, g_j)\}_{j\in\mathbb{N}} \quad \text{if} \quad |D(f_i - g_j)| = 0 \ \ \nu - \text{a.e. on } A_i \cap B_j.$$

Consider the map $|\cdot|_* : \mathrm{PCM}_p/\!\sim \longrightarrow L^p(\mathbb{X}, \nu)$ given by

$$|\{(f_i, A_i)\}_{i\in\mathbb{N}}|_* := |Df_i|$$

ν-everywhere on A_i. It is called *pointwise norm* on $\mathrm{PCM}_p/\!\sim$. We define the norm $\|\cdot\|$ in $\mathrm{PCM}_p/\!\sim$ as

$$\|\{(f_i, A_i)\}_{i\in\mathbb{N}}\|^p = \sum_{i\in\mathbb{N}} \int_{A_i} |Df_i|^p$$

and set $L^p(T^*\mathbb{X})$ to be the closure of $\mathrm{PCM}_p/\!\sim$ with respect to this norm. The space $L^p(T^*\mathbb{X})$ is called the *cotangent module* and its elements will be called *p-cotangent vector fields*. It is a $L^p(\nu)$-normed module; we denote by $L^q(T\mathbb{X})$ the dual module of $L^p(T^*\mathbb{X})$, namely $L^q(T\mathbb{X}) := \mathrm{HOM}(L^p(T^*\mathbb{X}), L^1(\mathbb{X}, \nu))$, which is a $L^q(\nu)$-normed module. The elements of $L^q(T\mathbb{X})$ will be called *q-vector fields* on $\mathbb{X}$. The duality between $\omega \in L^p(T^*\mathbb{X})$ and $L \in L^q(T\mathbb{X})$ will be denoted by $\omega(X) \in L^1(\mathbb{X}, \nu)$. Since the module $L^p(T^*\mathbb{X})$ is reflexive, we can identify

$$L^q(T\mathbb{X})^* = L^p(T^*\mathbb{X}),$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

Definition 2.4 Given $f \in W^{1,p}(\mathbb{X}, d, \nu)$ we can define its *differential* df as an element of $L^p(T^*\mathbb{X})$ as

$$df = (f, \mathbb{X}).$$

Clearly, the operation of taking the differential is linear as an operator from $W^{1,p}(\mathbb{X}, d, \nu)$ to $L^p(T^*\mathbb{X})$; moreover, from the definition of the norm in $L^p(T^*\mathbb{X})$ it is clear that this operator is bounded with norm equal to one. Furthermore, again from the definition of the pointwise norm, it is clear that

$$|df|_* = |Df| \qquad \nu\text{-a.e. on } \mathbb{X} \text{ for all } f \in W^{1,p}(\mathbb{X}, d, \nu).$$

Now, following [8], we define the divergence of a vector field, in the case when it can be represented by an integrable function. For $\frac{1}{r} + \frac{1}{s} = 1$, we set

$$\mathcal{D}^{q,r}(\mathbb{X}) = \left\{ X \in L^q(T\mathbb{X}) : \ \exists f \in L^r(\mathbb{X}, \nu) \quad \forall g \in W^{1,p}(\mathbb{X}, d, \nu) \cap L^s(\mathbb{X}, \nu) \quad \int_{\mathbb{X}} fg \, d\nu = -\int_{\mathbb{X}} dg(X) \, d\nu \right\}.$$

Here, the right hand side makes sense as an action of an element of $L^p(T^*\mathbb{X})$ on an element of $L^q(T\mathbb{X})$; the resulting function is an element of $L^1(\mathbb{X}, \nu)$. The function f, which is unique by the density of $W^{1,p}(\mathbb{X}, d, \nu)$ in $L^p(\mathbb{X}, \nu)$, will be called the (q, r)-*divergence* of the vector field X, and we shall write $\operatorname{div}(X) = f$. An exhaustive discussion on the uniqueness of the divergence and its dependence on the exponents can be found in [8, 12].

In the course of the paper, we will extensively rely on the first order differential structure presented above. It is well-defined on metric spaces which are complete and separable; moreover, at least a priori, the structure is not defined locally - the objects $T^*\mathbb{X}$ and $T\mathbb{X}$ are not well-defined (the notation $L^p(T^*\mathbb{X})$ and $L^q(T\mathbb{X})$ is purely formal) and there is no immediate way to localise it to an open set $\Omega \subset \mathbb{X}$. However, whenever $\Omega \subset \mathbb{X}$ is an open bounded set, then $\overline{\Omega}$ is also a complete and separable metric space; hence, the whole first-order differential structure described above may be defined on $\overline{\Omega}$ as well. However, under the assumptions of Theorem 2.2, we may identify Newton-Sobolev and BV functions on Ω and $\overline{\Omega}$ (see [13]). Then, on $\overline{\Omega}$ the Newton-Sobolev space is equivalent to the Sobolev space defined by test-plans as in [11] and [8], so we may also define the differential structure on Ω if it is sufficiently regular; with a slight abuse of notation, we write $L^p(T^*\Omega)$ and $L^q(T\Omega)$, even though technically these objects are defined via an isometric extension to $\overline{\Omega}$.

However, under this identification, the divergence introduced above is not suitable for our purposes, because it takes into account the boundary effects. We need to use a notion of divergence which only sees the structure of X inside the open set Ω. To this end, we test the definition of the divergence using only functions which vanish at the boundary. Given an open bounded set $\Omega \subset \mathbb{X}$ which satisfies the assumptions of Theorem 2.2, for $\frac{1}{r} + \frac{1}{s} = 1$, we set

$$\mathcal{D}_0^{q,r}(\Omega) = \left\{ X \in L^q(T\Omega) : \ \exists f \in L^r(\Omega, \nu) \quad \forall g \in W_0^{1,p}(\Omega, d, \nu) \cap L^s(\Omega, \nu) \int_\Omega fg \, d\nu = -\int_\Omega dg(X) \, d\nu \right\},$$

where $W_0^{1,p}(\Omega, d, \nu)$ is the space of Sobolev functions in $W^{1,p}(\Omega, d, \nu)$ with zero trace. We again say that the (uniquely defined) function f is the divergence of X and we write $\operatorname{div}_0(X) = f$. The relationship between the two definitions of the divergence can be roughly described as follows: the divergence $\operatorname{div}(X)$ is the divergence $\operatorname{div}_0(X)$ plus a boundary term which has an interpretation of the normal trace. Finally, let us note that when the metric measure space is Euclidean equipped with the Lebesgue measure, the vector fields and differentials arising from this construction coincide with their standard counterparts defined in coordinates, see [11].

2.4. Anzellotti pairings and Gauss-Green formula

We now present the notion of Anzellotti pairings and a Gauss-Green formula for a bounded domain in a metric space, which will be our key tool in the identification of solutions; this is a generalisation of the classical results due to Anzellotti [5] in the Euclidean setting. In this subsection, we require that $\Omega \subset \mathbb{X}$ satisfies the assumptions of Theorem 2.2. Moreover, we require that Ω is a regular domain in the following sense: denote $\Omega_t = \{x \in \Omega : \operatorname{dist}(x, \Omega^c) \geq t\}$. An open set $\Omega \subset \mathbb{X}$ is a *regular domain* if it has finite perimeter and

$$|D\chi_\Omega|(\mathbb{X}) = \limsup_{t \to 0} \frac{\nu(\Omega \setminus \Omega_t)}{t}.$$

Suppose that $X \in L^\infty(T\Omega)$ and $u \in BV(\Omega, d, \nu)$. As in the case of classical Anzellotti pairings, we will additionally assume a joint regularity condition on u and X which makes the pairing well-defined. The condition is as follows: for $p \in [1, \infty)$, we have

$$\operatorname{div}_0(X) \in L^p(\Omega, \nu), \quad u \in BV(\Omega, d, \nu) \cap L^q(\Omega, \nu), \quad \frac{1}{p} + \frac{1}{q} = 1. \tag{2.2}$$

In other words, we assume that $X \in \mathcal{D}_0^{\infty,p}(\Omega)$ and $u \operatorname{div}_0(X)$ is integrable.

Definition 2.5 Suppose that the pair (X, u) satisfies condition (2.2). Then, given a Lipschitz function $f \in \mathrm{Lip}(\Omega)$ with compact support, we set

$$\langle (X, Du), f\rangle := -\int_\Omega u \,\mathrm{div}_0(fX)\, d\nu = -\int_\Omega u\, df(X)\, d\nu - \int_\Omega uf \mathrm{div}_0(X)\, d\nu.$$

It turns out that the functional (X, Du) can be represented by a Radon measure which is absolutely continuous with respect to $|Du|_\nu$. Moreover, for every Borel set $A \subset \Omega$ we have

$$\int_A |(X, Du)| \le \|X\|_\infty \int_A |Du|_\nu.$$

Moreover, the following Gauss-Green formula was proved in [13] (for an earlier result concerning Lipschitz functions see [8]). Given $X \in \mathcal{D}_0^{\infty,p}(\Omega)$, there exists a function $(X \cdot \nu_\Omega)^- \in L^\infty(\partial\Omega, |D\chi_\Omega|_\nu)$ such that for all $u \in BV(\Omega, d, \nu)$ with the property that the pair (X, u) satisfies condition (2.2) we have

$$\int_\Omega u \,\mathrm{div}_0(X)\, d\nu + \int_\Omega (X, Du) = -\int_{\partial\Omega} T_\Omega u\, (X \cdot \nu_\Omega)^-\, d|D_{\chi_\Omega}|_\nu.$$

Note that the sign on the right hand side is different from the usual Gauss-Green formula due to the fact that $(X \cdot \nu_\Omega)^-$ has an interpretation of the weak normal trace defined using the interior normal vector.

3. The total variation flow

We begin with the Neumann problem

$$\begin{cases} u_t(t,x) = \mathrm{div}\left(\frac{Du(t,x)}{|Du(t,x)|_\nu}\right) & \text{in } (0,T)\times\Omega; \\ \frac{\partial u}{\partial \eta} := \frac{Du}{|Du|_\nu}\cdot \eta = 0 & \text{on } (0,T)\times\partial\Omega; \\ u(0,x) = u_0(x) & \text{in } \Omega. \end{cases} \tag{3.1}$$

This Section is organised into two parts. In the first one, we consider the gradient flow of the total variation with Neumann boundary data and prove existence, uniqueness and characterisation of weak solutions for L^2 initial data. Then, we present the notion of entropy solutions to deal with L^1 initial data. From now on, we impose the following regularity assumptions on the domain: $\Omega \subset \mathbb{X}$ is an open bounded set; it is a regular domain; and both Ω and $\mathbb{X}\backslash\overline{\Omega}$ satisfy the assumptions of Theorem 2.2.

3.1. Weak solutions

Consider the energy functional $\mathcal{TV}_N : L^2(\Omega, \nu) \to [0, +\infty]$ defined by

$$\mathcal{TV}_N(u) := \begin{cases} |Du|_\nu(\Omega) & \text{if } u \in BV(\Omega, d, \nu) \cap L^2(\Omega, \nu); \\ \\ +\infty & \text{if } u \in L^2(\Omega, \nu) \setminus BV(\Omega, d, \nu). \end{cases}$$

It is clear that $\mathcal{TV}_N$ is convex and lower semi-continuous with respect to the $L^2(\Omega, \nu)$-convergence. Then, by the theory of maximal monotone operators (see [7]) there is a unique strong solution of the abstract Cauchy problem

$$\begin{cases} 0 \in u'(t) + \partial\mathcal{TV}_N(u(t)) & \text{for all } t \in [0,T]; \\ u(0) = u_0. \end{cases}$$

To characterize the subdifferential of $\mathcal{TV}_N$, we define the following operator.

Definition 3.1 $(u, v) \in \mathcal{A}_N$ if and only if $u, v \in L^2(\Omega, \nu)$, $u \in BV(\Omega, d, \nu)$ and there exists a vector field $X \in \mathcal{D}_0^{\infty,2}(\Omega)$ with $\|X\|_\infty \le 1$ such that the following conditions hold:

$$-\mathrm{div}_0(X) = v \quad \text{in } \Omega;$$

$$(X, Du) = |Du|_\nu \quad \text{as measures;}$$

$$(X \cdot \nu_\Omega)^- = 0 \qquad |D_{\chi_\Omega}|_\nu - \text{a.e. on } \partial\Omega.$$

The main result of this Section states that the operator $\mathcal{A}_\mathcal{N}$ coincides with the subdifferential of $\mathcal{TV}_\mathcal{N}$. To get this characterisation, we use the methods of convex duality, and in particular the Fenchel-Rockafellar duality theorem (see for instance [10]). It is possible to prove the characterisation using duality techniques, because the Gigli differential structure is linear and our assumptions on the domain guarantee that the trace operator is linear and bounded. The main reason to rely on duality theory is that the differential structure is (at least a priori) not defined locally and some of the Euclidean tools fail; for instance, it is not clear how to approximate a vector field with integrable divergence by more regular vector fields.

Theorem 3.2 *The set $D(\mathcal{A}_\mathcal{N})$ is dense in $L^2(\Omega,\nu)$ and*

$$\partial\mathcal{TV}_\mathcal{N} = \mathcal{A}_\mathcal{N}.$$

Our concept of solutions to the Neumann problem (3.1) is the following:

Definition 3.3 Given $u_0 \in L^2(\Omega,\nu)$, we say that u is a *weak solution* of the Neumann problem (3.1) in $[0,T]$, if $u \in C([0,T];L^2(\Omega,\nu)) \cap W^{1,2}_{\mathrm{loc}}(0,T;L^2(\Omega,\nu))$, $u(0,\cdot)=u_0$, and for almost all $t\in(0,T)$ we have $u(t)\in BV(\Omega,d,\nu)$ and there exist vector fields $X(t)\in\mathcal{D}^{\infty,2}_0(\Omega)$ with $\|X(t)\|_\infty\le 1$ such that the following conditions hold:

$$\mathrm{div}_0(X(t)) = u_t(t,\cdot) \quad \text{in } \Omega;$$

$$(X(t),Du(t)) = |Du(t)|_\nu \quad \text{as measures;}$$

$$(X(t)\cdot\nu_\Omega)^- = 0 \qquad |D_{\chi_\Omega}|_\nu - \text{a.e. on } \partial\Omega.$$

Then, using the classical theory of maximal monotone operators (see for instance [7]), as a consequence of Theorem 3.2 we have the following existence and uniqueness theorem.

Theorem 3.4 *For any $u_0\in L^2(\Omega,\nu)$ and all $T>0$, there exists a unique weak solution of the Neumann problem* (3.1) *in* $[0,T]$.

In a similar way, we can get existence and uniqueness of weak solutions to the Dirichlet problem for the total variation flow; we summarize the main results in the following Remark.

Remark 3.5 Consider the Dirichlet problem for the total variation flow

$$\begin{cases} u_t(t,x) = \mathrm{div}\left(\frac{Du(t,x)}{|Du(t,x)|_\nu}\right) & \text{in } (0,T)\times\Omega;\\ u(t,x) = f(x) & \text{on } (0,T)\times\partial\Omega;\\ u(0,x)=u_0(x) & \text{in } \Omega. \end{cases} \tag{3.2}$$

Given $u_0\in L^2(\Omega,\nu)$, we say that u is a *weak solution* of the Dirichlet problem (3.2) in $[0,T]$, if $u\in C([0,T];L^2(\Omega))\cap W^{1,2}_{\mathrm{loc}}(0,T;L^2(\Omega,\nu))$, $u(0,\cdot)=u_0$, and for almost all $t\in(0,T)$ there exist vector fields $X(t)\in\mathcal{D}^{\infty,2}_0(\Omega)$ with $\|X(t)\|_\infty\le 1$ such that the following conditions hold:

$$\mathrm{div}_0(X(t)) = u_t(t,\cdot) \quad \text{in } \Omega;$$

$$(X(t),Du(t)) = |Du(t)|_\nu \quad \text{as measures;}$$

$$(X(t)\cdot\nu_\Omega)^- \in \mathrm{sign}(T_\Omega u(t)-f) \qquad |D_{\chi_\Omega}|_\nu - \text{a.e. on } \partial\Omega.$$

Note that the sign of the normal trace of the vector field X is different than in the standard definition of weak solutions for the total variation flow in Euclidean spaces; this is due to the fact that $(X\cdot\nu_\Omega)^-$ corresponds to the choice of an interior unit normal and the standard Anzellotti normal trace is defined using the exterior unit normal. Then, for any $f\in L^1(\partial\Omega,\mathcal{H})$, $u_0\in L^2(\Omega,\nu)$ and $T>0$, there exists a unique weak solution of the Dirichlet problem (3.2) in $[0,T]$. Moreover, the following comparison principle holds: if u_1,u_2 are weak solutions for the initial data $u_{1,0},u_{2,0}\in L^2(\Omega,\nu)\cap L^q(\Omega,\nu)$ respectively, then

$$\|(u_1(t)-u_2(t))^+\|_q \le \|(u_{1,0}-u_{2,0})^+\|_q \quad \text{for all } 1\le q\le\infty.$$

3.2. Entropy solutions

We shift our attention to the Neumann problem for initial data in $L^1(\Omega,\nu)$. To this end, we make use of the notion of entropy solutions, which are defined using a family of inequalities involving truncations as in the Euclidean case; for $k>0$, denote by $T_k:\mathbb{R}\to\mathbb{R}$ the truncation function

$$T_k(r):=\begin{cases} k\,\operatorname{sign}(r) & \text{if } |r|>k;\\ r & \text{if } |r|\le k.\end{cases}$$

Definition 3.6 $(u,v)\in\mathcal{B}_{\mathcal{N}}$ if and only if $u,v\in L^1(\Omega,\nu)$, $T_k(u)\in BV(\Omega,d,\nu)$ for all $k>0$, and there exists a vector field $X\in\mathcal{D}_0^{\infty,1}(\Omega)$ with $\|X\|_\infty\le 1$ and $-\operatorname{div}_0(X)=v$ in Ω such that

$$\int_\Omega (w-T_k(u))\,v\,d\nu\le\int_\Omega (X,Dw)-\int_\Omega |DT_k(u)|_\nu$$

for all $w\in BV(\Omega,d,\nu)\cap L^\infty(\Omega,\nu)$ and $k>0$.

Proposition 3.7 *The following conditions are equivalent*:
(i) $(u,v)\in\mathcal{B}_{\mathcal{N}}$;
(ii) $u,v\in L^1(\Omega,\nu)$, $T_k(u)\in BV(\Omega,d,\nu)$ *for all* $k>0$, *and there exists a vector field* $X\in\mathcal{D}_0^{\infty,1}(\Omega)$ *with* $\|X\|_\infty\le 1$ *such that*

$$-div_0(X)=v \text{ in } \Omega;$$
$$\int_\Omega (X,DT_k(u))=\int_\Omega |DT_k(u)|_\nu \quad \text{for all } k>0;$$
$$(X\cdot\nu_\Omega)^-=0 \qquad |D_{\chi_\Omega}|_\nu-\text{a.e. on } \partial\Omega.$$

As a consequence, we have

$$\mathcal{B}_{\mathcal{N}}\cap(L^2(\Omega,\nu)\times L^2(\Omega,\nu))=\mathcal{A}_{\mathcal{N}}.$$

Theorem 3.8 *The operator* $\mathcal{B}_{\mathcal{N}}$ *is m-completely accretive in* $L^1(\Omega,\nu)$ *and homogeneous of degree zero. Moreover,* $D(\mathcal{B}_{\mathcal{N}})$ *is dense in* $L^1(\Omega,\nu)$.

In particular, the operator $\mathcal{A}_{\mathcal{N}}$ is also completely accretive as a restriction of $\mathcal{B}_{\mathcal{N}}$ to $L^2(\Omega,\nu)\times L^2(\Omega,\nu)$.

Definition 3.9 We say that $u\in C([0,T];L^1(\Omega,\nu))\cap W^{1,1}_{\mathrm{loc}}(0,T;L^1(\Omega,\nu))$ is an *entropy solution* of the Neumann problem (3.1) in $[0,T]$ with initial data $u_0\in L^1(\Omega,\nu)$, if for all $k>0$ we have $T_ku(t)\in BV(\Omega,d,\nu)$ and there exist vector fields $X(t)\in\mathcal{D}_0^{\infty,1}(\Omega)$ with $\|X(t)\|_\infty\le 1$ such that for a.e. $t\in[0,T]$ the following conditions hold:

$$\operatorname{div}_0(X(t))=u_t(t,\cdot) \quad \text{in } \Omega;$$
$$(X(t),DT_ku(t))=|DT_ku(t)|_\nu \quad \text{as measures};$$
$$(X(t)\cdot\nu_\Omega)^-=0 \qquad |D_{\chi_\Omega}|_\nu-\text{a.e. on } \partial\Omega.$$

Theorem 3.10 *For any* $u_0\in L^1(\Omega,\nu)$ *and all* $T>0$ *there is a unique entropy solution* $u(t)$ *of the Neumann problem* (3.1) *in* $[0,T]$. *Moreover, the following comparison principle holds*: *if* u_1,u_2 *are entropy solutions for the initial data* $u_{1,0},u_{2,0}\in L^q(\Omega,\nu)$, *respectively, then*

$$\|(u_1(t)-u_2(t))^+\|_q\le\|(u_{1,0}-u_{2,0})^+\|_q \quad \text{for all } 1\le q\le\infty. \tag{3.3}$$

The comparison principle given in equation (3.3) is a consequence of the complete accretivity of the operator $\mathcal{B}_{\mathcal{N}}$. Actually, this notion of solutions for L^1 initial data can also be applied not only for the Neumann problem, but also for the total variation flow defined on the whole space, considered for instance in [12]. This is formalised in the following Remark.

Remark 3.11 Suppose that $\nu(\mathbb{X})<\infty$. Consider the energy functional

$$\mathcal{TV}(u)=\begin{cases}|Du|_\nu & \text{if } u\in BV(\mathbb{X},d,\nu)\cap L^2(\mathbb{X},\nu);\\ +\infty & \text{if } u\in L^2(\mathbb{X},\nu)\setminus BV(\mathbb{X},d,\nu)\end{cases}$$

and its gradient flow in $L^2(\mathbb{X},\nu)$

$$\begin{cases} 0 \in u'(t) + \partial\mathcal{TV}(u(t)) & \text{for all } t \in [0,T]; \\ u(0) = u_0. \end{cases} \tag{3.4}$$

By theory of maximal monotone operators, there exists a unique solution of (3.4) for initial data $u_0 \in L^2(\mathbb{X},\nu)$ and its characterisation in terms of Anzellotti pairings was given in [12]. In a similar way as in this subsection, we may also introduce the notion of entropy solutions valid for initial data in $L^1(\mathbb{X},\nu)$; for the purpose of this Remark only, denote by (X, Du) the Anzellotti pairing introduced in [12], i.e. as in Definition 2.5, but with the divergence div in place of div_0.

With this understood, we say that $u \in C([0,T];L^1(\mathbb{X},\nu)) \cap W^{1,1}_{\mathrm{loc}}(0,T;L^1(\mathbb{X},\nu))$ is an *entropy solution* of the Cauchy problem (3.4) in $[0,T]$ with initial data $u_0 \in L^1(\mathbb{X},\nu)$, if for all $k > 0$ we have $T_k u(t) \in BV(\mathbb{X},d,\nu)$ and there exist vector fields $X(t) \in \mathcal{D}^{\infty,1}(\mathbb{X})$ with $\|X(t)\|_\infty \le 1$ such that for almost all $t \in [0,T]$ the following conditions hold:

$$\mathrm{div}(X(t)) = u_t(t,\cdot) \quad \text{in } \mathbb{X};$$
$$(X(t), DT_k u(t)) = |DT_k u(t)|_\nu \quad \text{as measures.}$$

For any $u_0 \in L^1(\mathbb{X},\nu)$ and $T > 0$, there exists a unique entropy solution of the Cauchy problem (3.4) in $[0,T]$, and it satisfies the comparison principle and estimates given in Theorem 3.10.

Acknowledgments. This research was funded partially by the Austrian Science Fund (FWF), grant ESP 88. The first author has also been partially supported by the OeAD-WTZ project CZ 01/2021. The second author has been partially supported by the Conselleria d'Innovació, Universitats, Ciència y Societat Digital, project AICO/2021/223. For the purpose of open access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

[1] L. Ambrosio and S. Di Marino, *Equivalent definition of BV spaces and total variation on metric measure spaces*. J. Funct. Anal. **266** (2014), 4150–4188.

[2] L. Ambrosio, N. Gigli, and G. Savaré, Gradient flows in metric spaces and in the space of probability measures, Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 2005.

[3] L. Ambrosio, N. Gigli and G. Savaré, *Density of Lipschitz function and equivalence of weak gradients in metric measure spaces*, Rev. Mat. Iberoam. **29** (2013), 969–996.

[4] F. Andreu, V. Caselles, and J.M. Mazón, Parabolic Quasilinear Equations Minimizing Linear Growth Functionals, Progress in Mathematics, vol. 223, Birkhäuser, 2004.

[5] G. Anzellotti, *Pairings between measures and bounded functions and compensated compactness*, Ann. Mat. Pura Appl. (4) **135** (1983), 293–318.

[6] A. Björn and J. Björn, *Nonlinear Potential Theory on Metric Spaces*, EMS Tracts in Mathematics, vol. 17, European Mathematical Society, Zürich, 2011.

[7] H. Brezis, Operateurs Maximaux Monotones. North Holland, Amsterdam, 1973.

[8] V. Buffa, G.E. Comi and M. Miranda Jr., *On BV functions and essentially bounded divergence-measure fields in metric spaces*, Rev. Mat. Iberoam. **38** (2022), no. 3, 883–946.

[9] M. G. Crandall and T. M. Liggett, *Generation of Semigroups of Nonlinear Transformations on General Banach Spaces*, Amer. J. Math. **93** (1971), 265–298.

[10] I. Ekeland, R. Temam, *Convex analysis and variational problems*, North-Holland Publ. Company, Amsterdam, 1976.

[11] N. Gigli, *Nonsmooth differential geometry - an approach tailored for spaces with Ricci curvature bounded from below*, Mem. Amer. Math. Soc. **251** (2018), no. 1196, v+161 pp.

[12] W. Górny and J.M. Mazón, *On the p-Laplacian evolution equation in metric measure spaces*, J. Funct. Anal. **283** (2022), 109621.

[13] W. Górny and J.M. Mazón, *The Anzellotti-Gauss-Green formula and least gradient functions in metric measure spaces*, preprint (2021), available at arXiv:2105.00432.

[14] W. Górny and J.M. Mazón, *The Neumann and Dirichlet problems for the total variation flow in metric measure spaces*, Adv. Calc. Var. (2022), ahead of print, doi.org/10.1515/acv-2021-0107.

[15] P. Lahti and N. Shanmugalingam, *Trace theorems for functions of bounded variation in metric spaces*, J. Funct. Anal. **274** (2018), no. 10, 2754–2791.

[16] M., Miranda Jr., *Functions of bounded variation on "good" metric spaces*. J. Math. Pures Appl. **82** (2003), 975–1004.

[17] L. Maly, N. Shanmugalingam, M. Snipes, *Trace and extension theorems for functions of bounded variation*, Ann. Scuola Norm-Sci. **18** (1) (2018), 313–341.

[18] L. Rudin, S. Osher and E. Fatemi, *Nonlinear Total Variation based Noise Removal Algorithms*. Physica D. **60** (1992), 259–268.

Singularly perturbed reaction-diffusion problems with non-smooth initial and/or boundary data

José Luis Gracia[1], **Eugene O'Riordan**[2]
1. *IUMA and Department of Applied Mathematics. Universidad de Zaragoza, Spain*
2. *School of Mathematical Sciences, Dublin City University, Ireland*

Abstract

This paper reviews a series of papers by the authors dealing with different approaches to generating a parameter-uniform global approximation to the solution of singularly perturbed reaction-diffusion problems with non-smooth data. In addition to the usual layers that appear in the solution due to the presence of the singular perturbation parameter, additional layers can appear when the data for the problem are not sufficiently smooth. All of the approaches involve finite difference operators on Shishkin meshes, coupled with some additional feature in the algorithm to deal with the new effects generated by the lack of regularity in the continuous solution.

1. Introduction

Singularly perturbed problems are often used as mathematical models describing physical processes in many areas of applied sciences. They are characterized by the presence of a small positive parameter multiplying the highest derivatives of the differential equation, causing their solutions to typically have large derivatives in narrow subregions of the domain, called layers. Due to the presence of the layers, classical numerical methods are not appropriate and parameter-uniform numerical methods are required [2, 18]. The convergence properties of a parameter-uniform numerical method are not adversely affected by the presence of singular perturbation parameters in the differential equation. Throughout the paper, $0 < \varepsilon \leq 1$ denotes a singular perturbation parameter, which can take arbitrary small values.

In the literature two approaches have been used to design a parameter-uniform numerical method: fitted operator [16] and fitted mesh methods [2]. Fitted mesh methods have been widely used over the last few decades and the most common incorporate either Bakhvalov [1] or Shishkin meshes [21]. The latter meshes, are piecewise uniform and are fine in the layer regions. The mesh is constructed using *a priori* information of the singularly perturbed nature of the solution. In this respect, it is important to note that it is not possible to construct a fitted operator method on a uniform mesh for a class of linear parabolic reaction-diffusion problems [19] that includes the singularly perturbed heat equation $-\varepsilon u_{xx} + u_t = f(x,t)$ for $(x,t) \in (0,1) \times (0,1]$.

In this paper we shall only consider finite difference schemes defined on uniform and Shishkin meshes. These schemes will give an approximation to the solution of the problem at the mesh nodes. A global approximation in the whole domain can be generated from the nodal values using an interpolation operator. Global approximations are desirable when approximating the solution of singularly perturbed problems due to the multiscale character of the solution. We cannot fail to point out that for some classes of singularly perturbed problems, there are fitted operator methods on uniform meshes that are nodally but not globally parameter-uniformly convergent [2].

In general, singularly perturbed problems with smooth data have boundary layers. If either the coefficients of the differential equation are discontinuous or the source term has a point source [17, 20], then the solution can also exhibit interior layers. The analysis of the asymptotic behaviour of the solution provides information about the location and width of the layers, and this information is crucial in designing layer-adapted *a priori meshes*, such as, for example, the piecewise-uniform Shishkin meshes.

Most papers assume that all problem data (including the initial and boundary conditions) are smooth in the error analysis to establish parameter-uniform error bounds on the numerical approximations. In addition, when singularly perturbed parabolic problems posed on a domain $\bar{Q}$ are considered, it is also usually assumed that second level compatibility conditions between the initial and boundary conditions are satisfied in order to guarantee that the solution is in $C^{4+\gamma}(\bar{Q})$ (which denotes the space of all functions whose spatial derivatives up to fourth order and time derivatives up to second order are Hölder continuous of degree γ.)

The aim of this paper is to review the main results and conclusions of our research for singularly perturbed linear parabolic initial-boundary-value problems with non-smooth data. A collection of these results are now presented in this single paper and our interest focusses on the cases that either the initial and boundary conditions do not satisfy zero level compatibility conditions or the initial or boundary conditions are discontinuous. The difficulties for approximating these three problem classes are well-known as standard fitted mesh methods are not parameter uniformly convergent [12–14] for problems with non-smooth data. In the case of reaction-diffusion problems, one deals with a bi-singular problem where a classical singularity is entwined with the singular nature

of the differential operator when $\varepsilon \ll 1$. The scenario is much more complex if a convective term (i.e., $a(x,t)u_x$) is present in the differential equation. Then, the singularity is transported along a characteristic of the reduced problem (which is formally obtained by setting $\varepsilon = 0$ in the differential equation) causing an interior layer, whose location moves with time. When the final time is large enough, this interior layer interacts with a boundary layer causing serious difficulties from the theoretical and numerical point of view. We shall confine our discussion to singularly perturbed problems of reaction-diffusion type although some comments are made in the case of convection-diffusion problems.

Three approaches are considered in this paper for these singularly perturbed problems with non-smooth data and the technical details can be found in [5,6,11,12,14,22]. In § 2 singularly perturbed problems of reaction-diffusion type are regularized by replacing the discontinuous data with a smooth function and the regularized problem is approximated with a fitted mesh method [5]. A second alternative approach for this problem class is considered in § 3; from the values computed with a nodally (but not globally) fitted mesh method defined on a uniform mesh [14], a global approximation to the solution is obtained by using a special interpolation operator which is exact for the error function [6]. Finally, in § 4 a numerical/analytical approach [11, 12] is considered. First, the main singular component associated with the discontinuity is identified and separated off from the solution u, then the remainder, which is smoother than u, is approximated with a standard finite difference scheme on a Shishkin mesh.

Notation: Throughout, C denotes a generic positive constant that is independent of the singular perturbation parameter ε and all discretization parameters. The L_∞ norm on the domain D will be denoted by $\|\cdot\|_D$.

2. First Approach: Regularization of the data

Consider the following problem: Find $u(x,t)$ such that

$$Lu := u_t - \varepsilon u_{xx} + b(x,t)u = f(x,t), \quad (x,t) \in Q := (0,1)\times(0,T], \tag{2.1a}$$

$$u(0,t) = \phi_L(t),\ u(1,t) = \phi_R(t), \quad 0 < t \le T, \quad u(x,0) = \phi(x), \quad 0 \le x \le 1, \tag{2.1b}$$

where f and b are smooth functions. Without loss of generality, assume that $b(x,t) \ge \beta > 0$ for $(x,t) \in \bar{Q}$, otherwise consider the transformation $v(x,t) = u(x,t)e^{-\alpha_0 t}$ with $\alpha_0 > 0$ a constant sufficiently large.

If the initial condition ϕ and the boundary conditions are smooth functions and we assume that they satisfy second order compatibility conditions at the point $(0,0)$

$$\phi_L(0) = \phi(0), \quad \phi_L'(0) - \varepsilon\phi''(0) + b(0,0)\phi(0,0) = f(0,0), \tag{2.2a}$$

$$(f - L\phi)_t(0,0) - \varepsilon(f - L\phi)_{xx}(0,0) = 0, \tag{2.2b}$$

and similar conditions are satisfied at $(1,0)$, then $u \in C^{4+\gamma}(\bar{Q})$. In order to analyze the uniform convergence of a numerical scheme, the solution is decomposed into a regular component and boundary layer components [15]

$$u = v + w_L + w_R, \tag{2.3}$$

with $u, v, w \in C^{4+\gamma}(\bar{Q})$ and they satisfy for $0 \le k + 2m \le 4$

$$\left|\frac{\partial^{k+m}}{\partial x^k \partial t^m} v(x,t)\right| \le C(1 + \varepsilon^{1-k/2}), \quad (x,t) \in \bar{Q}, \tag{2.4a}$$

$$\left|\frac{\partial^{k+m}}{\partial x^k \partial t^m} w_L(x,t)\right| \le C\varepsilon^{-k/2} e^{-x\sqrt{\beta}/\sqrt{\varepsilon}}, \quad (x,t) \in \bar{Q}, \tag{2.4b}$$

$$\left|\frac{\partial^{k+m}}{\partial x^k \partial t^m} w_R(x,t)\right| \le C\varepsilon^{-k/2} e^{-(1-x)\sqrt{\beta}/\sqrt{\varepsilon}}, \quad (x,t) \in \bar{Q}, \tag{2.4c}$$

showing the presence of two boundary layers near $x = 0$ and $x = 1$ with a width of order $O(\sqrt{\varepsilon})$. Let N and M be two positive integers. Use backward Euler and standard central differences

$$L^{N,M}u_i^j := D_t^- u_i^j - \varepsilon\delta_x^2 u_i^j + b(x_i,t_j)u_i^j = f(x_i,t_j), \quad (x_i,t_j) \in Q^{N,M} \tag{2.5}$$

defined on the mesh $\bar{Q}^{N,M} = \{(x_i,t_j),\ 0 \le i \le N,\ 0 \le j \le M\}$ to generate an approximation to the solution of problem (2.1). The discrete operators D_t and δ_x^2 are defined by

$$D_t^- u_i^j := \frac{u_i^j - u_i^{j-1}}{t_j - t_{j-1}}, \quad \delta_x^2 u_i^j := \frac{2}{h_i + h_{i+1}}\left(\frac{u_{i+1}^j - u_i^j}{h_{i+1}} - \frac{u_i^j - u_{i-1}^j}{h_i}\right)$$

with $h_i := x_i - x_{i-1}$. If one considers a uniform mesh for the temporal variable and a piecewise uniform Shishkin mesh for the spatial variable

$$[0,\sigma] \cup [\sigma, 1-\sigma] \cup [1-\sigma, 1], \quad \sigma := \min\left\{\frac{1}{4}, m_0 \frac{\sqrt{\varepsilon}}{\sqrt{\beta}} \ln N\right\}, \tag{2.6}$$

where m_0 is an arbitrary positive constant and the N grid points are distributed in the ratio $N/4 : N/2 : N/4$, then the following error estimate is satisfied [15]

$$|u(x_i, t_j) - u_i^j| \leq C(N^{-1} \ln N)^2 + M^{-1}).$$

If the initial and/or boundary data are discontinuous, the previous error analysis is no longer valid. Even further, this method does not produce parameter-uniform numerical approximations [13, 14]. For example, consider problem (2.1) with

$$\phi(x) = 1 \quad \phi_L(t) = 0,$$

and sufficient compatibility conditions are satisfied at the point $(1, 0)$. Observe that $\phi(0) \neq \phi_L(0)$.

In order to obtain an accurate approximation to the solution of problem (2.1), the first approach is based on replacing the initial condition with a smooth function so that the solution of the new problem $u_{reg} \in C^{4+\gamma}(\bar{Q})$. Consider the following initial condition for the regularized problem

$$\phi(x;\varepsilon) = \left(1 - e^{\frac{-1}{\sqrt{\varepsilon}}}\right)^{-p} \left(1 - e^{\frac{-x}{\sqrt{\varepsilon}}}\right)^{p}, \; p \geq 4, \quad x \in (0,1).$$

Observe that this initial condition tends to a discontinuous initial condition as the singular perturbation parameter tends to zero. The solution of the regularized problem u_{reg} is decomposed into a regular v and singular w, z components, which satisfy for $0 \leq k + 2m \leq 4$ and $(x,t) \in \bar{Q}$

$$\left|\frac{\partial^{k+m}}{\partial x^k \partial t^m} v(x,t)\right| \leq C(1 + \varepsilon^{1-k/2}), \; \left|\frac{\partial^{k+m}}{\partial x^k \partial t^m} w(x,t)\right| \leq C\varepsilon^{-k/2} e^{-(1-x)\sqrt{\beta}/\sqrt{\varepsilon}}, \; \left|\frac{\partial^{k+m}}{\partial x^k \partial t^m} z(x,t)\right| \leq C\varepsilon^{-k/2} e^{-x\sqrt{\beta}/\sqrt{\varepsilon}}.$$

If the regularized problem is approximated with the standard scheme (2.5) on the Shishkin mesh (2.6), one can prove the following error estimates [5, Theorem 2] for $(x_i, t_j) \in \bar{Q}^{N,M}$

$$|u_{reg}(x_i, t_j) - (u_{reg})_i^j| \leq C((N^{-1} \ln N)^2 + N^{-1}\varepsilon^{1/2} + M^{-1}). \tag{2.7}$$

In [5] it is illustrated that the approximations $(u_{reg})_i^j$ are only accurate approximations to the solution u outside an ε dependent neighbourhood of the point $(0,0)$. In other words, this approach will not generate parameter-uniform global approximations to the solution u of the original problem.

Problem (2.1) with a discontinuous initial condition at the point $(d, 0)$ with $d \geq C\sqrt{\varepsilon}$ is also considered in [5]. This problem exhibits an interior layer along $x = d$ of width $O(\sqrt{\varepsilon})$ and an appropriate regularized problem is constructed to be approximated with a numerical method. The solution of the regularized problem is also decomposed into several components showing the asymptotic behaviour of the solution, but the regular component and the singular component associated with the interior layer of this problem are discontinuous functions along $x = d$. The Shishkin mesh condenses near the boundary and interior layers

$$[0, d-\sigma_1] \cup [d-\sigma_1, d+\sigma_2] \cup [d+\sigma_2, 1-\sigma_2] \cup [1-\sigma_2, 1], \tag{2.8}$$

where

$$\sigma_1 = \min\left\{\frac{d}{2}, 2\sqrt{\frac{\varepsilon}{\beta}} \ln N\right\}, \quad \sigma_2 = \min\left\{\frac{1-d}{4}, 2\sqrt{\frac{\varepsilon}{\beta}} \ln N\right\}.$$

The N grid points are distributed in the ratio $N/8 : N/4 : N/4 : N/4 : N/8$. If this problem is approximated with the standard scheme (2.5) on the Shishkin mesh (2.8), then one derives similar estimates to (2.7).

This approach has been also used in [3, 4, 7] to approximate problems of convection-diffusion with a discontinuous initial condition at the point $(d, 0)$ where d is either independent of ε or $d = O(\varepsilon^p)$ with $p < 1/2$. In this case, the classical singularity travels along the characteristic of the reduced problem emanating from the point $(d, 0)$. This causes an interior layer with a width of order $O(\sqrt{\varepsilon})$. In order to approximate the solution and prove the uniform convergence of a numerical method, the problem is transformed by using the mapping

$$X(x,t) := \begin{cases} \dfrac{d}{d(t)} x, & \text{if } x \leq d(t), \\ 1 - \dfrac{1-d}{1-d(t)}(1-x), & \text{if } x \geq d(t), \end{cases} \tag{2.9}$$

where $d(t)$ is the solution of the initial value problem

$$d'(t) = a(d(t), t),\ t > 0,\ d(0) = d. \tag{2.10}$$

The transformed domain consist of two rectangular subdomains $\bar{Q}^- := [0, d] \times [0, T]$ and $\bar{Q}^+ := [d, 1] \times [0, T]$. The solution is decomposed into a regular v, boundary w and interior z layer components, where w is a smooth function and v and z are discontinuous functions, but sufficiently smooth in each subdomain $\bar{Q}^-$ and $\bar{Q}^+$. Backward Euler method and upwinding approximation on a rectangular Shishkin mesh in the transformed domain—which is aligned with the interior layer in the original coordinate system—are used for the numerical approximation of this problem and it is proved that it is a parameter uniformly convergent scheme.

Problems with a delta-function present in the initial condition have also been examined. The initial condition chosen for the regularized problem is given by

$$\phi(x; \varepsilon) = g_1(x) + g_2(x) e^{-\theta \frac{(x-d)^2}{\varepsilon}}, \quad d = O(1);$$

where g_1, g_2 are two smooth functions. Observe that the initial condition involves a Gaussian profile with a standard deviation determined by two parameters ε and θ. Problems of reaction-diffusion and convection-diffusion with this type of initial condition are considered in [9] and [8], respectively. In both problems the asymptotic behaviour of the solution is analysed by using an appropriate decomposition of the solution. The components of the solution in the case of the reaction-diffusion problem are smooth unlike for the convection-diffusion problem. In the case of a reaction-diffusion problem [9], it is examined in detail the effect of the scale width of the initial layer on the solution if it is either thinner ($\theta \geq 1$) or wider ($\theta < 1$) than the scale induced by the differential equation. It is proved that the solution has, in addition of the typical boundary layers, an interior layer of width $O(\sqrt{\varepsilon/\theta})$. This information is used to design the Shishkin mesh for the finite difference scheme and its uniform convergence is proved. In addition, the error estimates reveal the effect of the width of the initial layer on the rate of convergence of the scheme. If a convective term ($a u_x$) is present in the differential equation, then the pulse travels along the characteristic of the reduced problem generating an interior layer. In [8] this problem is approximated when $\theta = O(1)$ by using the mapping (2.9) to align the mesh with the trajectory of the interior layer. Error estimates are given for a numerical scheme which combines backward Euler method and upwinding approximation on a Shishkin mesh, proving that the method is almost first-order uniformly convergent, due to the presence of a logarithmic factor.

3. Second Approach: Fitted operator method: Nodal and global convergence

We present an alternative approach (proposed in [13, 14]) to dealing with singularly perturbed problems with non-smooth data. Consider problem (2.1) where the reaction term $b = b(t)$ depends only on the time variable and the initial condition is discontinuous at $x = d$ with $d = O(1)$. Similar results can be deduced if the singularity is caused by an incompatibility in the problem data. Suppose the later situation and it is incompatible at the point $(0, 0)$. Estimates of the solution are deduced via a decomposition of the solution $u = v + z$, where the regular v and singular z components satisfy

$$\left| \frac{\partial^{k+m}}{\partial x^k \partial t^m} v(x,t) \right| \leq C \left(1 + \varepsilon^{1-k} t^{1/2-(m+k/2)} e^{-\frac{x}{2\varepsilon\sqrt{t}}} \right), \tag{3.1a}$$

$$\left| \frac{\partial^{k+m}}{\partial x^k \partial t^m} z(x,t) \right| \leq C \left(1 + \varepsilon^{-k} t^{-(m+k/2)} e^{-\frac{x}{2\varepsilon\sqrt{t}}} \right). \tag{3.1b}$$

These bounds reveal that the derivatives of the solution are unbounded in the neighbourhood of the point $(0, 0)$ and in the neighbourhood of the generated interior layer. The solution can also have boundary layers, but they are avoided here.

Problem (2.1) with an incompatibility at $(0, 0)$ is now approximated with the fitted scheme proposed in [14]. This scheme is defined on a uniform mesh and the discrete operator is given by

$$-\varepsilon \kappa(x,t) \delta_x^2 u_i^j + b(t_j) U + D_t^- u_i^j = f(x_i, t_j), \quad (x_i, t_j) \in Q^{N,M}, \tag{3.2}$$

where the fitting coefficient κ is defined by

$$\kappa(x,t) := \frac{D_t^-(\omega(x,t) + u_0(x,t)) + b(t)(\omega(x,t) + u_0(x,t))}{\varepsilon \delta_x^2 (\omega(x,t) + u_0(x,t))}, \quad (x,t) \in Q^{N,M}, \tag{3.3}$$

where $u_0(x,t) = -x^3 - 6\varepsilon x t$ and

$$\omega(x,t) := \frac{1}{2} e^{-\int_{r=0}^{t} b(r)\, dr} \operatorname{erf}\left(\frac{x}{2\sqrt{\varepsilon t}} \right),\ \operatorname{erf}(\zeta) := \frac{2}{\sqrt{\pi}} \int_{r=0}^{\zeta} \exp(-r^2)\, dr. \tag{3.4}$$

For $t = 0$, the error function is defined by continuous extension at $x = 0$. The fitting coefficient κ is chosen such that the scheme $D_t^- \omega - \varepsilon\kappa\delta_x^2\omega + b(t)\omega = 0$, i.e., this scheme is exact for the function $\omega(x,t)$.

Under certain conditions, including $b = b(t)$, the fitted scheme satisfies the following error estimates [14]

$$|u(x_i,t_j) - u_i^j| \le C(N^{-\nu} + M^{-\nu}),$$

with $\nu \in (0, 1/3)$, which are more pessimistic than the computed orders of convergence observed in the numerical experiments. These estimates show that this method converges nodally to the solution of problem (2.1), nevertheless one cannot generate a global approximation in the whole domain using these nodal values and bilinear interpolation

$$\bar{U}(x,t) := \sum_{i=0,j=1}^{N,M} u_i^j \varphi_i(x)\eta_j(t),$$

where $\varphi_i(x)$ is the standard hat function centered at $x = x_i$ and $\eta_j(t) = M(t - t_{j-1})$, $t \in [t_{j-1}, t_j)$. The lack of global convergence of the fitted scheme is not caused by the accuracy of the nodal approximations. To support this observation, we compare the solution u with its bilinear interpolant $\bar{u}$ generated on the Shishkin mesh (2.6). Using the estimates (3.1), we obtain

$$\|u - \bar{u}\|_{R_{i,j}} \le \begin{cases} CN^{-1} + CM^{-1}N^{-1}, & \text{if } x_i \ge \sigma, \\ C(N^{-1}\ln N)^2\delta^{-1} + CM^{-2}\delta^{-2}, & \text{if } t_j \ge \delta \text{ and } x_i < \sigma, \end{cases}$$

where $R_{i,j} := (x_i, x_{i+1}) \times (t_j, t_{j+1})$. These estimates show that the fitted piecewise-uniform Shishkin mesh and bilinear interpolation only give global parameter-uniform convergence accuracy outside of the corner layer region $(x,t) \in (0, \sqrt{\varepsilon}\ln N) \times (0, \delta), \delta > 0$, although we have accurate approximations of the solution at all mesh points, even if they are very close to the corner $(0,0)$. In [6] it is proposed a nonlinear interpolant operator which is exact for the constant function 1 and the error function in order to obtain a global approximation to the solution in the whole domain. This interpolant operator in the cell $R_{i,j}$ is given by

$$\bar{U}(x,t) := \sum_{l,m=0}^{1} u_{i+1}^{j+m} T(t; x_{i+l}, t_{j+m}) S(x,t; x_{i+l}), \tag{3.5}$$

where

$$T(t; x_{i+l}, t_{j+m}) := \frac{\omega(x_{i+l},t) - \omega(x_{i+l}, t_{j+1-m})}{\omega(x_{i+l},t_{j+m}) - \omega(x_{i+l}, t_{j+1-m})}, \quad S(x,t; x_{i+l}) := \frac{\omega(x,t) - \omega(x_{i+1-l},t)}{\omega(x_{i+l},t) - \omega(x_{i+1-l},t)}.$$

We have observed global approximations to the solution of all the test problems considered with non-smooth data when the fitted scheme (3.2) on the Shishkin mesh (2.6) and the nonlinear interpolation operator (3.5) are used. Nevertheless, the proof of the parameter uniform global convergence of this numerical method is an open question.

4. Third Approach: Analytical/numerical approach

In the third approach the main singular component $s(x,t)$ associated with the singularity is identified. This singular function involves the error function. The other term

$$y = u - s \quad \text{where} \quad Ly = f - Ls, \tag{4.1}$$

and the initial/boundary data associated with the function y are continuous functions. This function y is approximated using a numerical method.

Nevertheless, all the difficulties have not disappeared as the right-hand side $f - Ls$ is, in general, a non-smooth function. In addition, it is only satisfied the zero order compatibility condition or the initial/boundary data are piecewise smooth functions, and this can cause a small reduction in the order of convergence of the numerical methods [22]. This reduction is not only a theoretical issue, but it is observed in the computed orders of convergence. In [12] three problem classes are analysed and they are described below and for the sake of simplicity, we assume that $b = b(t)$, if not otherwise indicated.

Consider first the case that the initial and boundary conditions are incompatible at $(0,0)$, i.e., $\phi(0) \ne \phi_L(0)$. For this problem class, the function s is given by

$$s(x,t) = \phi(0^+)e^{-b(0)t}\,\mathrm{erf}\left(\frac{x}{2\sqrt{\varepsilon t}}\right).$$

The function y is decomposed into $y = v + w_L + w_R$, where the regular v and right boundary w_R components are in $C^{4+\gamma}(\bar{Q})$ and they satisfy similar estimates to (2.4). The component $w_L \in C^{2+\gamma}(\bar{Q})$ and it satisfies

$$\left|\frac{\partial^{i+j}}{\partial x^i \partial t^j} w_L(x,t)\right| \le C\varepsilon^{-(i/2)} e^{-\frac{\mu}{2}\frac{x}{\sqrt{T\varepsilon}}}, \quad 0 \le i+2j \le 2; \quad (x,t) \in \bar{Q},$$
$$\left|\frac{\partial^{i}}{\partial x^i} w_L(x,t)\right| \le \frac{C}{\varepsilon(\sqrt{\varepsilon t})^{i-2}} e^{-\frac{\mu}{2}\frac{x}{\sqrt{T\varepsilon}}}, \quad i = 3,4; , \quad (x,t) \in Q;$$
$$\left|\frac{\partial^{2}}{\partial t^2} w_L(x,t)\right| \le \frac{C}{t}, \quad (x,t) \in Q,$$

with $0 < \mu < 1$. Thus, the solution has two layers near $x = 0$ and $x = 1$ of width $O(\sqrt{\varepsilon})$ and the standard scheme (2.5) on the Shishkin mesh (2.6) is used to approximate the function y. Moreover, the higher derivatives of the boundary layer components are singular initially, which means that the theoretical error analysis requires a nonstandard approach.

Secondly, consider the case that the initial condition is discontinuous at $(d, 0)$ with $d = O(1)$. The singular component s is now given by

$$s(x,t) = \frac{[\phi](d)}{2} e^{-b(0)t} \operatorname{erf}\left(\frac{x-d}{2\sqrt{\varepsilon t}}\right),$$

where $[\phi](d) := \phi(d^+) - \phi(d^-)$. The function y is further decomposed into $y = v + w_L + w_R + w_I$, where the left w_L and right w_R boundary components are continuous functions and the regular v and interior layer w_I components are discontinuous functions, but smooth enough in each subdomain $\bar{Q}^- := [0,d] \times [0,T]$ and $\bar{Q}^+ := [d,1] \times [0,T]$. The interior layer component satisfies

$$\left|\frac{\partial^{i+j}}{\partial x^i \partial t^j} w_I(x,t)\right| \le C\varepsilon^{-(i/2)} e^{-\frac{\mu|x-d|}{2\sqrt{\varepsilon T}}}, \ 0 \le i+2j \le, \ \mu < 1, \ (x,t) \in \bar{Q}^- \cup \bar{Q}^+,$$
$$\left|\frac{\partial^{i}}{\partial x^i} w_I(x,t)\right| \le \frac{C}{\varepsilon(\sqrt{\varepsilon t})^{i-2}} e^{-\frac{\mu}{2}\frac{|x-d|}{\sqrt{T\varepsilon}}}, \quad i = 3,4, \quad (x,t) \in Q^- \cup Q^+;$$
$$\left|\frac{\partial^{2}}{\partial t^2} w_I(x,t)\right| \le \frac{C}{t}, \quad (x,t) \in Q^- \cup Q^+.$$

In addition to the boundary layers, the solution has an interior layer in the vicinity of $x = d$ of width $O(\sqrt{\varepsilon})$. The standard scheme (2.5) is now defined on the Shishkin mesh (2.8).

We consider now the third problem class where the boundary condition ϕ_L is discontinuous at $x = d$ with $d = O(1)$. The singular function for this problem is

$$s(x,t) = [\phi](d) H(t-d) \left(1 - e^{-b(0)(t-d)} \operatorname{erf}\left(\frac{x}{2\sqrt{\varepsilon(t-d)}}\right)\right), \quad \text{where} \quad H(x) := \begin{cases} 0, & \text{for } x < 0, \\ 1, & \text{for } x \ge 0. \end{cases}$$

The function y is decomposed into $y = v + w_R + w_L$, with $v, w_R \in C^{4+\gamma}(\bar{Q})$ and $w_L \in C^{2+\gamma}(\bar{Q})$ and the character of the function y for this problem class is the same as for the first problem class. Then, the standard scheme (2.5) is defined on the Shishkin mesh (2.6).

In [12] it is proved that the numerical schemes proposed for the three problem classes are uniformly and globally convergent, and they have second order in space and first order in time, except for a logarithmic factor.

This approach has been also applied to other problem classes observing global convergence in all of them. Among other problems, we have considered in [11] the following problem class

$$Lu := \varepsilon(u_t - u_{xx}) + b(x,t)u = f(x,t), \quad (x,t) \in Q, \tag{4.2a}$$
$$u(0,t) = \phi_L(t), \ u(1,t) = \phi_R(t), \quad 0 < t \le T, \quad u(x,0) = \phi(x), \quad 0 \le x \le 1, \tag{4.2b}$$

with $b(x,t) \ge \beta > 0$ and f, b, ϕ, ϕ_L and ϕ_R are smooth functions, but $\phi(0) \ne \phi_L(0)$. The solution has boundary layers in the vicinity of $x = 0$ and $x = 1$, a classical discontinuity at $(0,0)$ caused by the discontinuous data and an initial layer in the vicinity of $t = 0$, which is generated from the fact that the coefficient of the time derivative is ε in this problem class. For this problem, the function y in (4.1) is defined by

$$s(x,t) = (\phi_L(0) - \phi(0)) e^{-b(0,0)t/\varepsilon} \operatorname{erfc}\left(\frac{x}{2\sqrt{t}}\right),$$

where erfc$(\cdot)$ is the complementary error function. The component y solution is decomposed into $y = v+w_L+w_R+w_I+w_{IB}$; where v, w_L and w_R are the regular and boundary layers components which satisfy similar estimates to (2.4). The component w_I is associated to the initial layer and it satisfies the estimates

$$\left|\frac{\partial^{i+j}}{\partial x^i \partial t^j} w_I(x,t)\right| \leq C\varepsilon^{-i/2}\varepsilon^{-j} e^{-\frac{\beta t}{\varepsilon}},\ 0 \leq i+2j \leq 4; \tag{4.3}$$

and for the initial-boundary layer component w_{IB} we have

$$\begin{aligned}
\left|\frac{\partial^2 w_{IB}}{\partial x^2}(x,t)\right| + \left|\frac{\partial w_{IB}}{\partial t}(x,t)\right| &\leq C + \frac{C}{\varepsilon}\left(e^{-\sqrt{\frac{\beta}{\varepsilon}}x} + e^{-\sqrt{\frac{\beta}{\varepsilon}}(1-x)} + e^{-\frac{\beta t}{\varepsilon}}\right),\\
\left|\frac{\partial^4 w_{IB}}{\partial x^4}(x,t)\right| + \left|\frac{\partial^2 w_{IB}}{\partial t^2}(x,t)\right| &\leq \frac{C}{\varepsilon} + C\frac{\varepsilon}{t}e^{-\frac{\beta t}{\varepsilon}}\\
&\quad + \frac{C}{\varepsilon^2}\left(e^{-\sqrt{\frac{\beta}{\varepsilon}}x} + e^{-\sqrt{\frac{\beta}{\varepsilon}}(1-x)} + e^{-\frac{\beta t}{\varepsilon}}\right).
\end{aligned}$$

We have that $v, w_L, w_R, w_I \in C^{4+\gamma}(\bar{Q})$ but $w_{IB} \in C^{2+\gamma}(\bar{Q})$. This information is used to define the numerical scheme to approximate the function y and in the error analysis of the scheme. Consider the discrete operator

$$L^{N,M}u_i^j := \varepsilon(D_t^- u_i^j - \delta_x^2 u_i^j) + b(x_i,t_j)u_i^j = f(x_i,t_j), \quad (x_i,t_j) \in Q^{N,M}, \tag{4.4}$$

on the Shishkin mesh (2.6) for the spatial variable. Estimates (4.3) show that there is a layer near $t = 0$ and we have also considered a piecewise mesh for the temporal variable

$$[0,\tau] \cup [\tau,T], \quad \text{where} \quad \tau := \min\left\{\frac{1}{2}, \frac{\varepsilon}{\beta}\ln M\right\} \tag{4.5}$$

and the M grid points are equally distributed between the intervals $[0,\tau]$ and $[\tau,T]$. In [11, Theorem 4] error bounds are given and they prove that the numerical method converges globally and uniformly with second order in space and first order in time (except for a logarithmic factor.)

In [10] several numerical results are given when this strategy is applied to parabolic problems in two space dimensions. For example, consider the following test problem

$$\begin{aligned}
Lu &:= \varepsilon(u_t - u_{x_1x_1} - u_{x_2x_2}) + (1+x_1+x_2+t)u\\
&= 4(x_1(1-x_1)+x_2(1-x_2)), \quad (x_1,x_2,t) \in \Omega\times(0,1],\\
u(x_1,x_2,t) &= 0, \quad (x_1,x_2) \in \partial\Omega, \quad t \in (0,1],\\
u(x_1,x_2,0) &= g(x_1,x_2) = \sin\left(\frac{5\pi}{4}x_1 + \frac{3\pi}{4}\right)\sin\left(\frac{5\pi}{4}x_2 + \frac{3\pi}{4}\right),\ (x_1,x_2) \in \Omega,
\end{aligned}$$

where $\Omega = (0,1)^2$. Observe that this problem has an incompatibility between the initial condition $g(x_1,x_2)$ and the boundary condition along the faces $x_1 = 0$ and $x_2 = 0$. For this problem, the function s in (4.1) is defined as

$$s(x_1,x_2,t) := g(x_1,x_2)e^{-b(0,0,0)t/\varepsilon}\,\mathrm{erf}\left(\frac{x_1}{2\sqrt{t}}\right)\mathrm{erf}\left(\frac{x_2}{2\sqrt{t}}\right).$$

The function y is approximated with the standard extension of scheme (2.5) to the two dimensional case on a Shishkin mesh which is given as in (2.6) for each variable x_1 and x_2, and the piecewise mesh (4.5) for the time variable. The uniform orders of convergence computed with this method suggest that this method converges globally and uniformly to the function y.

5. Further remarks

The first approach is purely numerical; uses a classical finite difference operator on a standard Shishkin mesh and generates a global approximation with simple bilinear interpolation. However, it generates a globally accurate approximation to the solution of a regularized problem, which is close to the solution of the original problem only outside an ε-dependent neighbourhood of the location of the singularity induced by the non-smooth data. This first approach is relatively easy to implement and the technique could be extended to other classes of singularly perturbed problems with non-smooth data. To generate a globally accurate approximation throughout the entire domain, the second approach uses a special fitted finite difference operator coupled with a sophisticated form of interpolation. The resulting global parameter-uniform accuracy has, at present, only been observed numerically and the form of

interpolation used limits the potential of extending this approach to other classes of problems. The final approach is a mixed analytical/numerical technique which uses a significant amount of a priori knowledge about the nature of the singularity induced by the lack of smoothness in the data. Nevertheless, this final approach has been extended to a wider class of singularly perturbed problems with non-smooth data and does produce parameter-uniform globally accurate approximations.

Acknowledgements

The research of J.L. Gracia has been partially supported by the Institute of Mathematics and Applications (IUMA), the project PID2019-105979GB-I00 and the Diputación General de Aragón (E24-17R).

References

[1] N.S. Bakhvalov, On the optimization of methods for solving boundary value problems with boundary layers (in Russian), Zh. Vychisl. Mat. i Mat. Fis., **9**, 1969, 841–859.

[2] P.A. Farrell, A.F. Hegarty, J.J.H. Miller, E. O'Riordan and G.I. Shishkin, Robust computational techniques for boundary layers, CRC Press, 2000.

[3] J.L. Gracia and E. O'Riordan, A singularly perturbed convection diffusion parabolic problem with an interior layer. Proceedings of the International Conference on Boundary and Interior Layers - Computational and Asymptotic Methods, BAIL 2010, Zaragoza, July 2010, C. Clavero, J.L. Gracia, F.J. Lisbona (Eds.), Lecture Notes in Comput. Sci. Engineering, **81**, Springer, 2011, 139–146.

[4] J.L. Gracia and E. O'Riordan, A singularly perturbed convection–diffusion problem with a moving interior layer, Int. J. Numer. Anal. Mod. **9** 2012, 823-843.

[5] J.L. Gracia and E. O'Riordan, A singularly perturbed parabolic problem with a layer in the initial condition, Appl. Math. Comp. **219** (2012) 498–510.

[6] J.L. Gracia and E. O'Riordan, A singularly perturbed reaction-diffusion problem with incompatible boundary-initial data, Lecture Notes in Computer Science, I. Dimov, I. Farago, and L. Vulkov (Eds.): Numerical Analysis and Its Applications: 5th International Conference, NAA 2012, Lozenetz, Bulgaria, June 15-20, 2012, Revised Selected Papers, v. 8236, 303–310. Springer, Heidelberg (2013).

[7] J.L. Gracia and E. O'Riordan, Interior layers in a singularly perturbed time dependent convection–diffusion problem, Int. J. Numer. Anal. Mod. **11**, 2014, 358–371.

[8] J.L. Gracia and E. O'Riordan, A singularly perturbed convection-diffusion problem with a moving pulse, J. Comput. Appl. Math. **321**, 2017, 371–388.

[9] J.L. Gracia and E. O'Riordan, Singularly perturbed initial-boundary value problems with a pulse in the initial condition. Proceedings of the International Conference on Boundary and Interior Layers - Computational and Asymptotic Methods, BAIL 2016, Beijing (China), August 2016, Z. Huang, M. Stynes, Z. Zhang (Eds.). Lecture Notes in Comput. Sci. Engineering **120**, Springer, 87–99, (2017).

[10] J. L. Gracia and E. O'Riordan, Numerical methods for singularly perturbed parabolic problems with incompatible boundary-initial data in two space dimensions, Proceedings of the International Conference on Boundary and Interior Layers - Computational and Asymptotic Methods, BAIL 2018, Glasgow, Barrenechea, Gabriel R., Mackenzie, John, (Eds.), Lecture Notes in Computational Science and Engineering, **135**, Springer, 171–182, 2020.

[11] J.L. Gracia and E. O'Riordan, Parameter-uniform numerical methods for singularly perturbed parabolic problems with incompatible boundary-initial data, Appl. Numer. Math. **146**, 2019, 436–451.

[12] J. L. Gracia and E. O'Riordan, Singularly perturbed reaction-diffusion problems with discontinuities in the initial and/or the boundary data, *J. Comput. Appl. Math.*, **370**, 112638, 17pp, (2020).

[13] P.W. Hemker and G.I. Shishkin, Approximation of parabolic PDEs with a discontinuous initial condition, *East-West J. Numer. Math*, **1**, 1993, 287–302.

[14] P.W. Hemker and G.I. Shishkin, Discrete approximation of singularly perturbed parabolic PDEs with a discontinuous initial condition, *Comp. Fluid Dynamics*, **2**, 1994, 375–392.

[15] P.W. Hemker, G.I. Shishkin and L.P. Shishkina, ε-uniform schemes with high-order time-accuracy for parabolic singular perturbation problems, *IMA J. Numer. Anal.* **20**, 2000, 99–121.

[16] J.J.H. Miller, E. O'Riordan, G.I. Shishkin, *Fitted Numerical Methods for Singular Perturbation Problems*, World Scientific, Singapore, 2012 (revised edition).

[17] E. O'Riordan and G.I. Shishkin, Singularly perturbed parabolic problems with non-smooth data, *J. Comput. Appl. Math.*, **166**, 2004, 233-245.

[18] H.-G. Roos, M. Stynes and L- Tobiska, *Robust numerical methods for singularly perturbed differential equations*, volume 24 of Springer Series in Computational Mathematics. Springer-Verlag, Berlin, second edition, 2008. Convection-diffusion-reaction and flow problems.

[19] G.I. Shishkin, Approximation of solutions of singularly perturbed boundary value problems with a parabolic boundary layer, *USSR Comput. Maths. Math. Phys.*, **29**, 1989, 1–10.

[20] G.I. Shishkin, A difference scheme for a singularly perturbed equation of parabolic type with discontinuous coefficients and concentrated factors, *Zh. Vychisl. Mat. i Mat. Fiz.*, **29**, 1989, 1277–1290.

[21] G.I. Shishkin, *Discrete approximation of singularly perturbed elliptic and parabolic equations*. Russian Academy of Sciences, Ural Section, Ekaterinburg (1992).

[22] G.I. Shishkin, Grid approximation of singularly perturbed parabolic reaction-diffusion equations with piecewise smooth initial-boundary conditions, Math. Model, Anal., **12**, 2007, 235-254.

Observability and control of parabolic equations on networks

Jone Apraiz[1], **Jon Asier Bárcena-Petisco**[2]

1. jone.apraiz@ehu.eus, *Department of Mathematics, University of the Basque Country UPV/EHU, Barrio Sarriena s/n, 48940, Leioa, Spain*

2. jonasier.barcena@ehu.eus, *Department of Mathematics, University of the Basque Country UPV/EHU, Barrio Sarriena s/n, 48940, Leioa, Spain*

Abstract

In this article, we show some results we obtained related to observability and control of parabolic equations on networks. By using a novel Carleman inequality, we found that the observability of the entire network could be achieved under certain hypothesis about the position of the observation domain. The main difficulty we tackled, due to the existence of loops, was to avoid entering into a circular fallacy, notably in the construction of the auxiliary function for the Carleman inequality. The difficulty was overcomed with a careful treatment of the boundary terms on the junctions. Finally, we used the observability to prove the null controllability of the network and to obtain the Lipschitz stability for an inverse problem consisting on retrieving a stationary potential in the parabolic equation from measurements on the observation domain.

1. Introduction

Network theory can be useful for studying complex systems such as those that arise, for example, in physical sciences, engineering, economics and sociology. These systems can be modeled as networks, also known as metric graphs, and their elements and interactions or links are identified respectively by vertices and edges. During the last decades, the use of networks has been helpful and effective, among others, in the study of pipes, neural systems (the brain can be thought of as a network of neurons), the flow of traffic on roads, the global economy and the human circulatory system (see, for example, [4, Chapter 9], [5, 8, 16, 18]).

In this work, we consider the propagation of heat on a network with loops. We seek to control these networks by acting in its interior with a source term, and to estimate the solutions with an observation domain located in the interior of the network. Indeed, the main purpose of this research is to extend the results of [13] to networks with loops. This is relevant considering that loops arise naturally in pipe systems, transport systems, etc.

Recent important works involving the control of parabolic equations on networks are the followings: [11], where the controllability of the discretized heat equation is studied, [6], where bilinear controls are analyzed on networks, [17], where the optimal control is studied in time-fractional diffusion equations and, [2], where the controllability is analyzed with vanishing viscosity. Note that the literature related to the controllability of hyperbolic equations on networks is more extensive, on which we may highlight the book [7] and the paper [12]. Particularly, [7] mainly analyzes the problem of propagation, observation and control of waves on planar one-dimensional networks, using groundbreaking developments related to non-harmonic Fourier series, Diophantine approximation, graph theory and wave propagation techniques (d'Alembert formula, for example).

1.1. Basic definitions

We first define some concepts related to graph theory that we use in this work. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph.

- An edge $e \in \mathcal{E}$ that is incident to the vertices v and $\tilde{v} \in \mathcal{V}$ is expressed as $e = v\tilde{v}$, where v and $\tilde{v}$ are the *ends* of e. The set of ends of e is denoted by $\mathcal{V}(e)$. Similarly, for every vertex $v \in \mathcal{V}$, $\mathcal{E}(v)$ denotes the *set of edges incident to* v. The *degree of a vertex* $v \in \mathcal{V}$, denoted by $d(v)$, is $|\mathcal{E}(v)|$.
- $\mathcal{V}_0 = \{v : |\mathcal{E}(v)| \geq 2\}$ denotes the *set of inner vertices*, and $\mathcal{V}_\partial = \mathcal{V} \setminus \mathcal{V}_0$ denotes the *set of boundary vertices* of the graph.

- A sequence of vertices $v_0v_1...v_{n-1}v_n$ such that $v_i \in \mathcal{V}$ for all $i = 0, ..., n$ and such that $v_{i-1}v_i \in \mathcal{E}$ for all $i = 1, ..., n$ is called a *walk*. If all the vertices are distinct, it is called a *path*, and if all the vertices are distinct except $v_0 = v_n$ it is called a *cycle*.
- A graph is *connected* if there is a walk joining each pair of vertices.
- A graph isomorphic to $(\{v_0, ..., v_n\}, \{v_0v_1, ..., v_{n-1}v_n\})$ for some $n \in \mathbb{N}$ is called *path graph*.

We define a *network* as a tuple $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{I})$, where $\mathcal{V}$ is a finite *set of vertices*, $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the *set of edges*, and $\mathcal{I}$ is the *identification* of each edge $e = v\tilde{v}$ and its ends v and $\tilde{v}$ with a closed interval $[0, L^e]$ and the ends 0 and L^e respectively. Formally, the identification can be viewed as a function from $\mathcal{E}$ to $X \times \mathcal{V} \times \mathcal{V}$, where X is the set of compact intervals. Notably, $I(e) = ([0, L^e], v, \tilde{v})$, where v is the end of e identified with 0, and $\tilde{v}$ is the other end of e, which we identify with L^e. This definition is also referred in the literature as *metric graph*. With the identification $\mathcal{I}$, for every edge $e \in \mathcal{E}$, we define the following numbers for the vertex v identified with 0 and for the vertex $\tilde{v}$ identified with L^e:

$$n^e(v) = -1 \text{ and } n^e(\tilde{v}) = 1.$$

This allows to define the operator $\partial_{n^e(v)} y = n^e(v)\partial_x y^e(v)$, which can be shorten to $\partial_{n^e} y$ when the vertex is clear. Usually, we make a small abuse of notation and do not write the identification $\mathcal{I}$ explicitly when we denote the network $\mathcal{G}$.

In this article we work in the functional spaces $W^{k,p}_{pw}(\mathcal{E})$, which denotes the set of functions that belong to $W^{k,p}(e)$ for all $e \in \mathcal{E}$, $k \in \mathbb{N}$ and $1 \leq p \leq \infty$, and $W^{1,p}(\mathcal{E}) := W^{k,p}_{pw}(\mathcal{E}) \cap C^0(\overline{\mathcal{E}})$. Here "pw" stands for piecewise. Similar definitions apply to $H^k_{pw}(\mathcal{E})$ and $H^k(\mathcal{E})$. In this context, given a function $f \in W^{k,p}_{pw}(\mathcal{E})$, we define by $\partial_x f$ the derivative in each of the edges. Clearly, if $f \in W^{k,p}_{pw}(\mathcal{E})$, then $\partial_x f \in W^{k-1,p}_{pw}(\mathcal{E})$.

1.2. The controllability result

The problem that we study here is the dynamics of the flux and the control, which can be modelled by the following parabolic system:

$$\begin{cases} a\partial_t y - \mu\partial^2_{xx} y + b\partial_x y + cy = f1_\omega, & \text{in } (0,T) \times \mathcal{E}, \\ y = 0, & \text{on } (0,T) \times \mathcal{V}_\partial, \\ y^{e_i} = y^{e_j}, & \text{on } (0,T) \times \mathcal{V}_0, \ \forall e_i, e_j \in \mathcal{E}(v), \\ \displaystyle\sum_{e \in \mathcal{E}(v)} \mu^e \partial_{n^e} y^e = \gamma y, & \text{on } (0,T) \times \mathcal{V}_0, \\ y(0,\cdot) = y_0, & \text{in } \mathcal{E}. \end{cases} \tag{1.1}$$

In this model y denotes the flux of the heat on the entire network. Throughout this article, we denote the restriction of a function to an edge e by adding the superscript e. In addition, a and μ are positive coefficients and b and c are coefficients which characterize the properties of the pipes of the network (roughness or properties of the heat flux, for example). Moreover, γ is a real coefficient measuring the flux of the heat on junctions, and $\omega \subset \mathcal{E}$ is the control domain. Here, when writing $\omega \subset \mathcal{E}$ we make a small abuse of notation to mean $\omega \subset \cup_{e \in \mathcal{E}} e$.

In addition, by coefficients we mean functions which model the properties of the systems like the heat diffusivity and, unless stated otherwise, depend on the time and spatial variables.

It is trivial to prove in system (1.1) the usual energy estimations in $L^2(0,T; H^1(\mathcal{E})) \cap C^0([0,T]; L^2(\mathcal{E}))$ and regularity result in $L^2(0,T; H^2_{pw}(\mathcal{E})) \cap H^1(0,T; L^2(\mathcal{E}))$ for parabolic equations. This can be done by multiplying the first equation of the system by y and y_t and integrating it in $(0,T) \times \mathcal{E}$ (see [9] for a particular case).

In order to solve the controllability and inverse problems with respect to the parabolic system (1.1), we study the observability properties of the adjoint system, which is given by:

$$\begin{cases} -a\partial_t\varphi - \mu\partial_{xx}^2\varphi - \partial_x b\varphi - b\partial_x\varphi + c\varphi = 0, & \text{in } (0,T)\times\mathcal{E}, \\ \varphi = 0, & \text{on } (0,T)\times\mathcal{V}_\partial, \\ \varphi^{e_i} = \varphi^{e_j}, & \text{on } (0,T)\times\mathcal{V}_0,\ \forall e_i, e_j \in \mathcal{E}(v), \\ \displaystyle\sum_{e\in\mathcal{E}(v)} \mu^e \partial_{n^e}\varphi^e = \left(-\sum_{e\in\mathcal{E}(v)} n^e b^e + \gamma\right)\varphi, & \text{on } (0,T)\times\mathcal{V}_0, \\ \varphi(T,\cdot) = \varphi_T, & \text{in } \mathcal{E}. \end{cases} \tag{1.2}$$

System (1.2) might not be observable unless the control domain intersects a sufficient number of edges. In particular, in order to avoid some of those non-observable cases, we assume that the control domain intersects a sufficient number of edges:

Hypothesis 1 (Existence of an indexing function) *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a network and $\omega \subset \mathcal{E}$ an open subdomain. We suppose that*:

1. *ω intersects all the cycles of $\mathcal{G}$. That is, if $v_1, ..., v_n \in \mathcal{V}$ such that $e_1 := v_1v_2, ..., e_{n-1} := v_{n-1}v_n, e_n := v_nv_1$ satisfy $e_i \in \mathcal{E}$ for all $i = 1, ..., n$, then there is $k \in \{1, ..., n\}$ such that $e_k \cap \omega \neq \emptyset$.*

Moreover, we suppose that there exists a function $u : \{e \in \mathcal{E} : e \cap \omega = \emptyset\} \mapsto \mathcal{V}_0$ such that:

2. *u is injective.*
3. *e is incident to $u(e)$.*

Roughly speaking, the state of the equation in the edge e is controlled by ω if $e \cap \omega \neq \emptyset$, and by $u(e)$ otherwise, which is controlled by the rest of the adjacent edges. Identifying the right hypothesis, in the sense that allows us to prove the results without being too restrictive, is not trivial and is one of the contributions of our work. Indeed, the main breakthrough with respect to the previous work, and notably [13], is to make sure that we do not enter a circular reasoning fallacy. This is done with Hypothesis 1, as the proof of the controllability follows in a fluid way.

Remark 1.1 (Identification of edges) Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a network and let ω be a control domain such that Hypothesis 1 is satisfied with an indexing function u that we fix. In order to identify an edge e such that $\omega \cap e = \emptyset$ with an interval $[0, L^e]$, we establish that the end identified with L^e is the vertex $u(e)$. This assignment simplifies some computations in the proof of Proposition 2.1.

The main observability result in this work is a Carleman inequality (see Proposition 2.4 in Section 2.2). With that inequality, we prove the following null controllability result regarding open-loop control:

Theorem 1.2 (Controllability of the heat equation on networks with loops) *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a network satisfying Hypothesis 1, $a, \mu \in W^{1,\infty}((0,T); L^\infty(\mathcal{E})) \cap L^\infty((0,T); W^{1,\infty}_{pw}(\mathcal{E}))$ such that $\inf a, \inf \mu > 0$, $b \in L^\infty((0,T); W^{1,\infty}_{pw}(\mathcal{E}))$, $c \in L^\infty((0,T)\times\mathcal{E})$ and $\gamma \in L^\infty((0,T)\times\mathcal{V}_0)$. Then, there exists $C > 0$ such that for all $y_0 \in L^2(\mathcal{E})$ there is $f \in L^2((0,T)\times\omega)$ such that*:

$$\|f\|_{L^2((0,T)\times\omega)} \leq C\|y_0\|_{L^2(\mathcal{E})},$$

and the solution of (1.1) *satisfies $y(T,\cdot) = 0$.*

Theorem 1.2 is proved by duality with the results of Section 2 (see [1]).

1.3. Application to the resolution of inverse problems

Carleman estimates can also be used to obtain results in the field of inverse problems, which is an additional objective of our work. In fact, the link between Carleman inequalities and their applications is well known. Some important references regarding this topic include [14, 15], and detailed surveys are included in [3, 19].

In this work, we seek to generalize the results of [13] to systems with loops. With that purpose, let us consider the system:

$$\begin{cases} \partial_t y - \mu \partial_{xx}^2 y + py = 0, & \text{in } (0,T)\times\mathcal{E}, \\ y = 0, & \text{on } (0,T)\times\mathcal{V}_\partial, \\ y^{e_i} = y^{e_j}, & \text{on } (0,T)\times\mathcal{V}_0,\ \forall e_i, e_j \in \mathcal{E}(v), \\ \displaystyle\sum_{e\in\mathcal{E}(v)} \mu^e \partial_{n^e} y^e = \gamma y, & \text{on } (0,T)\times\mathcal{V}_0, \\ y(0,\cdot) = y_0, & \text{in } \mathcal{E}, \end{cases} \tag{1.3}$$

for μ a piecewise constant function, γ a real parameter, y_0 the initial state and p the static potential. Moreover, we denote by $y[p, y_0]$ the solution of (1.3).

Our objective is to recover the potential p by making measurements on the flux of the heat at a time $t_0 > 0$ and also on the observation domain ω but throughout the whole time interval $(0, T)$. In particular, we prove the following result:

Theorem 1.3 (Resolution of an inverse problem) *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a network satisfying Hypothesis 1, $p \in L^\infty(\mathcal{E})$, $r > 0$ and $y_0 \in L^2(\mathcal{E})$ such that $y[p, y_0] \in H^1(0,T; H^2_{pw}(\mathcal{E})) \cap H^2(0,T; L^2_{pw}(\mathcal{E}))$ and such that for some $t_0 \in (0,T)$ the following estimate holds*:

$$|y[p, y_0](t_0, \cdot)| \geq r \quad \text{in } \mathcal{E}. \tag{1.4}$$

Then, for any $m > 0$, there is a constant $C(m, r, T, \|\partial_t y[p, y_0]\|_{L^\infty((0,T)\times\mathcal{E})})$ such that for any $q \in L^\infty(\mathcal{E})$ satisfying:

$$\|q\|_{L^\infty(\mathcal{E})} \leq m,$$

we have:

$$\|q - p\|_{L^2(\mathcal{E})} \leq C\left(\|y[p, y_0](t_0, \cdot) - y[q, y_0](t_0, \cdot)\|_{H^2(\mathcal{E})} + \|y[p, y_0] - y[q, y_0]\|_{H^1(0,T;L^2(\omega))}\right). \tag{1.5}$$

The proof of Theorem 1.3 is an easy consequence of the Carleman inequality in Proposition 2.4 (see [1]).

2. The observability problem

In this section we explain the main results and tools we need in order to prove the observability inequality for system (1.2). With that purpose, in Section 2.1 we construct an auxiliary function of Fursikov-Imanuvilov type, and in Section 2.2, using appropriate weights, we show the observability of system (1.2) with a Carleman inequality.

2.1. Construction of the auxiliary function

In this section we construct an auxiliary function that is required to define the Fursikov-Imanuvilov weights in Section 2.2. Throughout this section we consider an open subdomain $\tilde{\omega} \subset \omega$ such that $\overline{\tilde{\omega}} \subset \omega$ and such that, for all $e \in \mathcal{E}$, $\tilde{\omega} \cap e \neq \emptyset$ if and only if $\omega \cap e \neq \emptyset$.

We need to make sure that for all edge e, if $e \cap \tilde{\omega} \neq \emptyset$, the maximum of η^e is achieved in $\tilde{\omega}$ and if $e \cap \tilde{\omega} = \emptyset$, the maximum of η^e is achieved on $u(e)$, being its derivative small near $u(e)$. As the "smallness" depends on the coefficients of the system, we get a family of auxiliary functions whose derivatives near $u(e)$ are as small as needed, and such that they are uniformly bounded in $W^{2,\infty}_{pw}(\mathcal{E})$.

Proposition 2.1 (Construction of the auxiliary function) *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a network, and $\tilde{\omega}$ be a domain satisfying Hypothesis 1 with the indexing function u. We identify the edges of $\mathcal{G}$ as in Remark 1.1. Then, there is $C > 0$ such that for all $\delta \in [0, 1]$ there exists a function η satisfying*:

1. *The function* $\eta \in C^0(\overline{\mathcal{E}}) \cap C^2_{pw}(\mathcal{E})$ *and* $\|\eta\|_{W^{2,\infty}_{pw}} \leq C$.
2. *For all edges* e *such that* $e \cap \tilde{\omega} = \emptyset$, *then*:
   - $\partial_x \eta^e \geq \delta$ *on* e,
   - $\partial_{n^e} \eta^e(0) = -\partial_x \eta^e(0) = -1$,
   - $\partial_{n^e} \eta^e(L^e) = \partial_x \eta^e(L^e) = \delta$.
3. *If an edge* e *that we identify with* $[0, L^e]$ *satisfies* $e \cap \tilde{\omega} \neq \emptyset$, *then*:
   - $|\partial_x \eta| = 1$ *on* $e \setminus \tilde{\omega}$,
   - $\partial_{n^e} \eta^e(0) = -\partial_x \eta^e(0) = -1$,
   - $\partial_{n^e} \eta^e(L^e) = \partial_x \eta^e(L^e) = -1$.

The proof of the existence of such function is based on an induction on the number of edges of $\mathcal{G}$. In order to prove Proposition 2.1, we need to study the case of an edge assuming we have some restrictions on the boundary. This is done with Lemmas 2.2 and 2.3, whose proofs are standard (see [1, 10]). In Lemma 2.2 we study the construction of the auxiliary function for edges that have no intersection with $\tilde{\omega}$.

Lemma 2.2 (Extension of the auxiliary function with one constraint) *Let* $\tilde{\omega} \subset \mathcal{E}$ *be an open subdomain,* $e \simeq [0, L^e]$ *be an edge such that* $\tilde{\omega} \cap e = \emptyset$, $R \in \mathbb{R}$, $p \in \{0, L^e\}$ *and* $\delta \in [0, 1]$. *Then, there is a function* $\eta^e \in C^2(\overline{e})$ *such that*:

- $\partial_x \eta^e \geq \delta$ *on* $[0, L^e]$,
- $\|\eta^e\|_{L^\infty(0,L^e)} \leq |R| + L^e$, $\|\partial_x \eta^e\|_{L^\infty(0,L^e)} \leq 1$, $\|\partial_{xx} \eta^e\|_{L^\infty(0,L^e)} \leq \frac{1}{L^e}$,
- $\partial_{n^e} \eta^e(0) = -\partial_x \eta^e(0) = -1$,
- $\partial_{n^e} \eta^e(L^e) = \partial_x \eta^e(L^e) = \delta$,
- $\eta^e(p) = R$.

Next, in the following Lemma, we study the construction of the auxiliary function for edges which intersect $\tilde{\omega}$.

Lemma 2.3 (Extension of the auxiliary function with two constraints) *Let* $\tilde{\omega} \subset \mathcal{E}$ *be an open subdomain,* $e \simeq [0, L^e]$ *be an edge such that* $\tilde{\omega} \cap e \neq \emptyset$ *and* $R_1, R_2 \in \mathbb{R}$. *Then, there is a function* $\eta^e \in C^2(\overline{e})$ *such that*:

- $\|\eta^e\|_{W^{2,\infty}_{pw}(0,L^e)} \leq C(|R_1|, |R_2|, L^e, \tilde{\omega})$, *for* C *increasing with* $|R_1|$ *and* $|R_2|$ *for a fixed* L^e *and* $\tilde{\omega}$,
- $|\partial_x \eta^e| = 1$ *on* $[0, L^e] \setminus \tilde{\omega}$,
- $\partial_{n^e} \eta^e(0) = -\partial_x \eta^e(0) = -1$,
- $\partial_{n^e} \eta^e(L^e) = \partial_x \eta^e(L^e) = -1$,
- $\eta^e(0) = R_1$, $\eta^e(L^e) = R_2$.

2.2. A new Carleman inequality

The auxiliary function constructed in the previous section allows us to define the usual Fursikov-Imanuvilov weights:

$$\alpha(t,x) := \frac{e^{6\lambda\|\eta\|_\infty} - e^{\lambda(4\|\eta\|_\infty + \eta(x))}}{t(T-t)}, \quad \xi(t,x) := \frac{e^{\lambda(4\|\eta\|_\infty + \eta(x))}}{t(T-t)}, \quad \forall\, (t,x) \in (0,T) \times \mathcal{E}, \tag{2.1}$$

where η is defined in Proposition 2.1 for $\tilde{\omega}$ an open domain compactly included in ω such that $e \cap \tilde{\omega} = \emptyset$ if and only if $e \cap \omega = \emptyset$ for all $e \in \mathcal{E}$, and for $\delta > 0$ a sufficiently small parameter that will be defined in the proof of Proposition 2.4 (see [1]) for absorbing boundary terms. Bearing this in mind, we state the next Carleman inequality:

Proposition 2.4 (A new Carleman inequality) *Let $a, \mu \in W^{1,\infty}((0,T); L^\infty(\mathcal{E})) \cap L^\infty((0,T); W^{1,\infty}_{pw}(\mathcal{E}))$ such that $\inf a, \inf \mu > 0$, $h \in L^\infty((0,T)\times\mathcal{V}_0)$ and $g \in L^2(Q)$. Then, there is $C > 0$ depending on $\mathcal{G}$, ω, a, and μ such that for all $\varphi_T \in L^2(\mathcal{E})$, $\lambda \geq C$ and $s \geq C(T + T^2)$ the following inequality is satisfied*:

$$s^3\lambda^4 \iint_Q e^{-2s\alpha}\xi^3|\varphi|^2 \,\mathrm{d}x\,\mathrm{d}t \leq C\left(s^3\lambda^4 \iint_{Q_\omega} e^{-2s\alpha}\xi^3|\varphi|^2 \,\mathrm{d}x\,\mathrm{d}t + \iint_Q e^{-2s\alpha}|g|^2 \,\mathrm{d}x\,\mathrm{d}t\right), \tag{2.2}$$

for α and ξ the weights defined in (2.1)*, and φ the solution of*:

$$\begin{cases} -a\partial_t\varphi - \mu\partial^2_{xx}\varphi = g, & \text{in } (0,T)\times\mathcal{E}, \\ \varphi = 0, & \text{on } (0,T)\times\mathcal{V}_\partial, \\ \varphi^{e_i} = \varphi^{e_j}, & \text{on } (0,T)\times\mathcal{V}_0,\ \forall e_i, e_j \in \mathcal{E}(v), \\ \displaystyle\sum_{e\in\mathcal{E}(v)} \mu^e \partial_{n^e}\varphi^e = h\varphi, & \text{on } (0,T)\times\mathcal{V}_0, \\ \varphi(T,\cdot) = \varphi_T, & \text{in } \mathcal{E}. \end{cases} \tag{2.3}$$

Acknowledgements

Both authors are supported by the Spanish Government's Ministry of Science, Innovation and Universities (MICINN), under grant PID2021-126813NB-I00 and the Basque Government, under grant IT1615-22.

References

[1] J. Apraiz, J. A. Bárcena-Petisco. Observability and control of parabolic equations on networks with loops. *hal-03501343*, 2022.

[2] J. A. Bárcena-Petisco, M. Cavalcante, G. M. Coclite, N. de Nitti, and E. Zuazua. Control of hyperbolic and parabolic equations on networks and singular limits. *hal-03233211*, 2021.

[3] M. Bellassoued and M. Yamamoto. *Carleman estimates and applications to inverse problems for hyperbolic systems*. Springer, 2017.

[4] V. D. Blondel, E. D. Sontag, M. Vidyasagar, and J. C. Willems. *Open Problems in Mathematical Systems and Control Theory*. Communication and Control Engineering Series. Springer, London, 1999.

[5] J. Brouwer, I. Gasser, and M. Herty. Gas pipeline models revisited: Model hierarchies, non-isothermal models and simulations on networks. *Multiscale Model. Simul.*, 9:601–623, 2011.

[6] P. Cannarsa, A. Duca, and C. Urbani. Exact controllability to eigensolutions of the bilinear heat equation on compact networks. *arXiv:2111.02250*, 2021.

[7] R. Dager and E. Zuazua. *Wave propagation, observation and control in 1-d flexible multistructures*. Volume 50 of Mathematics & Applications. Springer Verlag, Berlin, 2006.

[8] K. Egger and T. Kugler. Damped wave systems on networks: exponential stability and uniform approximations. *Numer. Math.*, 138(4):839–867, 2018.

[9] H. Egger and N. Philippi. On the transport limit of singularly perturbed convection-diffusion problems on networks. *Math. Methods Appl. Sci.*, 44, 2021.

[10] A. V. Fursikov and O. Y. Imanuvilov. *Controllability of evolution equations*. Volume 34 of Lecture Notes Series. Seoul National University, Research Institute of Mathematics, Global Analysis Research Center, Seoul, 1996.

[11] G. Notarstefano and G. Parlangeli. Controllability and observability of grid graphs via reduction and symmetries. *IEEE T. Automat. Contr.*, 58(7):1719–1731, 2013.

[12] F. M. Hante, G. Leugering, A. Martin, L. Schewe, and M. Schmidt. *Challenges in optimal control problems for gas and fluid flow in networks of pipes and canals: From modeling to industrial applications*. Industrial mathematics and complex systems, pages 77–122. Springer, 2017.

[13] L. Ignat, A. F. Pazoto, and L. Rosier. Inverse problem for the heat equation and the Schrödinger equation on a tree. *Inverse Prob.*, 28(1):015011, 2011.

[14] O. Y. Imanuvilov and M. Yamamoto. Lipschitz stability in inverse parabolic problems by the Carleman estimate. *Inverse Prob.*, 14:1229–1245, 1998.

[15] O. Y. Imanuvilov and M. Yamamoto. Global Lipschitz stability in an inverse hyperbolic problem by interior observations. *Inverse Prob.*, 17(4):717, 2001.

[16] L. E. Lagnese, G. Leugering, and E. J. P. G. Schmidt. *Modeling, Analysis and Control of Dynamic Elastic Multi-Link Structures*. Volume 19 of Systems Control: Foundations Applications. Springer Science+Business Media, New York, 1994.

[17] V. Mehandiratta, M. Mehra, and G. Leugering. Optimal control problems driven by time-fractional diffusion equations on metric graphs: optimality system and finite difference approximation. *SIAM J. Control and Optim.*, 59(6):4216–4242, 2021.

[18] M. Newman, A. L. Barabási, and D. J. Watts. *The Structure and Dynamics of Networks*. Volume 19 of Princeton Studies in Complexity. Princeton University Press, 2011.

[19] M. Yamamoto. Carleman estimates for parabolic equations and applications. *Inverse Prob.*, 25(12):123013, 2009.

Eigenvalue problems for the p–Laplacian in the critical range $1 < p < 2$

José C. Sabina de Lis
Departamento de Análisis Matemático and *IUEA*
Universidad de La Laguna

Abstract

A simple variational characterization for the *first nontrivial* eigenvalue to a family of nonlinear eigenvalue problems is discussed in this paper. The best representative is provided by the Neumann problem for the p–Laplacian operator,

$$\begin{cases} -\Delta_p u = \lambda |u|^{p-2}u & x \in \Omega, \\ \dfrac{\partial u}{\partial \nu} = 0 & x \in \partial\Omega, \end{cases}$$

where $\Omega \subset \mathbb{R}^N$ is a bounded $C^{1,\alpha}$ domain, ν is the outer unit normal and $\Delta_p u = \operatorname{div}(|\nabla u|^{p-2}\nabla u)$. The key point is the fact that the exponent p falls in the somehow critical range $1 < p < 2$. Research reported here is inspired by our work [32].

1. Introduction

The study of the spectrum of the p–Laplacian operator Δ_p, constrained by a set of standard boundary conditions, is one of the challenging subjects in nonlinear analysis ([30]). Most of the well–established achievements for the Laplacian Δ ($p = 2$) are still in an early stage of development for $p \neq 2$, to say the less. A tentative list of these subjects may be: distribution of the eigenvalues ([4], [18], [30]), isolation ([5], [29]), multiplicity ([24], [1], [37]), nodal sets ([16], [22]) and Fredholm alternative ([19], [34, 35]).

In order to fix the goals of the present paper let us review some general features on the Neumann problem,

$$\begin{cases} -\Delta_p u = \lambda |u|^{p-2}u & x \in \Omega, \\ \dfrac{\partial u}{\partial \nu} = 0 & x \in \partial\Omega, \end{cases} \tag{1.1}$$

where $\Omega \subset \mathbb{R}^N$ is a bounded smooth domain and $p > 1$. We begin with by recalling the definition of weak eigenvalue: a (weak) eigenfunction $u \in W^{1,p}(\Omega) \setminus \{0\}$ to (1.1) associated to the eigenvalue $\lambda \in \mathbb{R}$ is defined through the equality,

$$\int_\Omega |\nabla u|^{p-2}\nabla u \nabla v \, dx = \lambda \int_\Omega |u|^{p-2}uv \, dx, \tag{1.2}$$

which must be satisfied for all test functions $v \in W^{1,p}(\Omega)$.

It follows from (1.2) that eigenvalues λ to (1.1) must be nonnegative while $\lambda_1 = 0$ is the first (trivial) eigenvalue, which verifies

$$\lambda_1 = 0 = \inf_{u \in W^{1,p}(\Omega)} \frac{\int_\Omega |\nabla u|^p}{\int_\Omega |u|^p}.$$

It is in addition a simple eigenvalue (all possible associated eigenfunctions are constant) and the only one with the property of exhibiting one–signed eigenfunctions. In fact any other eigenfunction u associated to an eigenvalue $\lambda \neq 0$ must satisfy the average condition, $\int_\Omega |u|^{p-2}u = 0$. Hence u must change sign in Ω. Furthermore, $\lambda_1 = 0$ is *isolated* since if were $\lambda_n \to 0$ for a sequence λ_n of nontrivial eigenvalues, then a properly normalized sequence u_n of associated eigenfunctions would keep the sign in Ω for large n what is not possible.

On the other hand, an infinite amount of other eigenvalues to (1.1), under the form of an increasing sequence $\lambda = \lambda_n^{SL} \to \infty$, can be obtained by the minimax procedure,

$$\lambda_n^{LS} = \inf_{S \in \mathcal{S}_n} \sup_{u \in S} \frac{\int_\Omega |\nabla u|^p}{\int_\Omega |u|^p}.$$

Here, $\mathcal{S}_n = \{S \subset W^{1,p}(\Omega) : S \text{ compact}, \ S = -S, \ \gamma(S) \geq n\}$ where γ stands for the Krasnoselskii genus of A ([33]). This means that the class of positive eigenvalues is non void. As $\lambda_1 = \lambda_1^{LS} = 0$ is isolated, a *second* eigenvalue λ_2 to (1.1) exists and is defined as,

$$\lambda_2 = \min\{\lambda > 0 : \ \lambda \text{ is an eigenvalue to (1.1) }\}.$$

The main objective of this work is just presenting a *simple* variational characterization of λ_2. In addition, to obtain a similar result for a broader class of eigenvalue problems. To review background findings let us mention that a first contribution in this direction was obtained in [3] and states,

$$\lambda_2 = \lambda_2^{SL} = \inf_{S\in\mathcal{S}_2} \sup_{u\in S} \frac{\int_\Omega |\nabla u|^p}{\int_\Omega |u|^p}. \tag{1.3}$$

A further alternative expression to (1.3) is obtained by replacing $\mathcal{S}_2$ by the more friendly class Σ_2. It consists of the images $h(\mathbb{S}^1)$, $h : \mathbb{S}^1 \to W^{1,p}(\Omega)$, h being an odd continuous mapping, $\mathbb{S}^1$ the unit circumference ([21]).

A third representation of λ_2 is,

$$\lambda_2 = \inf_{\phi\in\Gamma} \sup_{s\in[-1,1]} \frac{\int_\Omega |\nabla\phi(s)|^p}{\int_\Omega |\phi(s)|^p}, \tag{1.4}$$

where $\Gamma = \{\phi : [-1,1] \to W^{1,p}(\Omega) \setminus \{0\} : \phi \text{ continuous}, \pm\phi(\pm 1) \geq 0\}$. Actually, (1.4) was shown in [7] for a more general kind of nonsymmetric problems including (1.1) as a particular case. Despite all these strange characterizations, there still exists a simpler expression for λ_2. Namely,

$$\lambda_2 = \inf\left\{ \frac{\int_\Omega |\nabla u|^p}{\int_\Omega |u|^p} : u \in W^{1,p}(\Omega) \setminus \{0\}\,,\ \int_\Omega |u|^{p-2}u = 0 \right\}. \tag{1.5}$$

It should be remarked that (1.5) exactly matches with the natural formula for the linear case $p = 2$. As a matter of fact, (1.5) is presented in [10] without proof under the status of a well–known result. Nevertheless, while the case $p \geq 2$ is straightforward, to show (1.5) in the critical case $1 < p < 2$ is far from obvious (see [11]). In fact, (1.5) is just shown in [26] only when $p \geq 2$. This paper is devoted to produce an independent proof of (1.5) when p falls in the critical regime $1 < p < 2$. We are also pursuing a similar result in a variety of problems ranging from Steklov conditions to the p–Laplacian on graphs.

The work is organized as follows. Section 2 states the differentiability of the variance functional. It constitutes the key tool for our proofs. Two different versions of the Neumann problem (1.1) are addressed in Sections 3 and 4, while the Stekloff problem is considered in Section 5. The eigenvalue problem for Δ_p in a compact and connected Riemannian manifold without boundary is presented in Section 6. An N–dimensional version of the periodic eigenvalue problem is analyzed in Section 7. The work concludes by studying the p–Laplacian on graphs in Section 8.

2. Variance functional

We begin with a basic result.

Lemma 2.1 *Let (X,μ) be a measurable space, $u \in L^p(X)$ with $p > 1$. Then there exists a unique $\tilde{u} \in \mathbb{R}$ so that*:

$$\|u - \tilde{u}\|_p = \inf_{t\in\mathbb{R}} \|u - t\|_p, \qquad \|u\|_p := \|u\|_{L^p(X)}.$$

Moreover, $t = \tilde{u}$ is characterized as the solution to equation $\int_\Omega |u-t|^{p-2}(u-t) = 0$. Furthermore, the functional $M : L^p(X) \to \mathbb{R}$, $M(u) = \tilde{u}$, is continuous.

Definition 2.2 Value $\tilde{u}$ is defined as the p–average of $u \in L^p(X)$ while $V_p(u) = \int_\Omega |u - \tilde{u}|^p$, is the variance of u.

Our main result in this section reads as follows ([32]).

Theorem 2.3 *Let (X,μ) be a measurable space. Then, the variance functional V_p is Fréchet differentiable in $L^p(X)$ for all $p > 1$. Moreover, its differential $DV_p(u)$ at u is represented as*:

$$\left\langle DV_p(u), v\right\rangle = p\int_X |u-\tilde{u}|^{p-2}(u-\tilde{u})v\,d\mu, \qquad v \in L^p(X). \tag{2.1}$$

Remark 2.4 Proof of Theorem 2.3 is straightforward if $p \geq 2$ and $u \in L^p(X)$ is not a constant function. In fact, the average functional $M(u) = \tilde{u}$ can be computed by solving in t the equation,

$$b(u-t) = 0, \qquad b(v) = \int_X |v|^{p-2}v, \qquad v \in L^p(X). \tag{2.2}$$

The Implicit Function Theorem then implies the existence of the differential $DM(u)$ so that,

$$\langle DM(u), v\rangle = \frac{\int_X |u-\tilde{u}|^{p-2}v}{\int_X |u-\tilde{u}|^{p-2}}, \qquad v \in L^p(X).$$

Thus, both the differentiability of DV_p and the expression (2.1) are obtained from the identity,

$$V_p(u) = \int_X |u - M(u)|^p.$$

Remark 2.5 The approach in Remark 2.4 can not be employed to manage the case $1 < p < 2$ since functional b in (2.2) in no more C^1. To deal with this critical case, variance V_p is regarded as the infimum of a family of functions $f_u(t)$,

$$V_p(u) = \inf_{t\in\mathbb{R}} f_u(t), \qquad f_u(t) = \int_X |u-t|^p,$$

having $u \in L^p(X)$ as a parameter. Then, the differentiability of V_p with respect to u is shown by directly using this variational representation. Reader is referred to [32] for full details.

3. The Neumann problem

In this section we are concerned with a slightly more general version of (1.1). Namely,

$$\begin{cases} -\Delta_p u = \lambda\, m(x)|u|^{p-2}u, & x \in \Omega, \\ \dfrac{\partial u}{\partial \nu} = 0, & x \in \partial\Omega, \end{cases} \tag{3.1}$$

where $m > 0$ a. e. in Ω and either $m \in L^\infty(\Omega)$ or $m \in L^r(\Omega)$ with $r > \left(\frac{p^*}{p}\right)' = \frac{N}{p}$ if $1 < p \le N$, $r = 1$ otherwise.

Point iii) of the next statement provides us with a proof of identity (1.5), a principal objective of this work. Notice that i), ii) are well–known and are included here for completeness. However, iii) requires a careful analysis in the critical regime $1 < p < 2$.

Theorem 3.1 *Let $\Omega \subset \mathbb{R}^N$ be a bounded $C^{0,1}$ domain and set*

$$\hat{\lambda} = \inf_{u\in\mathcal{M}_0\setminus\{0\}} \frac{\int_\Omega |\nabla u|^p\, dx}{\int_\Omega |u|^p\, m\, dx},$$

where $\mathcal{M}_0 = \{u \in W^{1,p}(\Omega) : \int_\Omega |u|^{p-2}u\, m\, dx = 0\}$. Then,

i) The infimum is achieved at some $\hat{u} \in \mathcal{M}_0$ and thus $\hat{\lambda} > 0$.

ii) Every eigenvalue $\lambda \neq 0$ to the Neumann problem (3.1) *satisfies $\lambda \ge \hat{\lambda}$. In particular, $\lambda = 0$ is an isolated eigenvalue.*

iii) $\hat{\lambda}$ is actually an eigenvalue *and therefore $\lambda_2 = \hat{\lambda}$.*

Proof While i) is shown by means of the direct methods in calculus of variations, assertion ii) is a consequence of the expression of $\hat{\lambda}$. As for point iii) first notice that $\widetilde{u-\tilde{u}} = 0$ for every $u \in W^{1,p}(\Omega)$ and so,

$$\mathcal{M}_0 = \{u \in W^{1,p}(\Omega) : \tilde{u} = 0\} = \{u - \tilde{u} : u \in W^{1,p}(\Omega)\}.$$

Accordingly, an alternative writing for $\hat{\lambda}$ is,

$$\hat{\lambda} = \inf_{v\in\mathcal{M}_0} \frac{\int_\Omega |\nabla v|^p}{\int_\Omega |v|^p} = \inf_{u\in W^{1,p}(\Omega)} \frac{\int_\Omega |\nabla(u-\tilde{u})|^p}{\int_\Omega |u-\tilde{u}|^p} = \inf_{u\in W^{1,p}(\Omega)} \frac{\int_\Omega |\nabla u|^p}{V_p(u)} =: \inf_{u\in W^{1,p}(\Omega)} Q(u),$$

where $u \notin \text{span}\,(1_\Omega)$. Since Q achieves its minimum at $u = \hat{u}$ then,

$$\langle DQ(\hat{u}), v\rangle = 0, \qquad v \in W^{1,p}(\Omega).$$

By using the differentiability of V_p, which holds regardless the value of $p > 1$ is, and (2.1) we get,

$$\int_\Omega |\nabla\hat{u}|^{p-2}\nabla\hat{u}\nabla v\, dx = \hat{\lambda}\int_\Omega |\hat{u}|^{p-2}\hat{u}v\, m\, dx, \qquad v \in W^{1,p}(\Omega),$$

and the conclusion follows from the fact $\hat{u}$ has zero average. □

Remark 3.2 An alternative proof assertion iii) can be obtained by means of the approach in [17] (see a similar reasoning in [11]).

4. Generalized Neumann problems

A more general version of (3.1) is provided by the problem,

$$\begin{cases} -\Delta_p u = \lambda m(x)|u|^{q-2}u & x \in \Omega, \\ \dfrac{\partial u}{\partial \nu} = 0 & x \in \partial\Omega, \end{cases} \tag{4.1}$$

where $1 < q < p^*$, $m > 0$ a. e. in Ω with $m \in L^\infty(\Omega)$ or either, $m \in L^r(\Omega)$, and $r > \left(\frac{p^*}{q}\right)' = \frac{Np}{N(p-q)+pq}$ if $1 < p \leq N$ together with $r = 1$ when $p > N$. The Dirichlet counterpart of (4.1) was formerly studied in [24] while a later discussion in $N = 1$ is contained in [20]. On the other hand, (4.1) is termed as a pq–generalized eigenvalue problem in [12].

Next result furnishes the existence of a principal branch of nontrivial solutions to (4.1). Its proof is obtained by combining Theorem 3.1 with the arguments in Section 3 (see [32]).

Theorem 4.1 *Assume that the exponent q and weight m satisfy the prescribed restrictions. Then problem* (4.1) *satisfies the following properties.*

i) *Existence of nontrivial solutions can only happen if $\lambda > 0$.*

ii) *Every nontrivial solution to* (4.1) *satisfies the average condition $\int_\Omega |u|^{q-2}u\, m dx = 0$.*

iii) *There exists a family of nontrivial solutions (λ, u_λ), $u_\lambda = \pm\left(\frac{\lambda}{\lambda_0}\right)^{\frac{1}{p-q}}(u - \tilde{u})$, $\lambda_0 = V_q(u)^{\frac{p}{q}-1}\mu$, where u is a minimizer to,*

$$\mu = \inf\{\int_\Omega |\nabla u|^p \ : \quad \int_\Omega |u|^q \, m\, dx = 1\,, \quad \int_\Omega |u|^{q-2}(u)\, m\, dx = 0\}. \tag{4.2}$$

Remark 4.2 An infinite collection of branches $\mathcal{F}_n = \{(\lambda, u_\lambda^{(n)})\}$ of solutions to (4.1) can be found by substituting the 'inf' procedure in (4.2) by a suitable mini–max argument.

5. The Stekloff problem

We are now interested in the Steklov eigenvalue problem,

$$\begin{cases} -\Delta_p u = 0, & x \in \Omega, \\ |\nabla u|^{p-2}\dfrac{\partial u}{\partial \nu} = \lambda |u|^{p-2}u, & x \in \partial\Omega, \end{cases} \tag{5.1}$$

which has received much attention in recent literature, as we are going to report below. Stekloff problem shares many features with Neumann problem. In fact, $\tilde{\lambda}_1 = 0$ is the unique principal eigenvalue since eigenfunctions $u \in W^{1,p}(\Omega)$ associated to eigenvalues $\lambda \neq 0$ must satisfy the average condition on the boundary, $\int_{\partial\Omega} |u|^{p-2}u = 0$. It is also an *isolated* eigenvalue, while an infinite sequence of eigenvalues $0 = \tilde{\lambda}_1 = \tilde{\lambda}_1^{LS} < \tilde{\lambda}_2^{LS} \leq \cdots \leq \tilde{\lambda}_n^{LS} \leq \cdots$, $\tilde{\lambda}_n^{LS} \to \infty$, is furnished by,

$$\tilde{\lambda}_n^{LS} = \inf_{S \in \mathcal{S}_n} \max_{u \in S} \frac{\int_\Omega |\nabla u|^p}{\int_{\partial\Omega} |u|^p},$$

where $\mathcal{S}_n = \{S \in W^{1,p}(\Omega) \setminus W_0^{1,p}(\Omega) : S = -S,\ S \text{ compact},\ \gamma(S) \geq n\}$.

Therefore, a second eigenvalue $\tilde{\lambda}_2 > 0$ to (5.1) can be defined in the same vein as in Section 1. Moreover, it can be shown that

$$\lambda_2 = \tilde{\lambda}_2 = \inf_{S \in \mathcal{S}_2} \max_{u \in S} \frac{\int_\Omega |\nabla u|^p}{\int_{\partial\Omega} |u|^p}.$$

Thus we get a first variational representation of $\tilde{\lambda}_2$. Our main goal in this section is to state a variational characterization of $\tilde{\lambda}_2$ patterned on (1.5). Proof of next theorem is achieved by using the ideas of Section 3.

Theorem 5.1 *The weighted eigenvalue problem,*

$$\begin{cases} -\Delta_p u = 0, & x \in \Omega, \\ |\nabla u|^{p-2}\dfrac{\partial u}{\partial \nu} = \lambda m(x)|u|^{p-2}u, & x \in \partial\Omega, \end{cases}$$

where $m(x) > 0$ a. e. in $\partial\Omega$ and either $m \in L^\infty(\partial\Omega)$ or $m \in L^r(\partial\Omega)$ with $r > \left(\frac{p_*}{p}\right)' = \frac{N-1}{p-1}$ if $1 < p \le N$, $r = 1$ otherwise, possesses a second (first nontrivial) eigenvalue which can be expressed as,

$$\tilde{\lambda}_2 = \inf\left\{ \frac{\int_\Omega |\nabla u|^p}{\int_{\partial\Omega} |u|^p \, m \, dS} : \int_{\partial\Omega} |u|^{p-2} u \, m \, dS = 0 \right\}. \tag{5.2}$$

Remark 5.2 Problem (5.1) has deserved a considerable amount of effort in recent times, see for instance [9, 23], [31], [25], [27], [28] and [6]. However it should be stressed that in all of these works, either the equation or the boundary condition are perturbed by a zero order term of the form $a(x)|u|^{p-2}u$. Thus, the first eigenvalue is no more trivial, the average condition $\int_{\partial\Omega} |u|^{p-2}u = 0$ fails and accordingly, formula (5.2) ceases to be valid in these perturbed problems.

Remark 5.3 Stekloff problem can be formulated for the so–called pseudo p–Laplacian operator,

$$\begin{cases} -\mathrm{div}(\Phi_p(\nabla u)) = 0, & x \in \Omega, \\ \\ \langle \Phi_p(\nabla u), \nu \rangle = \lambda |u|^{p-2}u, & x \in \partial\Omega, \end{cases} \tag{5.3}$$

where ν is again the outward unit normal but now $\Phi_p(\xi)$ stands for the field $\Phi_p(\xi) = (|\xi_1|^{p-2}\xi_1, \ldots, |\xi_N|^{p-2}\xi_N)$. Under this format the corresponding version of formula (5.2) was obtained in an independent way in [11]. Nevertheless, our approach here can be also used to show the natural version of (5.2) for problem (5.3).

6. Closed Riemannian manifolds

Let (M, g) be a compact and connected N–dimensional Riemannian manifold endowed with a Riemannian metric g which induces the associated volume element v_g on M. The p–Laplacian operator $\Delta_{p,M}$ on M is defined as $\Delta_{p,M} u = \mathrm{div}_g(|\nabla u|^{p-2}\nabla u)$, $u \in W^{1,p}(M)$, where the divergence div_g and gradient ∇u operators are understood in the sense of the metric g. The eigenvalue problem,

$$-\Delta_{p,M} u = \lambda |u|^{p-2}u, \tag{6.1}$$

for the p–Laplacian on M consists in finding pairs $(\lambda, u) \in \mathbb{R} \times W^{1,p}(M)$, $u \neq 0$, such that,

$$\int_M |\nabla u|^{p-2}\nabla u \nabla v \, dv_g = \lambda \int_M |u|^{p-2}uv \, dv_g, \qquad v \in W^{1,p}(M).$$

By the reasons just argued in Section 1, $\lambda_{1,M} = 0$ is the unique principal eigenvalue to $-\Delta_{p,M}$ which is isolated. Thus, the second (first nontrivial) eigenvalue $\lambda_{2,M}$ is well defined. By employing the standard calculus apparatus in (M, g) together with Theorem 2.3 one can shows the next result.

Theorem 6.1 *The 'first' nontrivial eigenvalue $\lambda_{2,M}$ of $-\Delta_{p,M}$ in a compact, connected N–dimensional Riemannian manifold (M, g) without boundary is expressed by,*

$$\lambda_2 = \inf\left\{ \frac{\int_M |\nabla u|^p \, dv_g}{\int_M |u|^p \, dv_g} : \quad u \in W^{1,p}(M), \int_M |u|^{p-2}u \, dv_g = 0 \right\}.$$

Remark 6.2 Theorem 6.1 was proved in [36] by an approximation argument. In fact, it is not achieved there that any minimizer $\hat{u}$ to the Rayleigh quotient is an eigenfunction. However, our approach here shows that all possible minimizers actually define eigenfunctions to (6.1).

7. A zero average flux problem

There is still available one more example falling in the scope of the ideas in Section 3. Namely,

$$\begin{cases} -\Delta_p u = \lambda |u|^{p-2}u, & x \in \Omega, \\ \quad u = c, & x \in \partial\Omega, \\ \\ \int_{\partial\Omega} |\nabla u|^{p-2} \dfrac{\partial u}{\partial \nu} = 0, \end{cases} \tag{7.1}$$

where c has the status of a free constant, not 'a priori' determined. In the case $p = 2$, problem (7.1) arises from the study of plasma equilibrium configurations ([8]). One can check that (7.1) is nothing else but the N–dimensional

version of the classical periodic problem ($N = 1$). Since boundary value c is an unknown the energy space for this problem is $X = W_0^{1,p}(\Omega) \oplus \mathbb{R}$. Thus, an eigen pair $(\lambda, u) \in \mathbb{R} \times X$ is defined through the equality,

$$\int_\Omega |\nabla u|^{p-2}\nabla u \nabla v = \lambda \int_\Omega |u|^{p-2}uv, \qquad \text{for all } v \in X.$$

Problem (7.1) exhibits $\hat{\lambda}_1 = 0$ as the only principal eigenvalue which is simple and isolated. Ljusternik–Schnirelmann theory provides us with an increasing sequence $\hat{\lambda}_n^{SL} \to \infty$ of eigenvalues so that $\hat{\lambda}_2^{SL}$ matches the 'second eigenvalue' $\hat{\lambda}_2$ (see the explanation in Section 1). Accordingly,

$$\hat{\lambda}_2 = \inf_{A \in \widehat{\mathcal{S}}_2} \sup_{u \in A} \frac{\int_\Omega |\nabla u|^p}{\int_\Omega |u|^p},$$

where $\widehat{\mathcal{S}}_2 = \{S \in X : S = -S,\ S \text{ compact},\ \gamma(S) \geq 2\}$. Next result proposes a more simpler expression in the line of previous cases.

Theorem 7.1 *First non trivial eigenvalue $\hat{\lambda}_2$ to* (7.1) *can be variationally expressed as,*

$$\hat{\lambda}_2 = \inf \left\{ \int_\Omega |\nabla u|^p : u \in X,\ \int_\Omega |u|^p = 1,\ \int_\Omega |u|^{p-2}u = 0 \right\}.$$

8. The p–Laplacian on graphs

A n–order graph is defined as a pair $\mathcal{G} = (\mathcal{V}, E)$ where $\mathcal{V} = \{v_1, \ldots, v_n\}$ are the 'vertices' of $\mathcal{G}$ and E is a distinguished family $E = \{e = \{u, v\} \subset \mathcal{V}\}$ of two–points sets of $\mathcal{V}$ called the 'edges' of $\mathcal{G}$. In fact, a vertix u is said to be connected (or 'adjacent') to another v provided that $e = \{u, v\} \in E$ ([15]).

A convenient way of *both* defining the edges and measuring the connectedness degree between vertices is by introducing a symmetric matrix of nonnegative weights $A = (\omega_{ij})$, $\omega_{ij} \geq 0$, $\omega_{ij} = \omega_{ji}$, $\omega_{ii} = 0$. This is the 'connectivity' or 'similarity' matrix. Under this convention $\{v_i, v_j\}$ defines an edge provided that $\omega_{ij} > 0$. A graph $\mathcal{G}$ is said to be *connected* if every couple $u, v \in \mathcal{V}$ can be related by means a family of edges $\{u, v_1\},\ \{v_1, v_2\}, \ldots, \{v_{i_{m-1}}, v\}$. Equivalently,

$$\omega_{\{u,v_1\}}\omega_{\{v_1,v_2\}} \cdots \omega_{\{v_{i_{m-1}},v\}} > 0.$$

The space $\mathcal{H}$ of functions $f : \mathcal{V} \to \mathbb{R}$ on a graph $\mathcal{G}$ consists of vectors $f \in \mathbb{R}^n$, so that $f = (f_i)$ where $f_i = f(v_i) =: f(i)$. On the other hand, the Dirichlet's form,

$$\mathcal{D}(f) = \frac{1}{2}\sum_{i,j} \omega_{ij}|f_i - f_j|^p,$$

is the equivalent to the Dirichlet integral $\mathcal{J}(u) := \int_\Omega |\nabla u|^p$ in the continuous case. Remark that the p–Laplacian is nothing else but $-\Delta_p u = \frac{1}{p} D\mathcal{J}(u)$. Thus we can already introducing both the definition of the p–Laplacian operator together with its eigenvalues ([2], [13, 14]). Please, kindly notice that a minus sign has been incorporated to the definition of the p–Laplacian so as to keep the analogies with partial differential operators.

Definition 8.1 The p–Laplacian operator $-\Delta_p : \mathcal{H} \to \mathcal{H}$ in $\mathcal{G}$ is defined as $-\Delta_p = \nabla \mathcal{D}_p$. In other words,

$$-\Delta_p(f)(i) = \sum_j \omega_{ij}|f_i - f_j|^{p-2}(f_i - f_j).$$

A (weighted) eigen pair (λ, f) to $-\Delta_p$ is defined as,

$$-\Delta_p(f)(i) = \lambda v_i |f_i|^{p-2}(f_i), \qquad f \neq 0,$$

where $v_i > 0$, $1 \leq i \leq n$, is a given set of weights.

The eigenvalues of $-\Delta_p$ are just the critical values of the Dirichlet form $\mathcal{D}(f)$ in the unit sphere $\mathcal{M}_1 = \{\sum_i v_i |f_i|^p = 1\}$. In other words, the critical values of the Rayleigh quotient,

$$\mathcal{Q}(f) = \frac{1}{2}\frac{\sum_{i,j} \omega_{ij}|f_i - f_j|^p}{\sum_i v_i |f_i|^p}.$$

On the other hand, the spectrum of $-\Delta_p$ is a compact set in $[0, \infty)$, while $\lambda_1 = 0$ is a trivial eigenvalue which has constant functions as the only associated eigenfunctions ($\mathcal{G}$ is connected). In addition, eigenfunctions f to eigenvalues $\lambda > 0$ must 'change sign' since they satisfy the 'zero average' condition:

$$\sum_i \nu_i |f_i|^{p-2}(f_i) = 0.$$

This means $\lambda_1 = 0$ is isolated and a second eigenvalue $\lambda_2 > 0$ should exist.

Next statement furnishes the right version of the existence assertions on eigenvalues to $-\Delta_p$ introduced in [2], [13]. In fact, existence results in these references are only be correctly stated in the regime $p \geq 2$. An explanation of the critical case $1 < p < 2$ is missing in the mentioned works.

Theorem 8.2 *Let $\mathcal{G}$ be a connected graph. Then, the second eigenvalue λ_2 of $-\Delta_p$ is given by,*

$$\lambda_2 = \min_{\mathcal{M}_0} \frac{\mathcal{D}(f)}{\sum_i \nu_i |f_i|^p},$$

where

$$\mathcal{M}_0 = \{f : \sum_i \nu_i |f_i|^{p-2}(f_i) = 0\}.$$

Similarly, the maximum eigenvalue is,

$$\lambda^* = \max_{\mathcal{M}_0} \frac{\mathcal{D}(f)}{\sum_i \nu_i |f_i|^p}.$$

Remark 8.3 Nodal regions associated to the eigenfunctions to λ_2 play a prominent rôle in spectral clustering of graphs. Reader is referred to [15] and [13] for a deeper account on this important subject.

Acknowledgements

This research has been partially supported by CIUCSD (Generalitat Valenciana) under project AICO/2021/223.

References

[1] Herbert Amann. Lusternik-Schnirelman theory and non-linear eigenvalue problems. *Math. Ann.*, 199:55–72, 1972.

[2] S. Amghibech. Eigenvalues of the discrete p-Laplacian for graphs. *Ars Combin.*, 67:283–302, 2003.

[3] A. Anane and N. Tsouli. On the second eigenvalue of the p-Laplacian. In *Nonlinear partial differential equations (Fès, 1994)*, volume 343 of *Pitman Res. Notes Math. Ser.*, pages 1–9. Longman, Harlow, 1996.

[4] Aomar Anane. *Etude des valeurs propres et de la résonance pour l'opérateur p–Laplacien*. Thése de doctorat. Université Libre de Bruxelles, 1987.

[5] Aomar Anane. Simplicité et isolation de la première valeur propre du p-laplacien avec poids. *C. R. Acad. Sci. Paris Sér. I Math.*, 305(16):725–728, 1987.

[6] Aomar Anane, Omar Chakrone, Belhadj Karim, and Abdellah Zerouali. Eigencurves for a Steklov problem. *Electron. J. Differential Equations*, pages No. 75, 8, 2009.

[7] M. Arias, J. Campos, M. Cuesta, and J.-P. Gossez. An asymmetric Neumann problem with weights. *Ann. Inst. H. Poincaré C Anal. Non Linéaire*, 25(2):267–280, 2008.

[8] Henri Berestycki and Haïm Brézis. On a free boundary problem arising in plasma physics. *Nonlinear Anal.*, 4(3):415–436, 1980.

[9] Julián Fernández Bonder and Julio D. Rossi. Existence results for the p-Laplacian with nonlinear boundary conditions. *J. Math. Anal. Appl.*, 263(1):195–223, 2001.

[10] Barbara Brandolini, Francesco Chiacchio, and Cristina Trombetti. Optimal lower bounds for eigenvalues of linear and nonlinear Neumann problems. *Proc. Roy. Soc. Edinburgh Sect. A*, 145(1):31–45, 2015.

[11] Lorenzo Brasco and Giovanni Franzina. An anisotropic eigenvalue problem of Stekloff type and weighted Wulff inequalities. *NoDEA Nonlinear Differential Equations Appl.*, 20(6):1795–1830, 2013.

[12] Lorenzo Brasco, Carlo Nitsch, and Cristina Trombetti. An inequality à la Szegö–Weinberger for the p-Laplacian on convex sets. *Commun. Contemp. Math.*, 18(6):1550086, 23, 2016.

[13] T. Bühler and M. Hein. Spectral clustering based on the graph p–laplacian. In *L. Bottou and M. Littman (eds.), Proc. of the* 26^{th} *Int. Conf. Mach. Learn. (ICML), Canada*, pages 81–88. Association for Computing Machinery, New York, United States, 2009.

[14] T. Bühler and M. Hein. Supplementary material. *http://www.ml.uni-saarland.de/Publications/BueHei09tech.pdf*, 2009.

[15] Fan R. K. Chung. *Spectral graph theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, Providence, RI, 1997.

[16] Mabel Cuesta, Djairo G. De Figueiredo, and Jean-Pierre Gossez. A nodal domain property for the p-Laplacian. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(8):669–673, 2000.

[17] B. Dacorogna, W. Gangbo, and N. Subía. Sur une généralisation de l'inégalité de Wirtinger. *Ann. Inst. H. Poincaré C Anal. Non Linéaire*, 9(1):29–50, 1992.

[18] P. Drábek. *Solvability and bifurcations of nonlinear equations*, volume 264 of *Pitman Research Notes in Mathematics Series*. Longman Scientific & Technical, Harlow; copublished in the United States with John Wiley & Sons, Inc., New York, 1992.

[19] Pavel Drábek, Pavel Krejčí, and Peter Takáč. *Nonlinear differential equations*, volume 404 of *Chapman & Hall/CRC Research Notes in Mathematics*. Chapman & Hall/CRC, Boca Raton, FL, 1999. Papers from the Seminar on Differential Equations held in Chvalatice, June 29–July 3, 1998.

[20] Pavel Drábek and Raúl Manásevich. On the closed solution to some nonhomogeneous eigenvalue problems with p-Laplacian. *Differential Integral Equations*, 12(6):773–788, 1999.

[21] Pavel Drábek and Stephen B. Robinson. Resonance problems for the p-Laplacian. *J. Funct. Anal.*, 169(1):189–200, 1999.

[22] Pavel Drábek and Stephen B. Robinson. On the generalization of the Courant nodal domain theorem. *J. Differential Equations*, 181(1):58–71, 2002.

[23] Julián Fernández Bonder and Julio D. Rossi. A nonlinear eigenvalue problem with indefinite weights related to the Sobolev trace embedding. *Publ. Mat.*, 46(1):221–235, 2002.

[24] J. P. García Azorero and I. Peral Alonso. Existence and nonuniqueness for the p-Laplacian: nonlinear eigenvalues. *Comm. Partial Differential Equations*, 12(12):1389–1430, 1987.

[25] Jorge García-Melián, Julio D. Rossi, and José C. Sabina de Lis. Multiplicity of solutions to a nonlinear elliptic problem with nonlinear boundary conditions. *NoDEA Nonlinear Differential Equations Appl.*, 21(3):305–337, 2014.

[26] Leszek Gasiński and Nikolaos S. Papageorgiou. *Nonlinear analysis*, volume 9 of *Series in Mathematical Analysis and Applications*. Chapman & Hall/CRC, Boca Raton, FL, 2006.

[27] An Lê. Eigenvalue problems for the p-Laplacian. *Nonlinear Anal.*, 64(5):1057–1099, 2006.

[28] Liamidi Leadi and Aboubacar Marcos. A weighted eigencurve for Steklov problems with a potential. *NoDEA Nonlinear Differential Equations Appl.*, 20(3):687–713, 2013.

[29] Peter Lindqvist. On the equation div $(|\nabla u|^{p-2}\nabla u) + \lambda |u|^{p-2}u = 0$. *Proc. Amer. Math. Soc.*, 109(1):157–164, 1990.

[30] Peter Lindqvist. A nonlinear eigenvalue problem. In *Topics in Mathematical Analysis*, volume 3 of *Ser. Anal. Appl. Comput.*, pages 175–203. World Sci. Publ., Hackensack, NJ, 2008.

[31] Sandra Martínez and Julio D. Rossi. Isolation and simplicity for the first eigenvalue of the p-Laplacian with a nonlinear boundary condition. *Abstr. Appl. Anal.*, 7(5):287–293, 2002.

[32] José C. Sabina de Lis. Remarks on the second Neumann eigenvalue. *Electron. J. Differential Equations*, pages Paper No. 13, 12, 2022.

[33] Michael Struwe. *Variational methods*. Modern Surveys in Mathematics. Springer-Verlag, Berlin, fourth edition, 2008.

[34] Peter Takáč. On the number and structure of solutions for a Fredholm alternative with the p-Laplacian. *J. Differential Equations*, 185(1):306–347, 2002.

[35] Peter Takáč. Nonlinear spectral problems for degenerate elliptic operators. In *Stationary partial differential equations. Vol. I*, Handb. Differ. Equ., pages 385–489. North-Holland, Amsterdam, 2004.

[36] Laurent Véron. Première valeur propre non nulle du p-laplacien et équations quasi linéaires elliptiques sur une variété riemannienne compacte. *C. R. Acad. Sci. Paris Sér. I Math.*, 314(4):271–276, 1992.

[37] E. Zeidler. The Ljusternik-Schnirelman theory for indefinite and not necessarily odd nonlinear operators and its applications. *Nonlinear Anal.*, 4(3):451–489, 1980.

The existence of well-balanced entropy stable numerical scheme for the Ripa model with the topography source term

Ludovic Martaud[1], **Christophe Berthon**[2]

1. *Laboratoire de Mathématiques Jean Leray, CNRS UMR 6629, Nantes Unversité, 2 rue de la Houssinière, BP 92208, 44322 Nantes, France, Email address:* ludovic.martaud@univ – nantes.fr
2. *Laboratoire de Mathématiques Jean Leray, CNRS UMR 6629, Nantes Unversité, 2 rue de la Houssinière, BP 92208, 44322 Nantes, France, Email address:* christophe.berthon@univ – nantes.fr

Abstract

In this work, we present a Godunov type scheme to approximate the solution of the Ripa model. The main objective is to design a scheme far from dry areas that verifies a fully discrete entropy inequality and preserves the lake at rest equilibrium. Some numerical experiments are carried out to illustrate the scheme relevance.

1. Introduction

This work deals with the numerical approximation of the weak solutions of the Ripa model with topography source term in one space dimension given by the following Cauchy problem:

$$\begin{aligned} &\partial_t \underbrace{\begin{pmatrix} h \\ hu \\ h\theta \end{pmatrix}}_{:=w} + \partial_x \underbrace{\begin{pmatrix} hu \\ hu^2 + \frac{g\varphi(\theta)}{2}h^2 \\ h\theta u \end{pmatrix}}_{:=f(w)} = \underbrace{\begin{pmatrix} 0 \\ -gh\varphi(\theta)\partial_x z \\ 0 \end{pmatrix}}_{:=S(w)}, \quad x \in \mathbb{R},\ t > 0, \\ &w(x, t=0) = w_0(x), \quad x \in \mathbb{R}, \end{aligned} \tag{1.1}$$

where $\varphi : \mathbb{R}_*^+ \to \mathbb{R}_*^+$ is a smooth invertible function. This model governs the water height $h > 0$ and the velocity $u \in \mathbb{R}$ of a fluid. The gravitation constant is $g > 0$ and $z : \mathbb{R} \to \mathbb{R}^+$ is a given time independent positive smooth topography function. The temperature of the fluid is $\varphi(\theta) > 0$. After [3], the system (1.1) is equivalent to the usual Ripa model in which $\varphi(\theta) = \theta$. The fluid is assumed to be far from the dry regions and the unknown state vector w is assumed to be in the convex set $\Omega = \left\{(h, hu, h\theta) \in \mathbb{R}^3 \mid h > 0, hu \in \mathbb{R}, h\theta > 0\right\}$. Finally, $w_0 : \mathbb{R} \to \Omega$ is a given measurable function of $L^1_{\text{loc}}(\mathbb{R})$. According to [3], if the function φ verifies

$$\varphi''(\theta)\varphi(\theta) - \varphi'(\theta)^2/2 > 0, \quad \varphi(\theta) - \theta\varphi'(\theta) + \theta^2\varphi''(\theta)/2 > 0, \quad \forall \theta \in \mathbb{R}_*^+, \tag{1.2}$$

then the system (1.1) can be endowed with a pair entropy-entropy flux (η, G) given by

$$\eta(w) = hu^2/2 + g\varphi(\theta)h^2/2, \quad G(w) = hu^3/2 + g\varphi(\theta)h^2u, \quad \forall w \in \Omega, \tag{1.3}$$

where $w \mapsto \eta(w)$ is a strictly convex function. As a consequence, the solutions of (1.1) has to satisfy in addition the following entropy inequality:

$$\partial_t\Big(\eta(w) + \psi(w)z\Big) + \partial_x\Big(G(w) + H(w)z\Big) \leq 0, \tag{1.4}$$

which is related to the physical energy and where we have set $\psi(w) := gh\varphi(\theta)$ and $H(w) := gh\varphi(\theta)u$. In addition, the presence of the source term $S(w)$ involves the existence of non-trivial stationary solutions verifying $\partial_x f(w) = S(w)$. Among these steady states, the lake at rest given by $u = 0$, $\theta = \text{cst}$, $h + z = \text{cst}$ is here of main importance.

From a numerical point of view, we consider uniform meshes in space $(x_{i+\frac{1}{2}})_{i\in\mathbb{Z}}$ in $\mathbb{R}$ of constant size $\Delta x > 0$. Thus, we have $x_{i+\frac{1}{2}} = x_{i-\frac{1}{2}} + \Delta x$ for all i in $\mathbb{Z}$. We also consider uniform meshes in time $(t^n)_{n\in\mathbb{N}}$ in $[0, +\infty)$ of constant size $\Delta t > 0$, and we have $t^{n+1} = t^n + \Delta t$ for all n in $\mathbb{N}$. On each cell $(x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$, the topography is discretized according to $z_i = \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} z(x)\,\mathrm{d}x$ for all i in $\mathbb{Z}$. At time t^n and on each cells the weak solutions of (1.1) are approximated with constant states w_i^n such that $w_i^n \approx \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} w(x, t^n)\,\mathrm{d}x$ for all i in $\mathbb{Z}$.

From a sequence $(w_i^n)_{i\in\mathbb{Z}}$ many strategies are well known to evaluate the updated sequence $(w_i^{n+1})_{i\in\mathbb{Z}}$ at time t^{n+1} (see for instance [2–4]). Here, a suitable scheme has to be well-balanced for the lake at rest solution. This property writes at a discrete level for all i in $\mathbb{Z}$

$$u_i^n = 0, \quad \theta_i^n = \text{cst}, \quad h_i^n + z_i = \text{cst}, \quad \text{then} \quad w_i^{n+1} = w_i^n. \tag{1.5}$$

A scheme has to verify, in addition, a discrete version of the entropy inequality (1.4) that reads

$$\eta(w_i^{n+1}) + g h_i^{n+1}\varphi(\theta_i^{n+1}) z_i \le \eta(w_i^n) + g h_i^n \varphi(\theta_i^n) z_i - \frac{\Delta t}{\Delta x}\big(\tilde{\mathcal{G}}_{i+\frac12} - \tilde{\mathcal{G}}_{i-\frac12}\big), \quad \forall i \in \mathbb{Z}, \tag{1.6}$$

where $\tilde{\mathcal{G}}_{i+\frac12}$ is a consistent approximation of $G + Hz$. In this regard, the authors of [3] introduced a well-balanced scheme satisfying a discrete entropy stability perturbed by an error therm of the form $O(\Delta x^2)$ while in [6] an entropy inequality is reached in the flat regions.

In this work, we consider schemes written under the from of a Godunov type scheme [4] that read

$$w_i^{n+1} = \frac{1}{\Delta x}\int_{x_{i-\frac12}}^{x_i} \tilde{w}\big((x - x_{i-\frac12})/\Delta t, w_{i-1}^n, w_i^n\big)\,\mathrm{d}x + \frac{1}{\Delta x}\int_{x_i}^{x_{i+\frac12}} \tilde{w}\big((x - x_{i+\frac12})/\Delta t, w_i^n, w_{i+1}^n\big)\,\mathrm{d}x, \quad \forall i \in \mathbb{Z}, \tag{1.7}$$

where $(\tilde{w}(\cdot, w_i^n, w_{i+1}^n))_{i\in\mathbb{Z}}$ is a set of juxtaposed approximated Riemann solvers without interaction. This non interaction is ensured by a restriction on the time step Δt also called the non interaction CFL condition. In the sequel, we propose to design a scheme satisfying a fully discrete entropy inequality (1.6) and well-balanced for the steady state at rest (1.5) in the wet regions.

2. A well-balanced entropy stable numerical scheme for the Ripa model

First, considering an interface between two constants states w_i^n and w_{i+1}^n, we establish an integral consistency relation and an entropy stability sufficient condition to be stated on the approximated Riemann solver $\tilde{w}(\cdot, w_{i+1}^n, w_i^n)$.

Lemma 2.1 *Consider the Riemann problem associated to the Ripa equations given by the system* (1.1) *with the following initial condition*

$$\big(w(x, t=0), z\big) = \begin{cases} (w_i^n, z_i) & \text{if} \quad x < x_{i+\frac12}, \\ (w_{i+1}^n, z_{i+1}) & \text{otherwise.} \end{cases} \tag{2.1}$$

Consider $\tilde{w}(\cdot, w_i^n, w_{i+1}^n)$, a Riemann solver that approximates the solutions of the above Riemann problem. Assume that a non interaction CFL condition holds. If

$$\frac{1}{\Delta x}\int_{x_i}^{x_{i+1}} \tilde{w}\big((x - x_{i+\frac12})/\Delta t, w_i^n, w_{i+1}^n\big)\,\mathrm{d}x = \frac{w_i^n + w_{i+1}^n}{2} - \frac{\Delta t}{\Delta x}\big(f(w_{i+1}^n) - f(w_i^n)\big) + \Delta t S_{i+\frac12}^n, \tag{2.2}$$

where $S_{i+\frac12}^n$ is a consistent discretization of $\frac{1}{\Delta x}\int_{x_i}^{x_{i+1}} S\big(w(x, t^n)\big)\,\mathrm{d}x$, then the Godunov type scheme (1.7) *is consistent with* (1.1) *and conservative on flat topography* ($z = \text{cste}$). *Moreover, if $\varphi''(\theta) > 0$ for all θ in $\mathbb{R}_*^+$ and if*

$$\frac{1}{\Delta x}\int_{x_i}^{x_{i+1}} \psi\big(\tilde{w}\big((x - x_{i+\frac12})/\Delta t, w_i^n, w_{i+1}^n\big)\big)\,\mathrm{d}x \le \frac{\psi(w_i^n) + \psi(w_{i+1}^n)}{2} - \frac{\Delta t}{\Delta x}\big(H(w_{i+1}^n) - H(w_i^n)\big), \tag{2.3}$$

$$\frac{1}{\Delta x}\int_{x_i}^{x_{i+1}} \eta\big(\tilde{w}\big((x - x_{i+\frac12})/\Delta t, w_i^n, w_{i+1}^n\big)\big)\,\mathrm{d}x \le \frac{\eta(w_i^n) + \eta(w_{i+1}^n)}{2} - \frac{\Delta t}{\Delta x}\big(G(w_{i+1}^n) - G(w_i^n)\big) - \frac{\Delta t}{\Delta x}\mathcal{H}_{i+\frac12}(z_{i+1} - z_i), \tag{2.4}$$

where

$$\begin{aligned} \mathcal{H}_{i+\frac12} \;:=\; & \frac{H(w_i^n) + H(w_{i+1}^n)}{2} - \frac{\Delta x}{4\Delta t}\big(\psi(w_{i+1}^n) - \psi(w_i^n)\big) + \frac{1}{2\Delta t}\int_{x_{i+\frac12}}^{x_{i+1}} \psi\big(\tilde{w}\big((x - x_{i+\frac12})/\Delta t, w_i^n, w_{i+1}^n\big)\big)\,\mathrm{d}x \\ & - \frac{1}{2\Delta t}\int_{x_i}^{x_{i+\frac12}} \psi\big(\tilde{w}\big((x - x_{i+\frac12})/\Delta t, w_i^n, w_{i+1}^n\big)\big)\,\mathrm{d}x, \end{aligned} \tag{2.5}$$

then the Godunov type scheme (1.7) *associated to the Riemann solver $\tilde{w}(\cdot,\cdot,\cdot)$ satisfies a fully discrete entropy inequality* (1.6).

Proof The condition (2.2) is standard and we refer to [1] for the proof of both consistency and conservation on flat topography. Here, we focus on the establishment of the discrete entropy stability. Since $\theta \mapsto \varphi(\theta)$ is strictly convex, it is easy to show that $w \mapsto \psi(w)$ is also convex. Hence, if we apply ψ to the Godunov type scheme (1.7), the Jensen inequality yields the following inequality:

$$\begin{aligned}
\psi(w_i^{n+1}) \quad \leq \quad & \frac{1}{2\Delta x}\int_{x_i}^{x_{i+\frac{1}{2}}} \psi\left(\tilde{w}((x-x_{i+\frac{1}{2}})/\Delta t, w_i^n, w_{i+1}^n)\right)\mathrm{d}x + \frac{1}{2\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_i} \psi\left(\tilde{w}((x-x_{i-\frac{1}{2}})/\Delta t, w_{i-1}^n, w_i^n)\right)\mathrm{d}x \\
+ \quad & \frac{1}{2\Delta x}\int_{x_{i-1}}^{x_i} \psi\left(\tilde{w}((x-x_{i-\frac{1}{2}})/\Delta t, w_{i-1}^n, w_i^n)\right)\mathrm{d}x + \frac{1}{2\Delta x}\int_{x_i}^{x_{i+1}} \psi\left(\tilde{w}((x-x_{i+\frac{1}{2}})/\Delta t, w_i^n, w_{i+1}^n)\right)\mathrm{d}x \\
- \quad & \frac{1}{2\Delta x}\int_{x_{i-1}}^{x_{i-\frac{1}{2}}} \psi\left(\tilde{w}((x-x_{i-\frac{1}{2}})/\Delta t, w_{i-1}^n, w_i^n)\right)\mathrm{d}x - \frac{1}{2\Delta x}\int_{x_{i+\frac{1}{2}}}^{x_{i+1}} \psi\left(\tilde{w}((x-x_{i+\frac{1}{2}})/\Delta t, w_i^n, w_{i+1}^n)\right)\mathrm{d}x.
\end{aligned}$$

Now, we use the condition of (2.3) to obtain

$$\begin{aligned}
\psi(w_i^{n+1}) \leq & \frac{1}{2\Delta x}\int_{x_i}^{x_{i+\frac{1}{2}}} \psi\left(\tilde{w}((x-x_{i+\frac{1}{2}})/\Delta t, w_i^n, w_{i+1}^n)\right)\mathrm{d}x + \frac{1}{2\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_i} \psi\left(\tilde{w}((x-x_{i-\frac{1}{2}})/\Delta t, w_{i-1}^n, w_i^n)\right)\mathrm{d}x \\
& + \frac{1}{2}\left(\frac{\psi(w_{i-1}^n)+\psi(w_i^n)}{2} - \frac{\Delta t}{\Delta x}\left(H(w_i^n) - H(w_{i-1}^n)\right)\right) + \frac{1}{2}\left(\frac{\psi(w_i^n)+\psi(w_{i+1}^n)}{2} - \frac{\Delta t}{\Delta x}\left(H(w_{i+1}^n) - H(w_i^n)\right)\right) \\
& - \frac{1}{2\Delta x}\int_{x_{i-1}}^{x_{i-\frac{1}{2}}} \psi\left(\tilde{w}((x-x_{i-\frac{1}{2}})/\Delta t, w_{i-1}^n, w_i^n)\right)\mathrm{d}x - \frac{1}{2\Delta x}\int_{x_{i+\frac{1}{2}}}^{x_{i+1}} \psi\left(\tilde{w}((x-x_{i+\frac{1}{2}})/\Delta t, w_i^n, w_{i+1}^n)\right)\mathrm{d}x.
\end{aligned}$$

Using the definition of $\mathcal{H}$ given by (2.5), the above inequality rewrites

$$\psi(w_i^{n+1}) \leq \psi(w_i^n) - \frac{\Delta t}{\Delta x}\left(\mathcal{H}_{i+\frac{1}{2}} - \mathcal{H}_{i-\frac{1}{2}}\right). \tag{2.6}$$

On the other hand, with η convex, using the Jensen inequality and the condition (2.4) the following estimates holds:

$$\begin{aligned}
\eta(w_i^{n+1}) \quad \leq \quad & \frac{1}{2\Delta x}\int_{x_i}^{x_{i+\frac{1}{2}}} \eta\left(\tilde{w}((x-x_{i+\frac{1}{2}})/\Delta t, w_i^n, w_{i+1}^n)\right)\mathrm{d}x + \frac{1}{2\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_i} \eta\left(\tilde{w}((x-x_{i-\frac{1}{2}})/\Delta t, w_{i-1}^n, w_i^n)\right)\mathrm{d}x \\
+ \quad & \frac{1}{2}\left(\frac{\eta(w_{i-1}^n)+\eta(w_i^n)}{2} - \frac{\Delta t}{\Delta x}\left(G(w_i^n) - G(w_{i-1}^n)\right) - \frac{\Delta t}{\Delta x}\mathcal{H}_{i-\frac{1}{2}}\left(z_i - z_{i-1}\right)\right) \\
+ \quad & \frac{1}{2}\left(\frac{\eta(w_i^n)+\eta(w_{i+1}^n)}{2} - \frac{\Delta t}{\Delta x}\left(G(w_{i+1}^n) - G(w_i^n)\right) - \frac{\Delta t}{\Delta x}\mathcal{H}_{i+\frac{1}{2}}\left(z_{i+1} - z_i\right)\right) \\
- \quad & \frac{1}{2\Delta x}\int_{x_{i-1}}^{x_{i-\frac{1}{2}}} \eta\left(\tilde{w}((x-x_{i-\frac{1}{2}})/\Delta t, w_{i-1}^n, w_i^n)\right)\mathrm{d}x - \frac{1}{2\Delta x}\int_{x_{i+\frac{1}{2}}}^{x_{i+1}} \eta\left(\tilde{w}((x-x_{i+\frac{1}{2}})/\Delta t, w_i^n, w_{i+1}^n)\right)\mathrm{d}x.
\end{aligned} \tag{2.7}$$

Since the topography z is positive, (2.6)$\times z_i$+(2.7) yields the fully discrete entropy inequality (1.6) with the following numerical entropy flux:

$$\begin{aligned}
\tilde{\mathcal{G}}_{i+\frac{1}{2}} \quad = \quad & \frac{G(w_{i+1}^n)+G(w_i^n)}{2} - \frac{\Delta x}{4\Delta t}\left(\eta(w_{i+1}^n) - \eta(w_i^n)\right) + \frac{1}{2\Delta t}\int_{x_{i+\frac{1}{2}}}^{x_{i+1}} \eta\left(\tilde{w}((x-x_{i+\frac{1}{2}})/\Delta t, w_i^n, w_{i+1}^n)\right)\mathrm{d}x \\
- \quad & \frac{1}{2\Delta t}\int_{x_i}^{x_{i+\frac{1}{2}}} \eta\left(\tilde{w}((x-x_{i+\frac{1}{2}})/\Delta t, w_i^n, w_{i+1}^n)\right)\mathrm{d}x + \mathcal{H}_{i+\frac{1}{2}}\frac{z_{i+1}+z_i}{2},
\end{aligned} \tag{2.8}$$

that achieves the proof. □

In Lemma 2.1, we require the strict convexity of the function $\theta \mapsto \varphi(\theta)$. Such functions that verify in addition the conditions (1.2) exist: for instance $\varphi(\theta) = \varepsilon e^{\theta}$ with $\varepsilon > 0$ an arbitrary constant or $\varphi(\theta) = \theta^3$ are possible choices.

According to Lemma 2.1, it is clear that an approximated Riemann solver $\tilde{w}(\cdot, w_{i+1}^n, w_i^n)$ is entirely locally defined around an interface $x_{i+\frac{1}{2}}$ having two constant states w_i^n, w_{i+1}^n on its either sides. For the sake of clarity in the notations, we now adopt an interface located in $x = 0$ and two constant states w_L, w_R respectively being on the

left, right side of this interface. We adopt a symmetric Riemann solver made of two intermediate states separated with a stationary wave. A such solver writes

$$\tilde{w}(\frac{x}{\Delta t}, w_L, w_R) = \begin{cases} w_L & \text{if} \quad \dfrac{x}{\Delta t} \leq -\lambda, \\ w_L^* & \text{if} \quad -\lambda < \dfrac{x}{\Delta t} \leq 0, \\ w_R^* & \text{if} \quad 0 < \dfrac{x}{\Delta t} \leq \lambda, \\ w_R & \text{if} \quad \lambda < \dfrac{x}{\Delta t}, \end{cases} \tag{2.9}$$

where $\lambda > 0$ stands for the artificial numerical diffusion and the two intermediate constant states $(w_\alpha^*)_{\alpha\in\{L,R\}} = (h_\alpha^*, h_\alpha^* u_\alpha^*, h_\alpha^* \theta_\alpha^*)^{\mathrm{T}}_{\alpha\in\{L,R\}}$ have to be characterized. Our objective is to choice the intermediate states w_L^* and w_R^* such that the conditions (2.2)-(2.3)-(2.4) hold. In this regard, considering a solver (2.9), the quantity $\mathcal{H}_{LR}$ given by (2.5) now reads

$$\mathcal{H}_{LR} = \frac{H(w_L) + H(w_R)}{2} - \frac{\lambda}{2}\big(\psi(w_R) - \psi(w_L)\big) + \frac{\lambda}{2}\big(\psi(w_R^*) - \psi(w_L^*)\big),$$

and assuming a non interaction CFL condition holds, the conditions (2.2)-(2.3)-(2.4) now write

$$\frac{w_L^* + w_R^*}{2} = w^{\mathrm{HLL}} + (0, \Delta x s_{LR}/(2\lambda), 0)^{\mathrm{T}}, \tag{2.10a}$$

$$\frac{\psi(w_L^*) + \psi(w_R^*)}{2} \leq \psi^{\mathrm{HLL}}, \tag{2.10b}$$

$$\frac{\eta(w_L^*) + \eta(w_R^*)}{2} + \frac{z_R - z_L}{4}\big(\psi(w_R^*) - \psi(w_L^*)\big) \leq \eta^{\mathrm{HLL}} + (\psi z)^{\mathrm{HLL}} - \psi^{\mathrm{HLL}}\frac{z_L + z_R}{2}, \tag{2.10c}$$

where s_{LR} is a consistent finite volume approximation of $\frac{1}{\Delta x}\int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} (-gh\varphi(\theta)\partial_x z)\,\mathrm{d}x$, and where we have defined the following notations:

$$\begin{aligned} w^{\mathrm{HLL}} := (h^{\mathrm{HLL}},\, (hu)^{\mathrm{HLL}},\, (h\theta)^{\mathrm{HLL}})^{\mathrm{T}} &= \frac{w_L + w_R}{2} - \frac{f(w_R) - f(w_L)}{2\lambda}, \\ \eta^{\mathrm{HLL}} &:= \frac{\eta(w_L) + \eta(w_R)}{2} - \frac{G(w_R) - G(w_L)}{2\lambda}, \\ \psi^{\mathrm{HLL}} &:= \frac{\psi(w_L) + \psi(w_R)}{2} - \frac{H(w_R) - H(w_L)}{2\lambda}, \\ (\psi z)^{\mathrm{HLL}} &:= \frac{\psi(w_L)z_L + \psi(w_R)z_R}{2} - \frac{H(w_R)z_R - H(w_L)z_L}{2\lambda}. \end{aligned} \tag{2.11}$$

In order to satisfy the relations (2.10), we propose to govern the intermediate states $(w_\alpha^*)_{\alpha\in\{L,R\}}$ by the following equations:

$$\frac{h_L^* + h_R^*}{2} = h^{\mathrm{HLL}}, \quad u_L^* = u_R^* \tag{2.12a}$$

$$\frac{h_L^* u_L^* + h_R^* u_R^*}{2} = (hu)^{\mathrm{HLL}} + \frac{\Delta x s_{LR}}{2\lambda}, \tag{2.12b}$$

$$\frac{h_L^* \theta_L^* + h_R^* \theta_R^*}{2} = h^{\mathrm{HLL}}\theta^{\mathrm{HLL}}, \tag{2.12c}$$

$$\frac{\psi(w_L^*) + \psi(w_R^*)}{2} = \psi(w^{\mathrm{HLL}})\Big(1 + \Big(\frac{c^{\mathrm{HLL}}\Delta t(\theta_R - \theta_L)}{(h\theta)^{\mathrm{HLL}}}\Big)^2\Big), \tag{2.12d}$$

$$h_R^* - h_L^* + \frac{h_L^* h_R^*}{h^{\mathrm{HLL}}}\frac{\varphi(\theta_R^*) - \varphi(\theta_L^*)}{\varphi(\theta^{\mathrm{HLL}})} = -\exp\Big(\frac{\lambda(\theta_R - \theta_L)^2}{c^{\mathrm{HLL}}(\theta^{\mathrm{HLL}})^2}\Big)\big(h_R^* - h_L^* + 2z_R - 2z_L\big), \tag{2.12e}$$

where we have introduced $\theta^{\mathrm{HLL}} := (h\theta)^{\mathrm{HLL}}/h^{\mathrm{HLL}} > 0$, $c^{\mathrm{HLL}} := \sqrt{gh^{\mathrm{HLL}}} > 0$ and $\psi(w^{\mathrm{HLL}}) := gh^{\mathrm{HLL}}\varphi(\theta^{\mathrm{HLL}}) > 0$. We will see in Theorem 2.2 that the above system endowed with a suitable choice of λ and Δt may ensure the required conditions (2.10).

The set of equations (2.12) is non linear and despite our efforts, we are not able to exhibit an explicit solution. Nevertheless, using (2.12a)-(2.12b), we deduce

$$u_L^* = u_R^* = \frac{1}{h^{\mathrm{HLL}}}\Big((hu)^{\mathrm{HLL}} + \Delta x s_{LR}/(2\lambda)\Big). \tag{2.13}$$

In the particular cases where $\theta_L = \theta_R$, the above definition of $u_L^* = u_R^*$ supplemented with

$$(h_R^*, h_L^*, \theta_L^*, \theta_R^*)^{\mathrm{T}} = (h^{\mathrm{HLL}} - (z_R - z_L)/2, h^{\mathrm{HLL}} + (z_R - z_L)/2, \theta^{\mathrm{HLL}}, \theta^{\mathrm{HLL}})^{\mathrm{T}}, \tag{2.14}$$

define a solution of (2.12). In addition, considering the cases where λ is large enough and $|z_R - z_L|$ is small enough, the Implicit Function Theorem can be applied to ensure the existence of regimes (formally, λ large and Δx small) such that the system (2.12) admits solutions. From a numerical point of view, the solutions of the system (2.12) may be easily computed with a fix point procedure or a standard Newton method as presented Section 3. As a consequence, we reasonably may assume the existence of solutions of the system (2.12). Now, we give our main result.

Theorem 2.2 *Consider the Riemann problem associated to the Ripa equations* (1.1) *with φ a strictly convex function satisfying the conditions* (1.2). *Consider $\tilde{w}(\cdot, w_L, w_R)$, an approximated Riemann solver in the form* (2.9) *made of two intermediates constant states w_L^*, w_R^* governed by the equations* (2.12). *Assume that a non interaction CFL condition holds. Consider a solution of* (2.12) *verifying $h_R^* - h_L^* + 2(z_R - z_L) \neq 0$ if $z_R \neq z_L$ and $\theta_R \neq \theta_L$. Consider the following discrete source term definition*:

$$\Delta x s_{LR} = -g h^{\mathrm{HLL}} \varphi(\theta^{\mathrm{HLL}})(z_R - z_L) - \frac{H(w_L) + H(w_R)}{c^{\mathrm{HLL}} h^{\mathrm{HLL}}}(z_R - z_L)^2, \tag{2.15}$$

where h^{HLL}, θ^{HLL}, c^{HLL} are defined in (2.11)-(2.12) *respectively. Then, the approximated Riemann solver satisfies the consistency condition* (2.2). *Moreover, for states w_L, w_R such that*

$$u_L = u_R = 0, \quad \theta_L = \theta_R, \quad h_L + z_L = h_R + z_R, \tag{2.16}$$

we have $u_L^* = u_R^* = 0$, $\theta_L^* = \theta_R^* = \theta_R = \theta_L$, $h_L^* = h_L$ and $h_R^* = h_R$. In addition, there exists λ large enough and Δt small enough, such that both entropy conditions (2.3)-(2.4) *are satisfied. As a consequence, the Godunov type scheme* (1.7) *associated to the solver $\tilde{w}(\cdot, \cdot, \cdot)$ is consistent, well-balanced for the lake at rest* (1.5) *and satisfies a fully discrete entropy inequality* (1.6) *with the numerical entropy flux detailed in* (2.8).

We impose the restriction $h_R^* - h_L^* + 2(z_R - z_L) \neq 0$ if $z_R \neq z_L$ and $\theta_R \neq \theta_L$ on the solutions of the system (2.12). Once again, according to our numerical experiments, this condition seems to be numerically satisfied. Now, we give the proof of the above statement.

Proof The consistency is a direct consequence of the conservation equations (2.12a)-(2.12c) used for the solver design. Now, we prove the well-balanced property. Since the conditions (2.16) hold for the lake at rest, let denote $\theta_L = \theta_R := \theta$ and using the equation (2.13) we have

$$u_L^* = u_R^* = \frac{2}{h_L + h_R}\Big(- \frac{g}{4\lambda}[h^2\varphi(\theta)] + \frac{\Delta x s_{LR}}{2\lambda}\Big) = -\frac{2}{h_L + h_R}\frac{g\varphi(\theta)}{4\lambda}\left((h_L + h_R)[h] + (h_L + h_R)[z]\right) = 0, \tag{2.17}$$

where we have set $[\cdot] := \cdot_R - \cdot_L$. Since $[\theta] = 0$, then a solution of (2.12a)-(2.12c)-(2.12d)-(2.12e) is given by (2.14). As a consequence, we have $\theta_L^* = \theta_R^* = \theta^{\mathrm{HLL}} = \theta$. Finally, for the water heights, since we have $[z] = -[h]$ for the lake at rest, we deduce form (2.14)

$$h_L^* = \frac{h_L + h_R}{2} + \frac{[z]}{2} = h_L, \quad \text{and,} \quad h_R^* = \frac{h_L + h_R}{2} - \frac{[z]}{2} = h_R,$$

that achieves to show the well balanced property. Concerning the entropy stability, we show that the inequalities (2.10b)-(2.10c) may be ensured with a suitable couple $(\lambda, \Delta t)$. Thanks to the strict convexity of $\theta \mapsto \varphi(\theta)$, with λ is large enough then straightforward computations yield to $\psi(w^{\mathrm{HLL}}) \leq \psi^{\mathrm{HLL}}$, and the case $\psi(w^{\mathrm{HLL}}) = \psi^{\mathrm{HLL}}$ occurs if and only if $[\theta] = 0$. As a consequence, using the equation (2.12d) and a continuity argument, it is possible to show that there exists Δt small enough such that

$$\frac{\psi(w_L^*) + \psi(w_R^*)}{2} = \psi(w^{\mathrm{HLL}})\Big(1 + \Big(\frac{c^{\mathrm{HLL}}\Delta t[\theta]}{(h\theta)^{\mathrm{HLL}}}\Big)^2\Big) \leq \psi^{\mathrm{HLL}}, \tag{2.18}$$

and thus the inequality (2.10b) holds. Now, we focus on (2.10c). If $[z] = 0$ then $w_L^* = w_R^* = w^{\mathrm{HLL}}$ is solution of the system (2.12). In this case, the solver (2.9) degenerates to the standard HLL solver [4] and the inequality (2.10c) is satisfied. In the case where $[z] \neq 0$, we show that the inequality (2.10c) reads

$$-\frac{g}{8}\varphi(\theta^{\mathrm{HLL}})\exp\Big(\frac{\lambda[\theta]^2}{c^{\mathrm{HLL}}(\theta^{\mathrm{HLL}})^2}\Big)[h^*+2z]^2+\frac{g}{8}\varphi(\theta^{\mathrm{HLL}})\Big(\frac{c^{\mathrm{HLL}}\Delta t[\theta]}{(h\theta)^{\mathrm{HLL}}}\Big)^2\Big([h^*+2z][h^*]+4(h^{\mathrm{HLL}})^2\Big)$$
$$\leq \eta^{\mathrm{HLL}}-\eta(\tilde{w}^{\mathrm{HLL}})+(\psi z)^{\mathrm{HLL}}-\psi^{\mathrm{HLL}}\frac{z_L+z_R}{2}, \quad (2.19)$$

with $\eta(\tilde{w}^{\mathrm{HLL}}) := (\tilde{hu}^{\mathrm{HLL}})^2/(2h^{\mathrm{HLL}}) + g(h^{\mathrm{HLL}})^2\varphi(\theta^{\mathrm{HLL}})/2$, $\tilde{hu}^{\mathrm{HLL}} := (hu)^{\mathrm{HLL}} + \Delta x s_{LR}/(2\lambda)$. Indeed, according to the definition of η given by (1.3), we have

$$\frac{\eta(w_L^*)+\eta(w_R^*)}{2}=\frac{1}{2}\left(\frac{h_R^*u_R^*+h_L^*u_L^*}{2}\frac{u_R^*+u_L^*}{2}+\frac{[h^*u^*][u^*]}{4}+\frac{h_L^*+h_R^*}{2}\frac{\psi(w_L^*)+\psi(w_R^*)}{2}+\frac{[h^*][\psi(w^*)]}{4}\right).$$

But, using the equations (2.12)-(2.13), the above equation reformulates

$$\frac{\eta(w_L^*)+\eta(w_R^*)}{2}=\eta(\tilde{w}^{\mathrm{HLL}})+\frac{g}{8}[h^*][h^*\varphi(\theta^*)]+\frac{h^{\mathrm{HLL}}}{2}\psi(w^{\mathrm{HLL}})\Big(\frac{c^{\mathrm{HLL}}\Delta t[\theta]}{(h\theta)^{\mathrm{HLL}}}\Big)^2. \quad (2.20)$$

Now we have to rewrite the quantity $[h^*\varphi(\theta^*)]$. In this regard, using the equations (2.12a)-(2.12d), a straightforward computation yields to

$$gh^{\mathrm{HLL}}\frac{\varphi(\theta_L^*)+\varphi(\theta_R^*)}{2}+g\frac{[h^*][\varphi(\theta^*)]}{4}=\psi(w^{\mathrm{HLL}})\Big(1+\Big(\frac{c^{\mathrm{HLL}}\Delta t(\theta_R-\theta_L)}{(h\theta)^{\mathrm{HLL}}}\Big)^2\Big).$$

As a consequence, we have

$$\begin{aligned}
[h^*\varphi(\theta^*)] &= [h^*]\frac{\varphi(\theta_L^*)+\varphi(\theta_R^*)}{2}+\frac{h_L^*+h_R^*}{2}[\varphi(\theta^*)],\\
&= [h^*]\left(\varphi(\theta^{\mathrm{HLL}})\Big(1+\Big(\frac{c^{\mathrm{HLL}}\Delta t[\theta]}{(h\theta)^{\mathrm{HLL}}}\Big)^2\Big)-\frac{[h^*][\varphi(\theta^*)]}{4h^{\mathrm{HLL}}}\right)+h^{\mathrm{HLL}}[\varphi(\theta^*)],\\
&= \varphi(\theta^{\mathrm{HLL}})\left([h^*]+h^{\mathrm{HLL}}\left(1-\left(\frac{[h^*]}{2h^{\mathrm{HLL}}}\right)^2\right)\frac{[\varphi(\theta^*)]}{\varphi(\theta^{\mathrm{HLL}})}\right)+[h^*]\varphi(\theta^{\mathrm{HLL}})\Big(\frac{c^{\mathrm{HLL}}\Delta t[\theta]}{(h\theta)^{\mathrm{HLL}}}\Big)^2,\\
&= \varphi(\theta^{\mathrm{HLL}})\left([h^*]+\frac{h_R^*h_L^*}{h^{\mathrm{HLL}}}\frac{[\varphi(\theta^*)]}{\varphi(\theta^{\mathrm{HLL}})}\right)+[h^*]\varphi(\theta^{\mathrm{HLL}})\Big(\frac{c^{\mathrm{HLL}}\Delta t[\theta]}{(h\theta)^{\mathrm{HLL}}}\Big)^2.
\end{aligned}$$

Plugging the above equation and the equation (2.20) in the inequality (2.10c), we obtain (2.19). Now, in order to enforce the condition (2.19), we have two distinct cases. The first one is $[\theta] \neq 0$. In this case, since $[h^*+2z]$ is assumed to be non null, then we deduce the existence of $(\lambda, \Delta t)$ such that the inequality (2.19) holds. The second case is $[\theta] = 0$. In this case, w_L^*, w_R^* are given by (2.13)-(2.14). Let denote $\theta := \theta_L = \theta_R$, so that the expected inequality (2.19) now writes

$$0 \leq \big(\eta^{\mathrm{HLL}}-\eta(\tilde{w}^{\mathrm{HLL}})\big)\big|_{[\theta]=0}+\frac{g\varphi(\theta)}{4}[z]\Big([h]-\frac{(hu)_L+(hu)_R}{\lambda}\Big)+\frac{g}{8}\varphi(\theta)[z]^2. \quad (2.21)$$

We show that the above inequality is satisfied for large enough λ. A direct computation provides

$$\big(\eta^{\mathrm{HLL}}-\eta(\tilde{w}^{\mathrm{HLL}})\big)\big|_{[\theta]=0}+\frac{g\varphi(\theta)}{4}[z]\Big([h]-\frac{(hu)_L+(hu)_R}{\lambda}\Big)+\frac{g}{8}\varphi(\theta)[z]^2 \underset{\lambda\to+\infty}{=} \frac{h_Rh_L[u]^2}{4(h_L+h_R)}$$
$$+\frac{g\varphi(\theta)}{8}[h+z]^2-\frac{[u]^2h_Rh_L(h_Lu_R-h_Ru_L)}{4\lambda(h_L+h_R)^2}-\frac{(hu)_L+(hu)_R}{4\lambda}\left(g\varphi(\theta)[z]+\frac{\Delta x s_{LR}}{h^{\mathrm{HLL}}}\right)+O\left(\frac{1}{\lambda^2}\right).$$

As soon as $[u] \neq 0$ or $[h+z] \neq 0$, (2.21) is satisfied with λ large enough. Now, if $[u] = 0$ and $[h+z] = 0$, denoting $u := u_L = u_R$, we have $\Delta x s_{LR}\Big|_{\substack{[\theta]=0,\\ [h+z]=0,\\ [u]=0}} = -gh^{\mathrm{HLL}}\varphi(\theta)[z] - gu\varphi(\theta)\frac{h_L+h_R}{c^{\mathrm{HLL}}h^{\mathrm{HLL}}}[z]^2$, that involves

$$\left(\eta^{\mathrm{HLL}}-\eta(\tilde{w}^{\mathrm{HLL}})+\frac{g\varphi(\theta)}{4}[z]\Big([h]-\frac{(hu)_L+(hu)_R}{\lambda}\Big)+\frac{g}{8}\varphi(\theta)[z]^2\right)\Big|_{\substack{[\theta]=0,\\ [h+z]=0,\\ [u]=0}}$$
$$\underset{\lambda\to+\infty}{=} gu^2\frac{(h_L+h_R)^2\varphi(\theta)}{4\lambda}\frac{[z]^2}{c^{\mathrm{HLL}}(h^{\mathrm{HLL}})^2}+O\left(\frac{1}{\lambda^2}\right).$$

Once again, we deduce that (2.21) is satisfied with λ large enough. The case $u = 0$, $[\theta] = 0$, $[h+z] = 0$ occurs only for the lake at rest in which both sides of the inequality (2.10c) goes to 0 thanks to the well-balanced property of the scheme. As a consequence, the two inequalities (2.10b)-(2.10c) always hold with a suitable choice of $(\lambda, \Delta t)$. Arguing Lemma 2.1, the proof is achieved. □

A numerical scheme (1.7) associated to the solver described in Theorem 2.2 verifies a discrete entropy stability but does not necessary preserve the convex set Ω. As a consequence we complete the solver with a conservative limitation procedure as done in [5]. The next section deals with the numerical experiments.

3. Numerical experiments

Theorem 2.2 ensures the existence of $(\lambda, \Delta t)$ such that the entropy stability is satisfied. From a practical point of view, we select a couple $(\lambda, \Delta t)$ as follows: for each interface, starting from

$$\lambda = \max_{\alpha \in \{L,R\}} (|u_\alpha \pm \sqrt{g h_\alpha \varphi(\theta_\alpha)}|, |u_\alpha|), \quad \Delta t = \frac{\Delta x}{2\lambda},$$

we solve the system (2.12) with a standard Newton method. If the solution satisfies the inequalities (2.18)-(2.19), then the above λ, Δt are accepted. Otherwise, we increase λ, we decrease Δt and we repeat the procedure until the two inequalities (2.18)-(2.19) are satisfied. We fix $g := 1$ and $\varphi(\theta) := \exp(\theta)$. The test case is a slightly modified version of the break dam introduced in [2]. The space domain is $[-1, 1]$. The bottom topography is given by

$$z(x) = \begin{cases} 2\cos(10\pi(x+0.3)) + 2, & \text{if } -0.4 \le x \le -0.2, \\ 0.5\cos(10\pi(x-0.3)) + 0.5, & \text{if } 0.2 \le x \le 0.4, \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

We consider the following initial condition:

$$(h, u, \varphi(\theta))^{\mathrm{T}}(x, t=0) = \begin{cases} (5-z, 0, 1.5)^{\mathrm{T}}, & \text{if } x \le 0, \\ (1-z, 0, 5)^{\mathrm{T}}, & \text{otherwise,} \end{cases} \tag{3.2}$$

that ensures that the fluid is always far from the dry areas. We lay down homogeneous Neumann conditions on both boundaries. We compute references solutions with the standard HLL scheme [4] coupled to the discrete source term (2.15) on a fine grid having 50 000 cells. The final time is 0.3. The Figure 1 displays the numerical results.

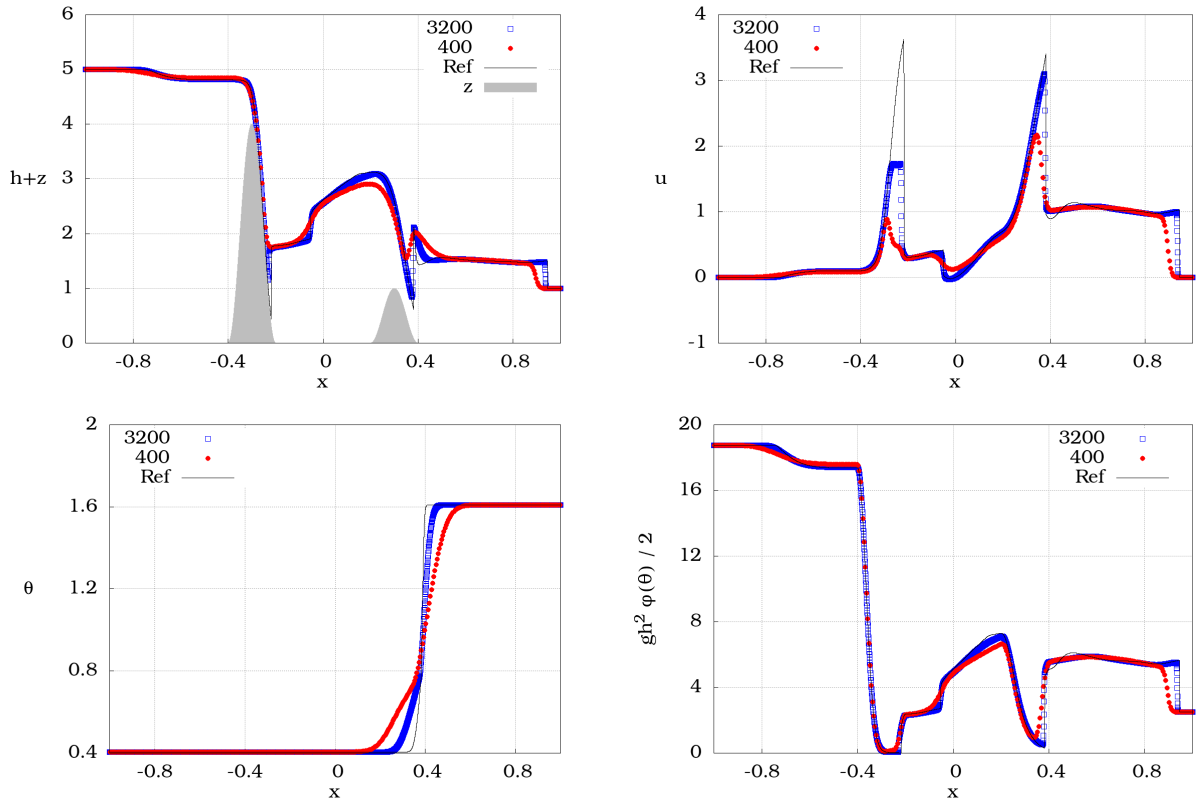


Fig. 1 Numerical results of the solver described in Theorem 2.2 and for the test case (3.1)-(3.2) at time $t = 0.3$, on a mesh made of 400 cells in red and made of 3 200 cells in blue.

We observe an acceptable agreement to the reference solution. The numerical solution is nevertheless diffusive. On a coarse grid, the front shock on the right is slightly misplaced and is localized before the reference position. This misplacement is widely corrected with finer grids. We measure τ defined as the ratio between the CPU time needed to run this problem with the scheme of Theorem 2.2 and the CPU time required to run the standard HLL scheme. We have $\tau = 2.15$. This increasing of the CPU time is due to the search of an admissible couple $(\lambda, \Delta t)$ and due to the non-linear set of equations (2.12) that have to be solved at each time step and on each interfaces.

4. Conclusion

We have introduced a well-balanced entropy stable scheme for the Ripa system in the wet regions. We have proved the well-balanced property and the fully discrete entropy stability verified by the scheme. The stability result depends on numerical viscosity λ and on time step Δt that are implicit in our main result. From a numerical point of view, the proposed scheme results are good but diffusive. As a consequence, the design of an efficient selection procedure of the couple $(\lambda, \Delta t)$ and the transition toward dry areas should be investigated. The existence of solutions of the system (2.12) and the solutions restriction assumed in Theorem 2.2 also should be investigated.

Acknowledgements

The authors thank the ANR project MUFFIN ANR-19-CE46-0004 for the financial support. The authors also thank Mehdi Badsi for fruitfully discussions.

References

[1] F. Bouchut. *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources*. Frontiers in Mathematics. Birkhäuser Verlag, Basel, 2004.

[2] A. Chertock, A. Kurganov, and Y. Liu. Central-upwind schemes for the system of shallow water equations with horizontal temperature gradients. *Numerische Mathematik*, 127, 08 2014.

[3] V. Desveaux, M. Zenk, C. Berthon, and C. Klingenberg. Well-balanced schemes to capture non-explicit steady states. Ripa model. *Mathematics of Computation*, 85:1, 10 2015.

[4] A. Harten, P.D. Lax, and B. Van Leer. On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. *SIAM review*, 25:35–61, 1983.

[5] V. Michel-Dansac, C. Berthon, S. Clain, and F. Foucher. A well-balanced scheme for the shallow-water equations with topography or Manning friction. *Journal of Computational Physics*, 335:115–154, 2017.

[6] C. Sánchez-Linares, T. Morales de Luna, and M. Castro. A HLLC scheme for Ripa model. *Applied Mathematics and Computation*, 272, 06 2015.

On the convergence of solutions of nonlinear elliptic problems with L^1 data

Antonio J. Martínez Aparicio
Universidad de Almería, Spain

Abstract

In this work we study the convergence of the sequence $\{u_n\}$ of weak solutions of the problems

$$\begin{cases} -\Delta_p u_n + \frac{1}{n}|f(x)|u_n = f(x) & \text{in } \Omega, \\ u_n = 0 & \text{on } \partial\Omega, \end{cases}$$

where Ω is a bounded open set of $\mathbb{R}^N$ $(N \geq 2)$, $-\Delta_p u$ is the usual p-Laplacian operator $(1 < p < \infty)$ and $f(x)$ is an $L^1(\Omega)$ function. This work has been motivated by the "Q-condition" result proven in [2] which guarantees the existence of u_n for every fixed $n \in \mathbb{N}$. We show that $\{u_n\}$ converges in some sense to u, the entropy solution of the problem

$$\begin{cases} -\Delta_p u = f(x) & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases}$$

1. Introduction

In this text we deal with some of the most relevant aspects of the paper [9]. For the sake of simplicity, we only study here the model problem. All this work revolves around the boundary value problem

$$\begin{cases} -\Delta_p u = f(x) & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{P}$$

where Ω is a bounded open set of $\mathbb{R}^N$ $(N \geq 2)$, $-\Delta_p u = -\text{div}(|\nabla u|^{p-2}\nabla u)$ is the p-Laplacian operator $(1 < p < \infty)$ and $f(x)$ is an $L^1(\Omega)$ function.

We stress that the problem (P) does not always have a weak solution in the $W_0^{1,1}(\Omega)$ space when the right-hand side is in $L^1(\Omega)$ and p is near 1; more specifically, when $p \leq 2 - \frac{1}{N}$. For the convenience of the reader, we have decided to include the Proposition 2.1 which contains an elemental proof of this fact.

To solve this question, many authors started to establish other more general concepts of solutions. In this line we find, for example, the work [6], where the authors use the notion of renormalized solution, or the paper [5], where the authors introduce the concept of entropy solution (see Definition 2.3). This last kind of solution will develop an essential role throughout the present work.

In [2] (see also [1]) Arcoya and Boccardo studied the regularizing effect that a lower order term could have on problem (P). Concretely, they considered the boundary value problem

$$\begin{cases} -\Delta_p u + b(x)u = f(x) & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{P_b}$$

with $0 \leq b(x) \in L^1(\Omega)$ and $f(x) \in L^1(\Omega)$. They showed that if the so-called Arcoya-Boccardo Q-condition is satisfied, i.e., if there is some $Q > 0$ such that

$$|f(x)| \leq Qb(x), \tag{1.1}$$

then there exists an unique weak solution $u \in W_0^{1,p}(\Omega) \cap L^\infty(\Omega)$ of (P_b). Moreover, they also proved that this solution is bounded in $L^\infty(\Omega)$ by Q, i.e., that

$$\|u\|_\infty \leq Q.$$

Therefore, these authors proved that this interplay between $b(x)$ and $f(x)$ provides a regularizing effect on problem (P).

Motivated by this result, for each $n \in \mathbb{N}$ we consider the problem

$$\begin{cases} -\Delta_p u_n + \frac{1}{n}|f(x)|u_n = f(x) & \text{in } \Omega. \\ u_n = 0 & \text{on } \partial\Omega. \end{cases} \tag{P_n}$$

See that its coefficients always satisfy the Q-condition (1.1) since

$$|f(x)| \leq n \cdot \tfrac{1}{n}|f(x)|,$$

so the results of [2] allows us to assure that for each $n \in \mathbb{N}$ there exists some $u_n \in W_0^{1,p}(\Omega) \cap L^\infty(\Omega)$ such that

$$\|u_n\|_\infty \leq n. \tag{1.2}$$

The main goal of this work is to study the behaviour of the sequence $\{u_n\}$ of weak solutions of (P_n) at infinity. We stress that similar studies can be done with other types of problems for which similar Q-condition results have been proven, such as the Hamilton-Jacobi equations (see [3]) or Dirichlet problems with distributional data (see [4]), among others.

Here below we state the main theorem of this work, which is extremely related with the entropy solution of (P) (see Definition 2.3). We recall that in [5] it is proved that this type of solution always exists and always is unique. Nevertheless, our result is an alternative proof of this fact where the main difference is that we do not have to approach the datum $f(x)$.

Theorem 1.1 *Assume that $f \in L^1(\Omega)$. Then the entropy solution u of (P) exists and the sequence $\{u_n\}$ of weak solutions of (P_n) satisfy that*

$$u_n \to u \text{ in measure.}$$

We point out that, in general, the sequence $\{u_n\}$ cannot be bounded in the $W_0^{1,1}(\Omega)$ space because that would imply the existence of a weak solution of (P) for every datum $f(x) \in L^1(\Omega)$, which is false in general (see Proposition 2.1). However, under the assumption that the weak solution of (P) exists it can be showed that $\{u_n\}$ converges strongly in $W_0^{1,p}(\Omega)$. The proof of this fact can be found in [9].

The scheme of this work is as follows. In Section 2 we include the proof of an elemental result about the nonexistence of weak solution for the problem (P) for at least one $f(x) \in L^1(\Omega)$, we recall the definition of weak and entropy solution of (P) and we give a brief review about the Marcinkiewicz spaces. Finally, in Section 3, we include a summarized version of the proof of the Theorem 1.1; the whole proof can be found in [9].

2. Preliminaries

We begin this section by giving an elemental proof of a well known result which partly motivated many authors to introduce new concepts of solutions different from the usual ones.

Proposition 2.1 *If $p \leq 2 - \frac{1}{N}$, then there exists some $f \in L^1(\Omega)$ such that the problem (P) does not have a solution in the $W_0^{1,1}(\Omega)$ space.*

Proof First of all, observe that $\Delta_p u$ can be seen as an element of $\left(W_0^{1,\frac{1}{2-p}}(\Omega)\right)' = W^{-1,\frac{1}{p-1}}(\Omega)$ when $u \in W_0^{1,1}(\Omega)$. Indeed, since $p < 2$, then $\frac{1}{p-1} > 1$ and thus we can apply Hölder's inequality to obtain that

$$\left|\int_\Omega |\nabla u|^{p-2}\nabla u \nabla \varphi\right| \leq \left(\int_\Omega |\nabla u|\right)^{p-1} \left(\int_\Omega |\nabla \varphi|^{\frac{1}{2-p}}\right)^{2-p} < +\infty \text{ when } \varphi \in W_0^{1,\frac{1}{2-p}}(\Omega).$$

So, if we assume that for every $f \in L^1(\Omega)$ the problem (P) has a solution u which belongs to $W_0^{1,1}(\Omega)$, from the equality $-\Delta_p u = f$ we can deduce that $L^1(\Omega) \subseteq W^{-1,\frac{1}{p-1}}(\Omega)$ and by this reason the duals of these spaces satisfy that $W_0^{1,\frac{1}{2-p}}(\Omega) \subset L^\infty(\Omega)$. But this inclusion is true if, and only if, $\frac{1}{2-p} > N$, i.e, if $p > 2 - \frac{1}{N}$, thus obtaining a contradiction. □

We stress that the concept of entropy solution of (P) solves the existence problem stated on the above result. We recall here its definition and the definition of weak solution of (P).

Definition 2.2 A function $u\colon \Omega \to \mathbb{R}$ is a *weak solution* for the problem (P) if $u \in W_0^{1,p}(\Omega)$ and

$$\int_\Omega |\nabla u|^{p-2}\nabla u \nabla \varphi = \int_\Omega f(x)\varphi$$

for every $\varphi \in W_0^{1,p}(\Omega) \cap L^\infty(\Omega)$.

Definition 2.3 A function $u\colon \Omega \to \mathbb{R}$ is an *entropy solution* for the problem (P) if $T_k(u) \in W_0^{1,p}(\Omega)$ for every $k > 0$ and

$$\int_\Omega |\nabla u|^{p-2}\nabla u \nabla T_k(u - \varphi) = \int_\Omega f(x) T_k(u - \varphi)$$

for every $\varphi \in W_0^{1,p}(\Omega) \cap L^\infty(\Omega)$ and every $k > 0$.

See that every weak solution of (P) is an entropy solution. Reciprocally, in [5, Corollary 4.3] it is proved that if the entropy solution belongs to $W_0^{1,p}(\Omega)$, then it is also a weak solution. This shows that, indeed, the concept of entropy solution is more general than the weak one.

In the proof of the Theorem 1.1 we use the Marcinkiewicz spaces. For the convenience of the reader, we recall here their definition and some of their properties. For $0 < q < \infty$, we denote by $\mathcal{M}^q(\Omega)$ the set of measurable functions $v\colon \Omega \to \mathbb{R}$ such that there exists $C > 0$ satisfying that

$$\operatorname{meas}\{|v| > k\} \leq \frac{C}{k^q},\ \forall k > 0. \tag{2.1}$$

This space is a complete quasi-normed space with the quasi-norm

$$\|v\|^q_{\mathcal{M}^q(\Omega)} = \inf\{C > 0 : (2.1) \text{ holds}\}.$$

We also recall that, since Ω is bounded, then

$$\mathcal{M}^{q_2}(\Omega) \hookrightarrow L^{q_1}(\Omega) \hookrightarrow \mathcal{M}^{q_1}(\Omega)$$

for $0 < q_1 < q_2 < \infty$.

Related with these spaces we state the following lemma whose proof can be found in [5, Lemma 4.1]. For any $k > 0$ we set $T_k(s) = \min\{k, \max\{s, -k\}\}$.

Lemma 2.4 ([5]) *Let $u\colon \Omega \to \mathbb{R}$ be a function such that $T_k(u) \in W_0^{1,p}(\Omega)$ for every $k > 0$ and*

$$\frac{1}{k}\int_{\{|u|<k\}} |\nabla u|^p \leq M$$

for some constant $M > 0$ and for every $k > 0$. Then $u \in \mathcal{M}^{p_1}(\Omega)$ for $p_1 = \frac{N(p-1)}{N-p}$ if $1 < p < N$ and for every $p_1 > 1$ if $p \geq N$. More precisely, there exists $C = C(M, N, p) > 0$ such that

$$\operatorname{meas}\{|u| > k\} \leq \frac{C}{k^{p_1}},\ \forall k > 0.$$

3. Convergence to the entropy solution

In this section we give a summarized version of the proof of the Theorem 1.1.

Proof (Proof of Theorem 1.1) First, let us remember that as u_n are weak solutions of (P_n), then for every $n \in \mathbb{N}$ and for every $\varphi \in W_0^{1,p}(\Omega) \cap L^\infty(\Omega)$ we have that

$$\int_\Omega |\nabla u_n|^{p-2}\nabla u_n \nabla \varphi + \frac{1}{n}\int_\Omega |f(x)| u_n \varphi = \int_\Omega f(x)\varphi. \tag{3.1}$$

Now we begin with the proof.

Step 1. $\{u_n\}$ is bounded on some Marcinkiewicz space.

Taking $T_k(u_n) \in W_0^{1,p}(\Omega) \cap L^\infty(\Omega)$ as test function in (3.1) we obtain for every $n \in \mathbb{N}$ and for every $k > 0$ that

$$\int_{\{|u_n|\leq k\}} |\nabla u_n|^{p-2}\nabla u_n \nabla T_k(u_n) + \frac{1}{n}\int_\Omega |f(x)| u_n T_k(u_n) = \int_\Omega f(x) T_k(u_n).$$

Observe that the second integral is nonnegative since $s\,T_k(s) \geq 0$ for every $s \in \mathbb{R}$, so, from the above equality we deduce that

$$\int_\Omega |\nabla T_k(u_n)|^p \leq \int_\Omega f(x) T_k(u_n) \leq k\|f\|_1,\ \forall n \in \mathbb{N},\ \forall k > 0. \tag{3.2}$$

Thus, we can apply Lemma 2.4 to assure that there exists a constant $C > 0$ depending only of N, p, α and f such that

$$\operatorname{meas}\{|u_n| > k\} \leq C k^{-\frac{N(p-1)}{N-p}}, \tag{3.3}$$

for every $n \in \mathbb{N}$ and every $k > 0$. As a consequence, we deduce that $\{u_n\}$ is bounded on the space $\mathcal{M}^{p_1}(\Omega)$ with $p_1 = \frac{N(p-1)}{N-p}$.

Step 2. $\{u_n\}$ converges in measure to some function u.

The key here is to prove that $\{u_n\}$ is Cauchy in measure by using the (3.3) estimate and that $\{T_k(u_n)\}$ is bounded in $W_0^{1,p}(\Omega)$ for every $k > 0$ by (3.2). Once this has been proven, this implies that there exists some measurable

function u such that $u_n \to u$ in measure. As a consequence, there exists a subsequence of $\{u_n\}$, still denoted by $\{u_n\}$, such that

$$u_n \to u \text{ a.e. in } \Omega.$$

Now, since for $k > 0$ fixed the sequence $\{T_k(u_n)\}$ is bounded in $W_0^{1,p}(\Omega)$ by (3.2) and $T_k(u)$ is its only possible almost everywhere limit because of the continuity of T_k, we can conclude that

$$\begin{array}{ll} T_k(u_n) \rightharpoonup T_k(u) & \text{in } W_0^{1,p}(\Omega), \\ T_k(u_n) \to T_k(u) & \text{in } L^p(\Omega), \\ T_k(u_n) \to T_k(u) & \text{a.e. in } \Omega. \end{array}$$

Observe that this implies that $T_k(u) \in W_0^{1,p}(\Omega)$ for every $k > 0$.

Step 3. $T_k(u_n)$ strongly converges to $T_k(u)$ in $W_0^{1,p}(\Omega)$ for every $k > 0$.

Following the ideas of [8], in order to obtain the strong convergence of the truncations in the $W_0^{1,p}(\Omega)$ space we choose

$$w_n = T_{2k}(u_n - T_h(u_n) + T_k(u_n) - T_k(u))$$

with $h > k > 0$ as test function in (3.1). After several calculations, we can prove that

$$\lim_{n\to\infty} \int_\Omega [|\nabla T_k(u_n)|^{p-2}\nabla T_k(u_n) - |\nabla T_k(u)|^{p-2}\nabla T_k(u)]\nabla(T_k(u_n) - T_k(u)) = 0.$$

This allows us to apply Lemma 5 of [7] to conclude that

$$T_k(u_n) \to T_k(u) \text{ strongly in } W_0^{1,p}(\Omega) \text{ for every } k > 0.$$

Step 4. u is the entropy solution of (P).

Let us take $T_k(u_n - \varphi)$ with $\varphi \in W_0^{1,p}(\Omega) \cap L^\infty(\Omega)$ and $k > 0$ as test function in (3.1). Observe that if we define $L = k + \|\varphi\|_\infty$, then we have that $\nabla T_k(u_n - \varphi) = 0$ on the set $\{|u_n| > L\}$, so we can write

$$\int_\Omega |\nabla u_n|^{p-2}\nabla u_n \nabla T_k(u_n - \varphi) = \int_\Omega |\nabla T_L(u_n)|^{p-2}\nabla T_L(u_n)\nabla T_k(u_n - \varphi)$$

and thus (3.1) with this test function can be rewriten as

$$\int_\Omega |\nabla T_L(u_n)|^{p-2}\nabla T_L(u_n)\nabla T_k(u_n - \varphi) + \frac{1}{n}\int_\Omega |f(x)|u_n T_k(u_n - \varphi) = \int_\Omega f(x)T_k(u_n - \varphi). \tag{3.4}$$

Since $T_L(u_n) \to T_L(u)$ strongly in $W_0^{1,p}(\Omega)$, then we have that $\nabla T_L(u_n) \to \nabla T_L(u)$ a.e. in Ω and, as a consequence of Lebesgue Theorem, we have that

$$|\nabla T_L(u_n)|^{p-2}\nabla T_L(u_n) \to |\nabla T_L(u)|^{p-2}\nabla T_L(u) \text{ in } L^{p'}(\Omega).$$

As we also have that $\nabla T_k(u_n - \varphi) \to \nabla T_k(u - \varphi)$ in $L^p(\Omega)$, we can assure that

$$\int_\Omega |\nabla T_L(u_n)|^{p-2}\nabla T_L(u_n)\nabla T_k(u_n - \varphi) \to \int_\Omega |\nabla T_L(u)|^{p-2}\nabla T_L(u)\nabla T_k(u - \varphi) = \int_\Omega |\nabla u|^{p-2}\nabla u \nabla T_k(u - \varphi).$$

If we bear in mind that $\frac{1}{n}|f(x)|g(u_n) \to 0$ in $L^1(\Omega)$ thanks to the (1.2) estimate, we can easily pass to the limit in (3.4) to obtain that

$$\int_\Omega |\nabla u|^{p-2}\nabla u \nabla T_k(u - \varphi) = \int_\Omega f(x)T_k(u - \varphi),$$

so we can conclude that u is the entropy solution of (P). Finally, observe that due to the uniqueness of the entropy solution we can assert that the whole original sequence $\{u_n\}$ converges in measure to u. □

Acknowledgements

Research supported by Ministerio de Ciencia, Innovación y Universidades (MCIU), Agencia Estatal de Investigación (AEI) and (FEDER) Fondo Europeo de Desarrollo Regional under Research Project PID2021-122122NB-I00. First, second and third author supported by Junta de Andalucía, Consejería de Transformación Económica, Industria, Conocimiento y Universidades-Unión Europea grant P18-FR-667. First and second author supported by Junta de Andalucía FQM-194 and first author also by CDTIME. Third and fourth author supported by Junta de Andalucía FQM-116 and by Junta de Andalucía, Consejería de Transformación Económica, Industria, Conocimiento y Universidades grant UAL2020-FQM-B2046. Second and fourth author also supported by Ministerio de Universidades (grants FPU21/04849 and FPU21/ 05578 respectively).

References

[1] D. Arcoya, L. Boccardo, *Regularizing effect of the interplay between coefficients in some elliptic equations*, J. Funct. Anal. 268 (2015), no. 5, 1153–1166.

[2] D. Arcoya, L. Boccardo, *Regularizing effect of L^q interplay between coefficients in some elliptic equations*, J. Math. Pures Appl. (9) 111 (2018), 106–125.

[3] D. Arcoya, L. Boccardo, *Regularizing effect of two hypotheses on the interplay between coefficients in some Hamilton-Jacobi equations*, Adv. Nonlinear Stud. 21 (2021), no. 2, 251–260.

[4] D. Arcoya, L. Boccardo, L. Orsina, *Regularizing effect of the interplay between coefficients in some nonlinear Dirichlet problems with distributional data*, Ann. Mat. Pura Appl. (4) 199 (2020), no. 5, 1909–1921.

[5] P. Bénilan, L. Boccardo, T. Gallouët, R. Gariepy, M. Pierre, J. L. Vázquez, *An L^1-theory of existence and uniqueness of solutions of nonlinear elliptic equations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) 22 (1995), no. 2, 241–273.

[6] L. Boccardo, D. Giachetti, J. I. Díaz, F. Murat, *Existence and regularity of renormalized solutions for some elliptic problems involving derivatives of nonlinear terms*, J. Differential Equations 106 (1993), no. 2, 215–237.

[7] L. Boccardo, F. Murat, J. P. Puel, *Existence of bounded solutions for nonlinear elliptic unilateral problems*, Ann. Mat. Pura Appl. (4) 152 (1988), 183–196.

[8] C. Leone, A. Porretta, *Entropy solutions for nonlinear elliptic equations in L^1*, Nonlinear Anal. 32 (1998), no. 3, 325–334.

[9] A. J. Martínez Aparicio, *Convergence of weak solutions of elliptic problems with datum in L^1*, submitted to publication.

Existence and regularity of solutions in a semilinear problem with singularity in the datum

José Carmona[1], Antonio J. Martínez Aparicio[1], Pedro J. Martínez-Aparicio[1], Miguel Martínez-Teruel[2]
1. *Universidad de Almería, Spain*
2. *Universidad de Granada, Spain*

Abstract

In this work, we analyze different cases for the lower order terms and, in each case, obtaining a regularizing effect on singular problems. The model of this problems is

$$\begin{cases} -\Delta u + a(x)g(u) = \dfrac{f(x)}{u^\gamma} & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

where Ω is a bounded open set of $\mathbb{R}^N$, $\gamma > 0$, $f(x)$ and $a(x)$ are nonnegative functions in $L^1(\Omega)$ and $g(s)$ is a continuous function. Imposing to the datum $f(x)$ an interaction either with the boundary of the domain or with the lower order term, we are able to prove the existence of solution.

1. Introduction

In this work, we study the following boundary value problem

$$\begin{cases} -\Delta u + a(x)g(u) = \frac{f(x)}{u^\gamma} & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{1.1}$$

Here, Ω is a bounded open subset of $\mathbb{R}^N$ ($N \geq 2$), $a(x), f(x) \in L^1(\Omega)$ are nonnegative functions and $g(s)$ is a continuous function.

The scope of this text is to analyze the existence of solutions to (1.1) in $H_0^1(\Omega)$ in a more general range of values of γ that has been studied until now (regularizing effect). We show that it is connected with the interplay of f with the boundary of Ω or with the lower order term. We point out that the hypotheses that we will impose are natural in order to obtain our principal results according to the existing literature on the problem (1.1).

In [6], the authors studied the problem (1.1) with $f(x)$ a positive Hölder continuous function in $\overline{\Omega}$ and they showed that this problem always has a classical solution which may not be in $H_0^1(\Omega)$. Concretely, they proved that the solution belongs to $H_0^1(\Omega)$ if, and only if, $\gamma < 3$.

In [3], the authors extensively study problem (1.1) with $f \in L^m(\Omega)$ for $m \geq 1$ and obtain existence results depending on γ and on the summability of f. For $\gamma = 1$ and $f \in L^1(\Omega)$, they proved the existence of a solution belonging to $H_0^1(\Omega)$. The same was proved for $\gamma < 1$, but this time more summability on f is needed, namely $f \in L^m(\Omega)$ with $m \geq C(N, \gamma) > 1$. Finally, for the case $\gamma > 1$ and $f \in L^1(\Omega)$ it was proved the existence of a solution u belonging to $H_{\text{loc}}^1(\Omega)$ satisfying that $u^{\frac{\gamma+1}{2}}$ belongs to $H_0^1(\Omega)$.

In [2], the authors partially improved the results in [3] for the case $\gamma > 1$ by adding more restrictive hypotheses. Specifically, for a regular domain and for $f \in L^m(\Omega)$ greater than a positive constant they proved the existence of a finite energy solution to (1.1), with $a(x)g(u) \equiv 0$, for every $1 < \gamma < \frac{3m-1}{m+1}$. These results seem optimal, since if f belongs to $L^\infty(\Omega)$ then we have that such solution is in $H_0^1(\Omega)$ for all $\gamma < 3$ (see [6]).

2. Main results

Our approach is twofold, on one hand we extend known results to the problem (1.1) and on the other hand we analyze the regularizing effect produced by different interplays of f, illustrated here according to whether γ is greater or less than one.

In the first case ($\gamma > 1$), we will assume that there exists $r > -1$ such that, the function $f(x)$ satisfies, for some $m_1 > 0$, that

$$f(x) \geq m_1 \varphi_1^r \text{ a.e. in } \Omega, \tag{2.1}$$

where φ_1 denotes a positive eigenfunction associated to the first eigenvalue of the operator $-\Delta$ with zero Dirichlet boundary conditions. Also, we will assume that $g\colon \mathbb{R} \to \mathbb{R}$ is continuous function verifying that

$$g(s) \text{ is nonnegative and increasing} \tag{2.2}$$

and that

$$a(x)g(s) \in L^1_{\text{loc}}(\Omega) \text{ for each } s \gg 0 \text{ fixed.} \tag{2.3}$$

Moreover, we will suppose that there exists some $0 < s_0 < 1$ and some $c_1, c_2 > 0$ such that

$$\begin{cases} a(x)g(s) \leq c_1 s^{\frac{r-2\gamma}{2+r}} & \text{if } r \geq 2\gamma, \\ a(x)g(s) \leq c_1 (s+c_2)^{\frac{r-2\gamma}{2+r}} & \text{if } r < 2\gamma, \end{cases} \tag{2.4}$$

for every $0 \leq s \leq s_0$ and almost every $x \in \Omega$.

Finally, the regularity result is obtained when there exists $m_2 > 0$ and an open neighborhood of $\partial\Omega$ in Ω, denoted by Γ, such that

$$f(x) \leq m_2 \varphi_1^r \text{ a.e. in } \Gamma. \tag{2.5}$$

The main result of the paper in the case $\gamma > 1$ is the following one.

Theorem 2.1 *Assume that Ω satisfies the interior sphere condition, $\gamma > 1$ and that $g(s)$ satisfies* (2.2) *and* (2.3). *Assume also that there exists $r > -1$ such that $0 \leq f(x) \in L^1(\Omega)$ satisfies* (2.1) *and $a(x)g(s)$ verifies* (2.4). *Then there exists $u \in H^1_{\text{loc}}(\Omega)$ solution to* (1.1) *such that $u^{\frac{\gamma+1}{2}} \in H^1_0(\Omega)$ and, if $\gamma < 2r+3$ and $f(x)$ satisfies* (2.5) *then $u \in H^1_0(\Omega)$.*

In the second case ($\gamma \leq 1$), we are inspired by [1]. We assume that $g\colon \mathbb{R} \to \mathbb{R}$ is a continuous function satisfying that

$$g \text{ is increasing, odd and we denote } g_\infty = \lim_{s \to +\infty} g(s). \tag{2.6}$$

We also assume the called "Q-condition":

$$\text{there exists } Q \in (0, g_\infty) \text{ such that } f(x) \leq Q a(x) \text{ a.e. in } \Omega. \tag{2.7}$$

We remark that (2.7) is natural in order to obtain more regularity as we can observe in [1] where the authors show this phenomenon thanks to this hypothesis for the first time in the literature.

Our main result of the paper in the case $\gamma \leq 1$ is the following one.

Theorem 2.2 *Assume that $\gamma \leq 1$ and g verifies* (2.6). *Assume also that $a(x)$ and $f(x)$ satisfy* (2.7). *Then the problem* (1.1) *has only one solution $u \in H^1_0(\Omega) \cap L^\infty(\Omega)$.*

Remark 2.3 The lower order term $a(x)g(u)$ can be generalized to a Carathéodory function $g(x,s)$. General results can be found in [4].

3. Preliminaries

The concept of solution we will adopt is gathered in the following definition.

Definition 3.1 A function $u \in H^1_{\text{loc}}(\Omega)$ such that $u \geq 0$ a.e. in Ω, $a(x)g(u) \in L^1_{\text{loc}}(\Omega)$, $\frac{f}{u^\gamma} \in L^1_{\text{loc}}(\Omega)$ is said to be a *supersolution* to problem (1.1) if

$$\int_\Omega \nabla u \nabla \phi + \int_\Omega a(x)g(u)\phi \geq \int_\Omega \frac{f}{u^\gamma}\phi, \ \forall 0 \leq \phi \in C^1_c(\Omega).$$

We say that u is a *subsolution* for problem (1.1) if the reverse inequality is satisfied and $u^\tau \in H^1_0(\Omega)$ for some $\tau > 0$.

A function $u \in H^1_{\text{loc}}(\Omega)$ is a *solution* for (1.1) if it is both a subsolution and a supersolution for such a problem. If, in addition, $u \in H^1_0(\Omega)$, we say that u is a *finite energy solution* for problem (1.1).

Let us clarify that the function $\frac{f}{u^\gamma}\phi$ takes the value $+\infty$ in the case $u = 0$ and $f\phi \neq 0$ while takes the value zero whenever $f\phi = 0$.

For any $k > 0$ we set $T_k(s) = \min\{k, \max\{s, -k\}\}$ and $G_k(s) = s - T_k(s)$.

In the next result we summarize the main existence results for

$$\begin{cases} -\Delta u_n + g_n(x, u_n) = \dfrac{f_n(x)}{\left(|u_n| + \frac{1}{n}\right)^\gamma} & \text{in } \Omega, \\ u_n = 0 & \text{on } \partial\Omega. \end{cases} \tag{3.1}$$

Lemma 3.2 *Assume that* $0 \le f_n \in L^\infty(\Omega)$ *and* $g_n(x,s)$ *is a Carathéodory function with* $g_n(x,s)s \ge 0$ *for every* $s \in \mathbb{R}$ *and a.e.* $x \in \Omega$ *and* g_n *is bounded for* s *in bounded sets. There exists* $0 \le u_n \in H_0^1(\Omega)$ *solution of* (3.1) *for every fixed* $n \in \mathbb{N}$. *In addition, for* $\gamma \le 1$ *we have that* $u_n \in L^\infty(\Omega)$. *Moreover, in the case* $\gamma > 1$, *the existence of solution* $u_n \in H_0^1(\Omega)$ *is deduced even for* $f_n \in L^1(\Omega)$.

The rest of the section is devoted to the case $\gamma > 1$ where we approximate the datum $f(x)$ by

$$f_n(x) = f(x) + \frac{\chi(r)}{n^{r(\gamma+1)/(2+r)}},$$

where r is given by (2.1), $\chi(r) = 0$ for $r \le 0$ and $\chi(r) = c_1 + 1$ for $r > 1$, where c_1 is given by (2.4). We also approximate the nonlinearity $a(x)g(s)$ by a suitable sequence of Carathéodory functions g_n defined in $\Omega \times \mathbb{R}$. Specifically, we define

$$g_n(x,s) = \begin{cases} T_n(a(x)g(s)), & s \ge \frac{1}{n}, \\ n\, s\, T_n(a(x)g(s)), & 0 < s < \frac{1}{n}, \\ 0, & s \le 0. \end{cases}$$

Observe that $g_n(x,s)$ is increasing in s for a.e. $x \in \Omega$ when (2.2) is satisfied and that $g_n(x,s) \le a(x)g(s)$ for $s \ge 0$.

Following the ideas of [2], in the next lemma we prove that a certain power of an approximation of φ_1 is a subsolution of (3.1).

Lemma 3.3 *Assume that* $\gamma > 1$, $g(s)$ *satisfies* (2.2) *and there exists* $r > -2$ *such that* $0 \le f(x) \in L^1(\Omega)$ *verifies* (2.1) *and* $a(x)g(s)$ *verifies* (2.4). *Then there exists* $n_0 \in \mathbb{N}$ *such that the function*

$$z_n(x) = \left(C\varphi_1(x) + \frac{1}{n^{(\gamma+1)/(2+r)}} \right)^{\frac{2+r}{\gamma+1}} - \frac{1}{n}$$

is a subsolution of (3.1) *for* $n \ge n_0$ *and for* $C > 0$ (*independent of* n) *sufficiently small. As a consequence,*

$$z_n \le u_n \text{ a.e. } in\ \Omega.$$

The Lemma 3.3 allows us to obtain that lower boundedness for the sequence $\{u_n\}_{n \ge n_0}$ and this suffices to prove the existence of a solution of (1.1).

Theorem 3.4 *Assume that* $\gamma > 1$ *and* $g(s)$ *satisfies* (2.2) *and* (2.3). *Assume also that there exists* $r > -1$ *such that* $0 \le f(x) \in L^1(\Omega)$ *satisfies* (2.1) *and* $a(x)g(s)$ *verifies* (2.4). *Then there exists* $u \in H^1_{\mathrm{loc}}(\Omega)$ *solution of* (1.1) *satisfying that* $u^{\frac{\gamma+1}{2}} \in H_0^1(\Omega)$. *Moreover,* $u_n \to u$ a.e. *in* Ω.

4. Regularizing effect due to the behavior of the data at the boundary of Ω

In this section we prove Theorem 2.1.

Proof Taking $\left(T_k(u_n) + \frac{1}{n}\right)^\theta - \frac{1}{n^\theta} \in H_0^1(\Omega) \cap L^\infty(\Omega)$ with $\theta > \max\left\{0, \gamma - \frac{(r+1)(\gamma+1)}{2+r}\right\}$ as test function in (3.1), we obtain, after dropping a positive term, that

$$\alpha\theta \int_\Omega \left(T_k(u_n) + \frac{1}{n}\right)^{\theta-1} |\nabla T_k(u_n)|^2 \le \int_\Omega f_n \left(u_n + \frac{1}{n}\right)^{\theta-\gamma}. \tag{4.1}$$

If we take $\theta < \gamma$, we can apply the Lemma 3.3 to deduce that

$$\int_\Omega f_n \left(u_n + \frac{1}{n}\right)^{\theta-\gamma} \le \int_\Omega f_n \left(C\varphi_1(x) + \frac{1}{n^{(\gamma+1)/(2+r)}} \right)^{\frac{(2+r)(\theta-\gamma)}{\gamma+1}}. \tag{4.2}$$

On one hand, there is $C_1 > 0$ such that $\varphi_1 > C_1$ in $\Omega \backslash \Gamma$ (Γ given by (2.5)) since $\varphi_1 > 0$ in Ω, $\varphi_1 \in C(\Omega)$ and $\Omega \backslash \Gamma$ is closed. Therefore, we have that

$$\int_{\Omega\backslash\Gamma} f_n \left(C\varphi_1(x) + \frac{1}{n^{(\gamma+1)/(2+r)}} \right)^{\frac{(2+r)(\theta-\gamma)}{\gamma+1}} \le C_2 \int_\Omega (f + 1 + c_1). \tag{4.3}$$

On the other hand, we can apply hypothesis (2.5) to obtain that

$$\int_\Gamma f_n \left(C\varphi_1(x) + \frac{1}{n^{(\gamma+1)/(2+r)}} \right)^{\frac{(2+r)(\theta-\gamma)}{\gamma+1}} \le C_1 \int_\Gamma \left(\varphi_1(x) + \frac{\chi(r)}{n^{(\gamma+1)/(2+r)}} \right)^{r + \frac{(2+r)(\theta-\gamma)}{\gamma+1}} < +\infty. \tag{4.4}$$

In this way, we can deduce from (4.1), (4.2), (4.3) and (4.4) that the sequence

$$\left\{\left(T_k(u_n)+\frac{1}{n}\right)^{\frac{\theta+1}{2}}-\frac{1}{n^{\frac{\theta+1}{2}}}\right\}$$

is bounded in $H_0^1(\Omega)$ by a constant independent of k. For this reason, we can use Fatou Lemma to assure that

$$\left\{\left(u_n+\frac{1}{n}\right)^{\frac{\theta+1}{2}}-\frac{1}{n^{\frac{\theta+1}{2}}}\right\}$$

is bounded in $H_0^1(\Omega)$ and thus, up to a subsequence, we can assume that it converges weakly in $H_0^1(\Omega)$. Since $u_n \to u$ a.e. in Ω, this weak limit has to be equal to $u^{\frac{\theta+1}{2}}$ and, consequently $u^{\frac{\theta+1}{2}} \in H_0^1(\Omega)$.

Finally, let us note that

$$\theta \in \left]\max\left\{0, \gamma-\frac{(r+1)(\gamma+1)}{2+r}\right\}, \gamma\right[\iff 1<\gamma<2r+3.$$

□

5. Regularizing effect thanks to the Q-condition

In this section we prove Theorem 2.2.

Proof Inspired by [1], we define the approximated problems

$$\begin{cases} -\Delta u_n + a_n(x)g(u_n) = \dfrac{f_n(x)}{\left(|u_n|+\frac{1}{n}\right)^{\gamma}} & \text{in } \Omega, \\ u_n = 0 & \text{on } \partial\Omega, \end{cases} \tag{5.1}$$

where

$$f_n(x)=\frac{f(x)}{1+\frac{1}{n}f(x)}, \qquad a_n(x)=\frac{a(x)}{1+\frac{Q}{n}a(x)}.$$

Since the Q-condition is hold, $a_n(x)$ and $f_n(x)$ are nonnegative functions and $g(s)s \geq 0$ for all $s \in \mathbb{R}$ by (2.6), we can apply Lemma 3.2 to assure the existence of $0 \leq u_n \in H_0^1(\Omega) \cap L^\infty(\Omega)$ solution of (5.1), i.e., satisfying

$$\int_\Omega \nabla u_n \nabla \phi + \int_\Omega a_n(x)g(u_n)\phi = \int_\Omega \frac{f_n(x)\phi}{\left(|u_n|+\frac{1}{n}\right)^{\gamma}}, \ \forall \phi \in H_0^1(\Omega)\cap L^\infty(\Omega). \tag{5.2}$$

The scheme of the rest of the proof is as follows:

Step 1. $\{u_n\}$ is bounded in $L^\infty(\Omega)$ and in $H_0^1(\Omega)$.

Step 2. Control of the right hand side integral of (5.2).

Step 3. Passing to the limit in (5.2).

Step 1. In this step we essentially apply the ideas of [1]. To obtain the boundedness of $\{u_n\}$ in $L^\infty(\Omega)$ we use $G_k(u_n) \in H_0^1(\Omega)\cap L^\infty(\Omega)$ as test function in (5.1), with $k=\max\{1, g^{-1}(Q)\}$ obtaining that

$$\alpha\int_\Omega |\nabla G_k(u_n)|^2 + \int_\Omega a_n(x)[g(u_n)-Q]G_k(u_n) \leq 0$$

and, as the second integral is nonnegative because $g(u_n) \geq Q$ on the set $\{u_n \geq k\}$, we can conclude that $\|G_k(u_n)\|_{H_0^1(\Omega)}=0$ and then $\{u_n\}$ is bounded in $L^\infty(\Omega)$ with $\|u_n\|_\infty \leq k$.

Now, using $u_n \in H_0^1(\Omega)\cap L^\infty(\Omega)$ as test function in (5.1) and using this boundedness of $\{u_n\}$ in $L^\infty(\Omega)$, we can deduce by (2.6) that

$$\alpha\int_\Omega |\nabla u_n|^2 \leq \int_\Omega f(x)k^{1-\gamma}.$$

Thus, $\{u_n\}$ is bounded in $H_0^1(\Omega)$. Therefore, there exists a subsequence, still denoted by $\{u_n\}$, which converges weakly in $H_0^1(\Omega)$ and a.e. to some $0 \leq u \in H_0^1(\Omega)$ with $\|u\|_\infty \leq k$.

Step 2. In this part we follow the ideas of [5]. We introduce for $\delta > 0$ the function

$$Z_\delta(s) = \begin{cases} 1, & \text{if } 0 \le s \le \delta, \\ -\frac{s}{\delta} + 2, & \text{if } \delta \le s \le 2\delta, \\ 0, & \text{if } 2\delta \le s. \end{cases}$$

Taking $Z_\delta(u_n)\phi \in H_0^1(\Omega) \cap L^\infty(\Omega)$ as test function in (5.1), where $\phi \in H_0^1(\Omega) \cap L^\infty(\Omega)$ with $\phi \ge 0$, one has

$$\int_\Omega \nabla u_n \nabla \phi \, Z_\delta(u_n) + \int_\Omega a_n(x) g(u_n) Z_\delta(u_n)\phi = \frac{1}{\delta} \int_{\{\delta \le u_n \le 2\delta\}} \nabla u_n \nabla u_n \phi + \int_\Omega \frac{f_n(x)}{\left(u_n + \frac{1}{n}\right)^\gamma} Z_\delta(u_n)\phi.$$

Since $Z_\delta(u_n) = 1$ in $\{u_n \le \delta\}$ and the first integral of the right hand side is positive, we can deduce the inequality

$$0 \le \int_{\{u_n \le \delta\}} \frac{f_n(x)}{\left(u_n + \frac{1}{n}\right)^\gamma} \phi \le \int_\Omega \nabla u_n \nabla \phi \, Z_\delta(u_n) + \int_\Omega a_n(x) g(u_n) Z_\delta(u_n)\phi.$$

Using that $\{u_n\}$ is bounded in $L^\infty(\Omega)$ and converges weakly in $H_0^1(\Omega)$ and a.e. in Ω to u, we can easily pass to the limit in n to obtain that

$$0 \le \limsup_{n \to +\infty} \int_{\{u_n \le \delta\}} \frac{f_n(x)}{\left(u_n + \frac{1}{n}\right)^\gamma} \phi \le \int_\Omega \nabla u \nabla \phi Z_\delta(u) + \int_\Omega a(x) g(u) Z_\delta(u)\phi.$$

Now, we pass to the limit as δ tends to 0 using that $g(0) = 0$ and that $\nabla u = 0$ a.e. in $\{u = 0\}$ since $u \in H_0^1(\Omega)$ allow us to conclude that

$$\limsup_{n \to +\infty} \int_{\{u_n \le \delta\}} \frac{f_n(x)}{\left(u_n + \frac{1}{n}\right)^\gamma} \phi \to 0 \text{ as } \delta \to 0. \tag{5.3}$$

Step 3. At this point, thanks to the boundness and convergence of u_n we can assure that

$$\int_\Omega \nabla u_n \nabla \phi + \int_\Omega a_n(x) g(u_n)\phi \to \int_\Omega \nabla u \nabla \phi + \int_\Omega a(x) g(u)\phi \ \forall \phi \in H_0^1(\Omega) \cap L^\infty(\Omega). \tag{5.4}$$

Now we choose $\delta_m \to 0$ such that $\text{meas}\{u = \delta_m\} = 0$ an we split the right hand side integral of (5.2) into two parts, namely

$$\int_\Omega \frac{f_n(x)}{\left(u_n + \frac{1}{n}\right)^\gamma} \phi = \int_{\{u_n \le \delta_m\}} \frac{f_n(x)}{\left(u_n + \frac{1}{n}\right)^\gamma} \phi + \int_\Omega \frac{f_n(x)}{(u_n + \frac{1}{n})^\gamma} \chi_{\{u_n > \delta_m\}} \phi. \tag{5.5}$$

Observe that thanks to (5.2), (5.4) and (5.5) and Lebesgue Theorem we have that

$$\lim_{n \to \infty} \int_{\{u_n \le \delta_m\}} \frac{f_n(x)}{\left(u_n + \frac{1}{n}\right)^\gamma} \phi = \int_\Omega \nabla u \nabla \phi + \int_\Omega a(x) g(u)\phi - \int_{\{u > \delta_m\}} \frac{f(x)}{u^\gamma} \phi$$

and, using (5.3), we obtain that

$$\lim_{m \to \infty} \int_{\{u > \delta_m\}} \frac{f(x)}{u^\gamma} \phi = \int_\Omega \nabla u \nabla \phi + \int_\Omega a(x) g(u)\phi. \tag{5.6}$$

In particular, using Fatou Lemma we deduce that $\frac{f(x)}{u^\gamma}\phi \in L^1(\{u > 0\})$ and then, using Lebesgue Theorem that

$$\lim_{m \to \infty} \int_{\{u > \delta_m\}} \frac{f(x)}{u^\gamma} \phi = \int_{\{u > 0\}} \frac{f(x)}{u^\gamma} \phi. \tag{5.7}$$

In addition, we can apply Fatou Lemma to obtain that

$$\int_{\{u=0\}} \frac{f(x)}{u^\gamma} \phi \le \limsup_{n \to +\infty} \int_{\{u_n \le \delta\}} \frac{f_n(x)}{\left(u_n + \frac{1}{n}\right)^\gamma} \phi, \ \forall \delta > 0,$$

which in view of (5.3) implies that

$$\int_{\{u=0\}} \frac{f(x)}{u^\gamma} \phi = 0 \text{ and } \int_{\{u>0\}} \frac{f(x)}{u^\gamma} \phi = \int_\Omega \frac{f(x)}{u^\gamma} \phi.$$

This, combined with (5.6) and(5.7) give us

$$\int_\Omega \nabla u \nabla \phi + \int_\Omega a(x)g(u)\phi = \int_\Omega \frac{f(x)}{u^\gamma}\phi, \ \forall \phi \in H_0^1(\Omega) \cap L^\infty(\Omega).$$

Moreover, $a(x)g(u) \in L^1(\Omega)$ since $u \in L^\infty(\Omega)$ and $\frac{f}{u^\gamma} \in L^1_{\text{loc}}(\Omega)$ since $\int_\Omega \frac{f(x)}{u^\gamma}|\phi| < +\infty$ for every $\phi \in H_0^1(\Omega) \cap L^\infty(\Omega)$. Thus, it is proved that the function $u \in H_0^1(\Omega) \cap L^\infty(\Omega)$ is a solution of (1.1), as desired. □

Acknowledgements

Research supported by Ministerio de Ciencia, Innovación y Universidades (MCIU), Agencia Estatal de Investigación (AEI) and (FEDER) Fondo Europeo de Desarrollo Regional under Research Project PID2021-122122NB-I00. First, second and third author supported by Junta de Andalucía, Consejería de Transformación Económica, Industria, Conocimiento y Universidades-Unión Europea grant P18-FR-667. First and second author supported by Junta de Andalucía FQM-194 and first author also by CDTIME. Third and fourth author supported by Junta de Andalucía FQM-116 and by Junta de Andalucía, Consejería de Transformación Económica, Industria, Conocimiento y Universidades grant UAL2020-FQM-B2046. Second and fourth author also supported by Ministerio de Universidades (grants FPU21/04849 and FPU21/ 05578 respectively).

References

[1] D. Arcoya and L. Boccardo, Regularizing effect of the interplay between coefficients in some elliptic equations, *J. Funct. Anal.* **268** (2015), no. 5, 1153–1166.

[2] D. Arcoya and L. Moreno-Mérida, Multiplicity of solutions for a Dirichlet problem with a strongly singular nonlinearity, *Nonlinear Anal.* **95** (2014), 281–291.

[3] L. Boccardo and L. Orsina, Semilinear elliptic equations with singular nonlinearities, *Calc. Var. Partial Differential Equations* **37** (2010), no. 3-4, 363–380.

[4] J. Carmona, A. J. Martínez Aparicio, P. J. Martínez-Aparicio, M. Martínez-Teruel, Regularizing effect in singular semilinear problems. Preprint, 2023.

[5] D. Giachetti, P. J. Martínez-Aparicio and F. Murat, A semilinear elliptic equation with a mild singularity at $u = 0$: Existence and homogenization, *J. Math. Pures Appl. (9)* **107** (2017), no. 1, 41–77.

[6] A. C. Lazer and P. J. McKenna, On a Singular Nonlinear Elliptic Boundary-Value Problem, *Proc. Amer. Math. Soc.* **111** (1991), no. 3, 721–730.

On a special class of boundary optimal control problems

Pablo Pedregal
Universidad de Castilla-La Mancha, Spain

Abstract

We introduce a special family of boundary optimal control problems in which we select, in a very precise way, the class of Dirichlet-boundary data that we would allow to compete. Not every possible datum is admitted. We therefore work in a subspace of the form

$$\mathbb{L} \equiv \mathbb{H} + H_0^1(\Omega) \subset H^1(\Omega)$$

for a proper subspace $\mathbb{H}$ in $H^1(\Omega)$. In this form, only boundary data found in $\mathbb{H}$ are admitted. Other ingredients are typical of a standard optimal control problem like the cost functional or the PDE-state equation. We will explore existence and optimality in this framework, and write a few helpful ideas about the numerical approximation of optimal solutions. The practical implementation is so special that it would have to wait, though, until some issues are resolved.

1. Introduction

Let $\Omega \subset \mathbb{R}^N$ be a regular, bounded domain, and let $\mathbb{H}$ be a non-trivial, closed subspace of $H^1(\Omega)$ such that

$$\mathbb{H} \cap H_0^1(\Omega)$$

is the trivial function. Under these circumstances, the subspace $\mathbb{H} + H_0^1(\Omega)$ is a direct sum

$$\mathbb{L} \equiv \mathbb{H} \oplus H_0^1(\Omega),$$

and the two (non-orthogonal) projections

$$\pi_1 : \mathbb{H} \oplus H_0^1(\Omega) \mapsto \mathbb{H}, \quad \pi_2 : \mathbb{H} \oplus H_0^1(\Omega) \mapsto H_0^1(\Omega)$$

are linear, continuous operators. The subspace $\mathbb{H}$ contains the set of feasible boundary conditions that we would like to consider, in the sense that functions $u \in \mathbb{L}$ are such that there is some $U \in \mathbb{H}$ with $u - U \in H_0^1(\Omega)$. Since

$$H_0^1(\Omega)^\perp = \{u \in H^1(\Omega) : -\Delta u + u = 0 \text{ in } \Omega\}$$

is the orthogonal complement of $H_0^1(\Omega)$ in $H^1(\Omega)$, we can regard $\mathbb{H}$ as a subspace of $H_0^1(\Omega)^\perp$, i.e. all functions in $\mathbb{H}$ can be assumed to be (weak) solutions of the linear, elliptic PDE

$$-\Delta u + u = 0 \text{ in } \Omega.$$

In this case, the two projections π_i, $i = 1, 2$, would be orthogonal projections. This perspective, however, is not important, and, in fact, it is of no practical interest.

We would like to consider the optimal control problem

$$\text{Minimize in } U \in \mathbb{H} : \quad I(U) = \int_\Omega \phi(\boldsymbol{x}, u(\boldsymbol{x}), \nabla u(\boldsymbol{x}))\, d\boldsymbol{x} \tag{1.1}$$

where u is the unique minimizer of the problem

$$\text{Minimize in } v \in H^1(\Omega) : \quad \int_\Omega \varphi(\boldsymbol{x}, v(\boldsymbol{x}), \nabla v(\boldsymbol{x}))\, d\boldsymbol{x}$$

under the fixed Dirichlet boundary condition

$$v = U \text{ on } \partial\Omega.$$

The operation taking

$$U \in \mathbb{H} \mapsto u \in H^1(\Omega),$$

can, under suitable hypotheses, be also written in the form

$$\begin{gathered} -\operatorname{div}[\varphi_{\mathbf{u}}(\boldsymbol{x}, u(\boldsymbol{x}), \nabla u(\boldsymbol{x}))] + \varphi_u(\boldsymbol{x}, u(\boldsymbol{x}), \nabla u(\boldsymbol{x})) = 0 \text{ in } \Omega, \\ u = U \text{ on } \partial\Omega. \end{gathered} \tag{1.2}$$

Note that $u \in \mathbb{L}$. Here

$$\begin{aligned} \phi(\boldsymbol{x}, u, \mathbf{u}) &: \Omega \times \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}, \\ \varphi(\boldsymbol{x}, u, \mathbf{u}) &: \Omega \times \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}, \end{aligned}$$

are densities enjoying typical conditions to ensure that problem (1.2) admits a unique solution u for each feasible $U \in \mathbb{H}$.

Though other possibilities can be considered, our main explicit example is the one examined in [8]. Our framework here is a generalization of that principal situation. Let $w \in H^1(\Omega)$ be a fixed, non-constant harmonic function in Ω

$$\Delta w = 0 \text{ in } \Omega,$$

and consider

$$\mathbb{H} = \{\psi(w) : \psi(w) \in H^1(\Omega)\},$$

for suitable functions $\psi(x)$ of a single variable. As a matter of fact, if we put

$$\tilde{\mathbb{H}} = \{\psi, \text{ measurable} : \int_J \psi(\lambda)^2 \omega_0(\lambda)\, d\lambda + \int_J \psi'(\lambda)^2 \omega_1(\lambda)\, d\lambda < \infty\}, \tag{1.3}$$

where:

1. $J = w(\Omega) \subset \mathbb{R}$, an interval (recall that w is harmonic);
2. ω_0 is the weight given by

$$\omega_0(\lambda) = \mathcal{H}^{N-1}(\{w = \lambda\} \cap \Omega) = \int_{\{w=\lambda\}\cap\Omega} 1\, dS(\boldsymbol{x});$$

3. ω_1 is the weight determined by

$$\omega_1(\lambda) = \int_{\{w=\lambda\}\cap\Omega} |\nabla w(\boldsymbol{x})|^2\, dS(\boldsymbol{x}),$$

then

$$\mathbb{H} = \{\psi(w) : \psi \in \tilde{\mathbb{H}}\}.$$

$\tilde{\mathbb{H}}$ is a weighted Sobolev space, and hence $\mathbb{H}$ is closed in $H^1(\Omega)$. The fact that the intersection

$$\mathbb{H} \cap H_0^1(\Omega)$$

is the trivial function is a consequence of the fact that the interval

$$J = w(\partial\Omega)$$

as well, given that w is harmonic (maximum principle). See [8] for more details.

There are several scenarios of increasing complexity that we can look at depending on the form of the two densities ϕ and φ.

1. One first, main paradigmatic example would be

$$\phi(\boldsymbol{x}, u, \mathbf{u}) = \frac{1}{2}|\mathbf{u}|^2 + \frac{1}{2}u^2, \quad \varphi(\boldsymbol{x}, u, \mathbf{u}) = \frac{1}{2}|\mathbf{u}|^2 - f(\boldsymbol{x})u,$$

   for a function $f \in L^2(\Omega)$, where both densities are quadratic in the variable $\mathbf{u}$.

2. We can allow general quadratic dependences on $\mathbf{u}$ for both ϕ and φ.

3. More general, non-quadratic dependence in $\mathbf{u}$ can be implemented in ϕ, but φ has to remain quadratic in $\mathbf{u}$.

4. If we take $\phi \equiv \varphi$, then the dependence on $\mathbf{u}$ does not have to be quadratic. This very particular situation in which $\phi \equiv \varphi$ is especially relevant. It was the one examined in [8].

5. If φ is non-quadratic, the problem is in need of relaxation. This is well-beyond the scope of this paper.

We plan to cover several steps in the analysis of this kind of problems:

1. We will make precise the functional analytical framework and hypotheses on the various ingredients to have a precise, well-defined problem.

2. We will show existence and, whenever appropriate, uniqueness of optimal solutions.

3. We will move on to explore optimality conditions.

4. Some simple ideas about the practical numerical implementation based on optimality will be indicated, as we plan to pursue the numerical approximation of some selected cases through a descent mechanism in the near future.

We will treat these various issues successively in subsequent sections. As our emphasis is on admissible boundary data, other ingredients are typical of a standard optimal control problem, and can be found in references like [2, 3, 6, 7, 9], or the pioneering work [5]. For standard material on Sobolev spaces, check [1] or [4].

2. Motivation

From the purely mathematical viewpoint, it is interesting to consider and study the proposed situation as one can learn new things that eventually may help in maturing old ideas.

From a more applied or practical perspective, one would rather discretize appropriately the boundary of the physical domain Ω, and allow for certain families of boundary data that are constant on the discretized subdomains, for example. These families are incorporated into the subspace $\mathbb{H}$, which would be finite-dimensional in this situation.

Another possibility is that there be parts of the boundary $\partial\Omega$ where the boundary datum must be an unknown constant. This is the idea behind the boundary datum of the form

$$u = \psi(w)$$

for a given fixed function w: points in $\partial\Omega$ in the same w-level will be assigned the same, unknown value at the boundary for feasible u's.

Another important motivation comes from inverse problems in conductivity in the 3D case. This was the main reason in [8] to consider variational problems of the form

$$\text{Minimize in } u \in \mathbb{L}: \quad \int_\Omega \phi(\boldsymbol{x}, u(\boldsymbol{x}), \nabla u(\boldsymbol{x}))\, d\boldsymbol{x}$$

that correspond to the equal-case $\phi \equiv \varphi$. The most natural way to approximate numerically these minimizers goes through the investigation of the family of boundary optimal control problems examined here. In fact, this numerical approximation has been one main point in pursuing this more general scenario.

3. Existence and uniqueness

Assume that subspace $\mathbb{H}$ of $H^1(\Omega)$ has been given with the properties indicated earlier. It is the class of boundary conditions that we are willing to admit in our optimization problem (it could be even finite-dimensional). The two densities ϕ and φ are given. For simplicity, we will take

$$\varphi = \varphi(\boldsymbol{x}, \mathbf{u}), \quad \phi = \phi(\boldsymbol{x}, u, \mathbf{u}),$$

as the more general situation does not pose any particular difficulty under suitable sets of hypotheses. We assume to begin with that $\varphi(\boldsymbol{x}, \mathbf{u})$ is quadratic, strictly convex in the variable $\mathbf{u}$, as usual, so that state equation (1.2)

$$\operatorname{div}[\varphi_{\mathbf{u}}(\boldsymbol{x}, \nabla u(\boldsymbol{x}))] = 0 \text{ in } \Omega, \quad u = U \text{ on } \partial\Omega, \tag{3.1}$$

is linear, well-posed and admits a unique solution. We therefore face the problem

$$\text{Minimize in } U \in \mathbb{H}: \quad I(U) = \int_\Omega \phi(\boldsymbol{x}, u(x), \nabla u(\boldsymbol{x}))\, d\boldsymbol{x} \tag{3.2}$$

subject to (3.1).

If, in addition to φ being quadratic in $\mathbf{u}$, so is density ϕ in $(u, \mathbf{u})$, the dependence of functional I in (3.2) on U is quadratic and coercive, and as such, beyond any other consideration, there is a unique optimal solution in $\mathbb{H}$. We are looking for a more general existence result.

Theorem 3.1 *In addition to the assumptions already indicated, if the density $\phi(\boldsymbol{x}, u, \mathbf{u})$ is convex in $\mathbf{u}$, and with quadratic growth at infinity with respect to pairs $(u, \mathbf{u})$, problem* (3.2) *admits an optimal solution $U \in \mathbb{H}$. If, in addition, ϕ is strictly convex in $(u, \mathbf{u})$ the optimal solution is unique.*

Proof Let $\{U_j\}$ be a minimizing sequence for the problem with $\{u_j\}$, the sequence of their corresponding states

$$\operatorname{div}[\varphi_{\mathbf{u}}(\boldsymbol{x}, \nabla u_j(\boldsymbol{x}))] = 0 \text{ in } \Omega, \quad u_j = U_j \text{ on } \partial\Omega.$$

By the coercivity of $\phi(\boldsymbol{x}, u, \mathbf{u})$ with respect to $(u, \mathbf{u})$, there is a weak limit (for a non-relabeled subsequence)

$$u_j \rightharpoonup u \text{ in } H^1(\Omega), \quad u_j \to u \text{ in } L^2(\Omega).$$

The decomposition

$$u_j = U_j + v_j, \quad v_j = u_j - U_j \in H^1_0(\Omega),$$

implies that

$$U_j = \pi_1 u_j, \quad v_j = \pi_2 u_j,$$

and both sequences $\{U_j\}$, and $\{v_j\}$, are bounded in $H^1(\Omega)$. For additional subsequences (not relabeled), we also have

$$U_j \rightharpoonup U, \quad v_j \rightharpoonup v$$

in $H^1(\Omega)$ with

$$U \in \mathbb{H}, \quad v \in H^1_0(\Omega)$$

(because both subspaces are weakly closed). By the uniqueness of limit, it ought to be true that

$$u = U + v, \quad U \in \mathbb{H}, \quad v \in H^1_0(\Omega),$$

and $u \in \mathbb{L}$. We claim that this weak limit u is indeed the state associated with the weak limit $U \in \mathbb{H}$. To have this crucial fact for our analysis is unavoidable to rely on the linearity of the state equation (3.1). Once this is granted, then it is immediate to have that (3.1) indeed holds for u and U, our weak limits. Then, the convexity of ϕ on $\mathbf{u}$ and its continuity with respect to u imply that

$$\begin{aligned} I(U) &= \int_\Omega \phi(\boldsymbol{x}, u(\boldsymbol{x}), \nabla u(\boldsymbol{x}))\, d\boldsymbol{x} \\ &\le \lim_{j\to\infty} \int_\Omega \phi(\boldsymbol{x}, u_j(\boldsymbol{x}), \nabla u_j(\boldsymbol{x}))\, d\boldsymbol{x} \\ &= \lim_{j\to\infty} I(U_j), \end{aligned}$$

and U is indeed a minimizer. The uniqueness under strict convexity is standard. □

The situation where ϕ and φ are the same densities is so special that it deserves a separate statement. In this case, the optimal control problem becomes

$$\text{Minimize in } u \in \mathbb{L} = \mathbb{H} \oplus H^1_0(\Omega) : \quad \int_\Omega \phi(\boldsymbol{x}, u(\boldsymbol{x}), \nabla u(\boldsymbol{x}))\, d\boldsymbol{x}. \tag{3.3}$$

Proposition 3.2 *Suppose $\phi(\boldsymbol{x}, u, \mathbf{u})$ satisfies the hypotheses in the previous theorem. If $\varphi \equiv \phi$, then problem* (3.3) *admits a solution. If convexity with respect to $\mathbf{u}$ is strengthened to strict convexity with respect to pairs $(u, \mathbf{u})$, the optimal solution is unique.*

This is the particular situation considered in [8]. Exactly the same ideas as in the proof of the previous statement lead to the existence of optimal solutions for the equal-case.

4. Optimality

As usual, we can write optimality conditions not necessarily linked to the available existence results. We therefore assume that problem (1.1) under (1.2) admits an optimal solution $U_0 \in \mathbb{H}$, and we will assume the necessary smoothness and regularity conditions on the two functions ϕ and φ for the following manipulations to be valid. We recover here the full dependence

$$\phi = \phi(\boldsymbol{x}, u, \mathbf{u}), \quad \varphi = \varphi(\boldsymbol{x}, u, \mathbf{u}).$$

Let $U_0 \in \mathbb{H}$ be a feasible element of our problem, and take $U \in \mathbb{H}$ a feasible variation such that the linear combination $U_0 + rU$ is feasible for every r. Put

$$\begin{gathered} -\operatorname{div}[\varphi_{\mathbf{u}}(\boldsymbol{x}, u_0(\boldsymbol{x}), \nabla u_0(\boldsymbol{x}))] + \varphi_u(\boldsymbol{x}, u_0(\boldsymbol{x}), \nabla u_0(\boldsymbol{x})) = 0 \text{ in } \Omega, \\ u_0 = U_0 \text{ on } \partial\Omega, \end{gathered}$$

and let u be the variation produced on u_0 due to the change that U produces on U_0, so that to first-order in r, we would have

$$\begin{gathered} -\operatorname{div}[\varphi_{\mathbf{u}}(\boldsymbol{x}, u_0(\boldsymbol{x}) + ru(\boldsymbol{x}), \nabla u_0(\boldsymbol{x}) + r\nabla u(\boldsymbol{x})] \\ +\varphi_u(\boldsymbol{x}, u_0(\boldsymbol{x}) + ru(\boldsymbol{x}), \nabla u_0(\boldsymbol{x}) + r\nabla u(\boldsymbol{x})) = 0 \text{ in } \Omega. \end{gathered}$$

Differentiating with respect to r, and evaluating at $r = 0$, we see that

$$\begin{gathered} -\operatorname{div}[\varphi_{\mathbf{u}u}(\boldsymbol{x}, u_0, \nabla u_0)u] - \operatorname{div}[\varphi_{\mathbf{uu}}(\boldsymbol{x}, u_0, \nabla u_0)\nabla u] \\ +\varphi_{uu}(\boldsymbol{x}, u_0, \nabla u_0(\boldsymbol{x}))u + \varphi_{u\mathbf{u}}(\boldsymbol{x}, u_0, \nabla u_0(\boldsymbol{x}))\nabla u = 0 \text{ in } \Omega, \end{gathered}$$

which can be simplified to

$$\begin{gathered} -\operatorname{div}[\varphi_{\mathbf{uu}}(\boldsymbol{x}, u_0, \nabla u_0)\nabla u] + [\varphi_{uu}(\boldsymbol{x}, u_0, \nabla u_0(\boldsymbol{x})) \\ -\operatorname{div}\varphi_{\mathbf{u}u}(\boldsymbol{x}, u_0, \nabla u_0)]u = 0 \text{ in } \Omega. \end{gathered} \tag{4.1}$$

The above is a short-form of the true calculation which would require to write

$$\begin{gathered} -\operatorname{div}[\varphi_{\mathbf{u}}(\boldsymbol{x}, u(\boldsymbol{x}, r), \nabla u(\boldsymbol{x}, r)] + \varphi_u(\boldsymbol{x}, u(\boldsymbol{x}, r), \nabla u(\boldsymbol{x}, r)) = 0 \text{ in } \Omega, \\ u(\cdot, r) = U_0 + rU \text{ on } \partial\Omega, \end{gathered}$$

with $u(\cdot, 0) = u_0$. Differentiation with respect to r, and later evaluation at $r = 0$, leads to (4.1) for $u = u_r(\cdot, 0)$ together with the boundary datum $u = U$ on $\partial\Omega$. Since the resulting problem for u is a linear, non-degenerate problem, we can conclude, through the implicit function theorem, that the computations performed are correct, and (4.1) along with the boundary datum $u = U$ on $\partial\Omega$ furnishes the correct perturbation u to u_0 produced by the perturbation U to U_0.

The same perturbation calculations with the cost functional pushes us to express the rate of change

$$\int_\Omega [\phi_u(\boldsymbol{x}, u_0, \nabla u_0)u + \phi_{\mathbf{u}}(\boldsymbol{x}, u_0, \nabla u_0)\nabla u]\, d\boldsymbol{x} \tag{4.2}$$

in terms of our variable ψ with the help of the co-state v. This is standard. The co-state v must be the unique solution of the problem

$$\begin{gathered} -\operatorname{div}[\nabla v \mathbf{A}_0(\boldsymbol{x}) + \mathbf{b}_0(\boldsymbol{x})] + a_0(\boldsymbol{x})v + b_0(\boldsymbol{x}) = 0 \text{ in } \Omega, \\ v = 0 \text{ on } \partial\Omega, \end{gathered} \tag{4.3}$$

where

$$\begin{gathered} \mathbf{A}_0(\boldsymbol{x}) = \varphi_{\mathbf{uu}}(\boldsymbol{x}, u_0(\boldsymbol{x}), \nabla u_0(\boldsymbol{x})), \\ \mathbf{b}_0(\boldsymbol{x}) = \phi_{\mathbf{u}}(\boldsymbol{x}, u_0(\boldsymbol{x}), \nabla u_0(\boldsymbol{x})), \\ a_0(\boldsymbol{x}) = \varphi_{uu}(\boldsymbol{x}, u_0(\boldsymbol{x}), \nabla u_0(\boldsymbol{x})) - \operatorname{div}\varphi_{\mathbf{u}u}(\boldsymbol{x}, u_0(\boldsymbol{x}), \nabla u_0(\boldsymbol{x})), \\ b_0(\boldsymbol{x}) = \phi_u(\boldsymbol{x}, u_0(\boldsymbol{x}), \nabla u_0(\boldsymbol{x}). \end{gathered}$$

Problem (4.3) can, equivalently written in the more standard form

$$\begin{gathered} -\operatorname{div}[\mathbf{A}_0(\boldsymbol{x})^T \nabla v + \mathbf{b}_0(\boldsymbol{x})] + a_0(\boldsymbol{x})v + b_0(\boldsymbol{x}) = 0 \text{ in } \Omega, \\ v = 0 \text{ on } \partial\Omega, \end{gathered} \tag{4.4}$$

where $\mathbf{A}_0(\boldsymbol{x})^T$ stands for the transpose matrix. Then it is a typical exercise to write (4.2) in the form

$$\int_{\partial\Omega} u(\mathbf{A}_0^T\nabla v + \mathbf{b}_0)\cdot\mathbf{n}\, dS(\boldsymbol{x}), \tag{4.5}$$

where we have taken into account (4.1) and (4.4). $\mathbf{n}$ is the unit, outer normal to $\partial\Omega$. The point is that now we have to respect the boundary form of feasible u's which is $u = U$. Indeed, (4.5) is written

$$\int_{\partial\Omega} U\,(\mathbf{A}_0^T\nabla v + \mathbf{b}_0)\cdot\mathbf{n}\, dS(\boldsymbol{x}). \tag{4.6}$$

In this generality, without specifying further the nature of $\mathbb{H}$, not much more can be said. If we put

$$\mathbb{H}|_{\partial\Omega} = \{\, U|_{\partial\Omega} : U \in \mathbb{H}\} \subset L^2(\partial\Omega), \tag{4.7}$$

then if all the integrals in (4.6) vanish, then

$$(\mathbf{A}_0^T\nabla v + \mathbf{b}_0)\cdot\mathbf{n} \in \mathbb{H}|_{\partial\Omega}^{\perp}$$

where $\perp$ means orthogonal complement in the Hilbert space $L^2(\partial\Omega)$. We have shown the following.

Theorem 4.1 *Suppose the integrand*

$$\varphi(\boldsymbol{x}, u, \mathbf{u}) : \Omega\times\mathbb{R}\times\mathbb{R}^N \to \mathbb{R}$$

is C^2 in pairs $(u, \mathbf{u})$, and

$$\phi(\boldsymbol{x}, u, \mathbf{u}) : \Omega\times\mathbb{R}\times\mathbb{R}^N \to \mathbb{R},$$

is C^1 in pairs $(u, \mathbf{u})$. If u_0 is a minimizer for our corresponding optimal control problem, then

$$(\mathbf{A}_0^T\nabla v_0 + \mathbf{b}_0)\cdot\mathbf{n} \in \mathbb{H}|_{\partial\Omega}^{\perp}$$

if $\mathbb{H}|_{\partial\Omega}$ is the subspace in (4.7)*, and orthogonality is interpreted in $L^2(\partial\Omega)$. Here $\mathbf{n}$ is the unit, outer normal to $\partial\Omega$,*

$$\mathbf{A}_0(\boldsymbol{x}) = \varphi_{\mathbf{uu}}(\boldsymbol{x}, u_0(\boldsymbol{x}), \nabla u_0(\boldsymbol{x})), \quad \mathbf{b}_0(\boldsymbol{x}) = \phi_{\mathbf{u}}(\boldsymbol{x}, u_0(\boldsymbol{x}), \nabla u_0(\boldsymbol{x})),$$

and v_0 is the associated co-state solution of (4.3) *for the additional functions*

$$a_0(\boldsymbol{x}) = \varphi_{uu}(\boldsymbol{x}, u_0(\boldsymbol{x}), \nabla u_0(\boldsymbol{x})) - \operatorname{div}\varphi_{\mathbf{u}u}(\boldsymbol{x}, u_0(\boldsymbol{x}), \nabla u_0(\boldsymbol{x})),$$
$$b_0(\boldsymbol{x}) = \phi_u(\boldsymbol{x}, u_0(\boldsymbol{x}), \nabla u_0(\boldsymbol{x}).$$

The converse is correct, as usual, in two main circumstances:

1. under convexity conditions for the integrand ϕ, and convex, quadratic for φ;
2. just under convexity conditions for the integrand ϕ, provided ϕ and φ are identical.

For our main explicit example

$$\mathbb{H} = \{\psi(w) : \psi(w) \in H^1(\Omega)\}, \tag{4.8}$$

with w a fixed, harmonic function in Ω, $\psi \in \tilde{\mathbb{H}}$ given in (1.3), and

$$J = w(\Omega) = w(\partial\Omega) \subset \mathbb{R},$$

our optimality result Theorem 4.1 can be much more explicit. All computations until (4.5) do not require to take into account the particular nature of $\mathbb{H}$. If we go back to (4.6) with $U = \psi(w)$, since the function w is given, and cannot be changed, to factor out our variable functions ψ from the integral in (4.6), we write it in the form

$$\int_J \int_{\{w=\lambda\}\cap\partial\Omega} \psi(w)\,(\mathbf{A}_0^T\nabla v + \mathbf{b}_0)\cdot\mathbf{n}\ dS|_{\{w=\lambda\}}\ d\lambda,$$

where

$$dS = dS|_{\{w=\lambda\}}\ d\lambda.$$

Hence, the previous integral can be finally written in the form

$$\int_J \psi(\lambda) \int_{\{w=\lambda\}\cap\partial\Omega} (\mathbf{A}_0^T\nabla v + \mathbf{b}_0)\cdot\mathbf{n}\ dS|_{\{w=\lambda\}}\ d\lambda. \tag{4.9}$$

If $U_0 = \psi_0(w)$ is indeed a minimizer of our problem, all of these directional derivatives should vanish, i.e. integrals in (4.9) ought to vanish for arbitrary ψ, and this leads immediately to the following corollary.

Corollary 4.2 *With the same notation and hypotheses as in Theorem* 4.1*, if u_0 is a minimizer for our corresponding optimal control problem with $\mathbb{H}$ given in* (4.8)*, then*

$$\int_{\{w=\lambda\}\cap\partial\Omega} (\mathbf{A}_0^T \nabla v_0 + \mathbf{b}_0)\cdot \mathbf{n}\ dS|_{\{w=\lambda\}\cap\partial\Omega} = 0$$

for every $\lambda \in J$.

5. Some ideas on the numerical approximation

To be specific, as it is usually required in practical approximation mechanisms, we will focus directly in our main situation (4.8). We become interested in designing a practical numerical procedure to approximate the unique optimal solution of the optimization problem

$$\text{Minimize in } \psi(x): \quad I(\psi) = \int_\Omega \phi(\boldsymbol{x}, u(\boldsymbol{x}), \nabla u(\boldsymbol{x}))\, d\boldsymbol{x} \tag{5.1}$$

where u is the unique solution of the PDE problem

$$\begin{gathered} -\operatorname{div}[\varphi_{\mathbf{u}}(\boldsymbol{x}, u(\boldsymbol{x}), \nabla u(\boldsymbol{x}))] + \varphi_u(\boldsymbol{x}, u(\boldsymbol{x}), \nabla u(\boldsymbol{x})) = 0 \text{ in } \Omega, \\ u = \psi(w) \text{ on } \partial\Omega. \end{gathered}$$

We assume all of the necessary assumptions, as examined in the previous sections, to justify the manipulations that follow. We are trying to implement a typical descent mechanism. The main step of such a procedure is to determine the steepest descent direction, which we treat in the sequel. The way in which this descent direction is defined is so special that its practical implementation is far from standard, and it is not clear how to use typical software packages for it. We will address this important issue in the near future.

By taking advantage of the calculations performed earlier about optimality, we can formally write

$$\langle I'(\psi_0), \psi\rangle = \int_J \psi(\lambda) \int_{\{w=\lambda\}\cap\partial\Omega} (\mathbf{A}_0^T \nabla v_0 + \mathbf{b}_0)\cdot \mathbf{n}\ dS|_{\{w=\lambda\}}\ d\lambda$$

where v_0 is the unique solution of

$$\begin{gathered} -\operatorname{div}[\mathbf{A}_0^T(\boldsymbol{x})\nabla v_0 + \mathbf{b}_0(\boldsymbol{x})] + a_0(\boldsymbol{x})v_0 + b_0(\boldsymbol{x}) = 0 \text{ in } \Omega, \\ v_0 = 0 \text{ on } \partial\Omega, \end{gathered}$$

and

$$\begin{gathered} \mathbf{A}_0(\boldsymbol{x}) = \varphi_{\mathbf{uu}}(\boldsymbol{x}, u_0(\boldsymbol{x}), \nabla u_0(\boldsymbol{x})), \\ \mathbf{b}_0(\boldsymbol{x}) = \phi_{\mathbf{u}}(\boldsymbol{x}, u_0(\boldsymbol{x}), \nabla u_0(\boldsymbol{x})), \\ a_0(\boldsymbol{x}) = \varphi_{uu}(\boldsymbol{x}, u_0(\boldsymbol{x}), \nabla u_0(\boldsymbol{x})) - \operatorname{div}\varphi_{\mathbf{u}u}(\boldsymbol{x}, u_0(\boldsymbol{x}), \nabla u_0(\boldsymbol{x})), \\ b_0(\boldsymbol{x}) = \phi_u(\boldsymbol{x}, u_0(\boldsymbol{x}), \nabla u_0(\boldsymbol{x}). \end{gathered}$$

Here

$$\begin{gathered} -\operatorname{div}[\varphi_{\mathbf{u}}(\boldsymbol{x}, u_0(\boldsymbol{x}), \nabla u_0(\boldsymbol{x}))] + \varphi_u(\boldsymbol{x}, u_0(\boldsymbol{x}), \nabla u_0(\boldsymbol{x})) = 0 \text{ in } \Omega, \\ u_0 = \psi_0(w) \text{ on } \partial\Omega, \end{gathered}$$

If we put

$$\Psi_0(\lambda) = \int_{\{w=\lambda\}\cap\partial\Omega} (\mathbf{A}_0^T \nabla v_0 + \mathbf{b}_0)\cdot \mathbf{n}\ dS|_{\{w=\lambda\}},$$

then the steepest descent direction will be the minimizer of the quadratic variational problem

$$\text{Minimize in } \psi: \quad \frac{1}{2}\|\psi\|^2 + \langle I'(\psi_0), \psi\rangle.$$

If we put

$$\|\psi\|^2 = \int_J [\psi(\lambda)^2 + \psi'(\lambda)^2]\, d\lambda, \quad \langle I'(\psi_0), \psi\rangle = \int_J \psi(\lambda)\Psi_0(\lambda)\, d\lambda,$$

it is easy to realize that the steepest descent direction is the unique solution of the one-dimensional problem

$$-\psi'' + \psi = -\Psi_0 \text{ in } J, \quad \psi'(\partial J) = 0. \tag{5.2}$$

Note that the local decay of I in (5.1) along this direction ψ trivially becomes

$$\int_J \psi(\lambda)\Psi_0(\lambda)\,d\lambda = -\|\psi\|^2.$$

In addition, at a point of minimum Ψ_0, identically vanishes, and so does the steepest descent direction ψ in (5.2). These simple calculations are the basis for a practical steepest descent method.

Acknowledgements

The author was partially supported by grants PID2020-116207GB-I00, and SBPLY/19/180501/000110.

References

[1] Brezis, H., Functional analysis, Sobolev spaces and partial differential equations. Universitext. Springer, New York, 2011.

[2] Casas, E., Mateos, M., Optimal control of partial differential equations. Computational mathematics, numerical analysis and applications, 3–59, SEMA SIMAI Springer Ser., 13, Springer, Cham, 2017.

[3] Lasiecka, I., Triggiani, R., Control theory for partial differential equations: continuous and approximation theories. I. Abstract parabolic systems. Encyclopedia of Mathematics and its Applications, 74. Cambridge University Press, Cambridge, 2000.

[4] Leoni, G., A first course in Sobolev spaces. Second edition. Graduate Studies in Mathematics, 181. American Mathematical Society, Providence, RI, 2017.

[5] Lions, J.-L., Optimal control of systems governed by partial differential equations. Translated from the French by S. K. Mitter Die Grundlehren der mathematischen Wissenschaften, Band 170 Springer-Verlag, New York-Berlin 1971.

[6] Manzoni, A., Quarteroni, A., Salsa, S., Optimal control of partial differential equations—analysis, approximation, and applications. Applied Mathematical Sciences, 207. Springer, Cham, [2021], ©2021.

[7] Martínez-Frutos, J., Periago Esparza, F., Optimal control of PDEs under uncertainty. An introduction with application to optimal shape design of structures. SpringerBriefs in Mathematics. BCAM SpringerBriefs. Springer, Cham; BCAM Basque Center for Applied Mathematics, Bilbao, 2018.

[8] Pedregal, P., On a new type of boundary condition. Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Mat. RACSAM 116 (2022), no. 1, Paper No. 43, 14 pp.

[9] Tröltzsch, F., Optimal control of partial differential equations. Theory, methods and applications. Translated from the 2005 German original by Jürgen Sprekels. Graduate Studies in Mathematics, 112. American Mathematical Society, Providence, RI, 2010.

An approach to Reduced Basis Large Eddy Simulation turbulence models based upon Kolmogorov's equilibrium turbulence theory

Cristina Caravaca García[1], Tomás Chacón Rebollo[1,2], Enrique Delgado Ávila[1], Macarena Gómez Mármol[1]
1. Departamento de Ecuaciones Diferenciales y Análisis Numérico, Universidad de Sevilla, Spain
2. Instituto de Matemáticas de la Universidad de Sevilla, Spain

Abstract

In this work we present a certified Reduced Basis for the unsteady Smagorinsky turbulence model, for which we introduce an *a posteriori* error indicator based upon the Kolmogorov turbulence theory, which introduces an expression for the energy cascade. The main idea of this estimator is that if the full order solution and the Reduced Order solution are close enough, then their flow energy spectrum within the inertial range should also be close. For this *a posteriori* error indicator, we also present some numerical tests which supports that the use of this indicator is helpful. We use as full-order model a finite element discretisation of the LES Smagorinsky model, and the Reduced Order Model, considering the inner pressure *supremizer*, for obtaining stable reduced velocity-pressure spaces.

1. Introduction

Reduced Order Modeling (ROM) has been successfully used in several fields to provide large reduction in computation times to solve Partial Differential Equations [9–11]. In fluid mechanics a popular strategy is to use POD to extract the dominant structures for high-Reynolds flow, which are then used in a Galerkin approximation of the underlying equations [10,12]. Application of the POD-Galerkin strategy to turbulent fluid flows remains a challenging area of research. By construction, ROMs generated using only the first most energetic POD basis functions are not endowed with the dissipative mechanisms associated to the creation of lower size, and less energetic, turbulent scales.

In this work we present an unsteady Reduced Basis (RB) Smagorinsky turbulence model. We address a turbulence model to consider realistic situations, as turbulent flows frequently appear in actual applications. The Smagorinsky model is a basic Large Eddy Simulation (LES) turbulence model, that provides accurate solutions for the large scales of the flow, and a part of the inertial spectrum. We address the *a posteriori* error analysis - based reduced order modeling of incompressible flow equations.

We present an *a posteriori* error indicator, based upon the Kolmogorov turbulence theory. Since the full-order model is intended to be a good approximation of the continuous problem, it should accurately approximate the energy spectrum of the continuous problem in the resolved part of the inertial spectrum. The key is to use the error deviation with respect to the full-order energy spectrum by the RB solution to select the new basis functions by the Greedy algorithm. To validate this indicator, we develop an academic test in which we compare the use of the Kolmogorov indicator as error estimator with the use of the exact error between the full order and reduced order solution, for the selection of the basis functions in the Greedy algorithm. This error estimation procedure has the advantage of applying to any kind of numerical discretisation, and to any physical time at which the turbulence is in statistical equilibrium. This allows to overcome the technical difficulties related to the building of error estimation for the Reduced Basis discretisation, as the ones developed in [3, 4, 6], based upon the Brezzi-Rappaz-Raviart theory [2].

Moreover, we use the Empirical Interpolation Method (EIM) to build reduced approximations of the non-linear eddy viscosity term, coming from the Smagorinsky model. This allows to linearize the eddy viscosity term, being able to efficiently decouple the RB problem in an offline/online procedure.

The structure of this work is as follows. In Section 2, we present the Finite Element (FE) discretization for the Smagorinsky turbulence model. Then, in Section 3 we present the Reduced Basis method for this problem. In Section 4, we present the linearization of the Smagorinsky term, recalling the Empirical Interpolation Method. We present the development of the *a posteriori* error indicator based upon the Kolmogorov turbulence theory in section 5, with the numerical test related in this part presented in ssection 6. Finally, in section 7, we present some conclusions.

2. Smagorinsky Finite element problem

In this section we present the unsteady Smagorinsky turbulence model, that is the basic LES turbulence model, in which the effect of the subgrid scales on the resolved scales is modeled by eddy diffusion terms [5]. We introduce a discretization by the Finite Element method using inf-sup stable velocity-pressure spaces.

Let Ω be a bounded domain of $\mathbb{R}^d$ ($d = 2, 3$), with Lipschitz-continuous boundary Γ. We assume that Γ is split into $\Gamma = \Gamma_D \cup \Gamma_N$ where Γ_D and Γ_N are two connected measurable sets of positive $(d-1)$-dimensional measure, with disjoint interiors. We intend to impose Dirichlet and Neumann boundary conditions on Γ_D and Γ_N, respectively.

We present a parametric unsteady Smagorinsky turbulence model, where we consider the Reynolds number as a physical parameter, denoted by $\mu \in \mathcal{D}$, where $\mathcal{D}$ is a compact sub-set of $\mathbb{R}$. Also, we consider the time interval $[0, T_f]$, with $T_f > 0$ a chosen finite time. Although the Smagorinsky model is intrinsically discrete, we present it in a continuous form in order to clarify its relationship with the Navier-Stokes equations: We search for a velocity field $\mathbf{w} : \Omega \times [0, T_f] \mapsto \mathbb{R}^d$ and a pressure function $p : \Omega \times [0, T_f] \mapsto \mathbb{R}$ such that

$$\begin{cases} \partial_t \mathbf{w} + \mathbf{w} \cdot \nabla \mathbf{w} + \nabla p - \nabla \cdot \left(\dfrac{1}{\mu} \nabla \mathbf{w} + \nu_T(\mathbf{w}) \nabla \mathbf{w} \right) = \mathbf{f} & \text{in } \Omega \times [0, T_f], \\ \nabla \cdot \mathbf{w} = \mathbf{0} & \text{in } \Omega \times [0, T_f], \\ \mathbf{w} = 0 & \text{on } \Gamma_D \times [0, T_f], \\ \mathbf{n} \cdot \left(\dfrac{1}{\mu} \nabla \mathbf{w} + \nu_T(\mathbf{w}) \nabla \mathbf{w} \right) = 0 & \text{on } \Gamma_N \times [0, T_f], \\ \mathbf{w} = 0 & \text{in } \Omega \times \{0\} \end{cases} \tag{2.1}$$

where $\mathbf{f}$ is the kinetic momentum source, $\nu_T(\mathbf{w})$ is the eddy viscosity defined as

$$\nu_T(\mathbf{w}) = C_S^2 \sum_{K \in \mathcal{T}_h} h_K^2 \left| \nabla \mathbf{w}_{|K} \right| \chi_K, \tag{2.2}$$

where $|\cdot|$ denotes the Frobenius norm in $\mathbb{R}^{d \times d}$, and C_S is the Smagorinsky constant.

To state the full-order discretization that we consider for problem (2.1), let us introduce the velocity and pressure spaces

$$Y = \{ \mathbf{v} \in H^1(\Omega)^d, \text{ s.t. } \mathbf{v}_{|\Gamma_D} = \mathbf{0} \}, \quad M = \{ q \in L^2(\Omega), \text{ s.t. } \int_\Omega q = 0 \}.$$

We assume $\mathbf{f} \in Y'$.

Let Y_h and M_h be two finite element subspaces of Y and M, respectively, that satisfy the discrete inf-sup condition, i.e.,

$$\|q_h\|_{0,2,\Omega} = \sup_{\mathbf{v}_h \in Y_h} \frac{(q_h, \nabla \cdot \mathbf{v}_h)_\Omega}{\|\nabla \mathbf{v}_h\|_{0,2,\Omega}}, \quad \forall q_h \in M_h, \tag{2.3}$$

We introduce the discretisation of the unsteady Smagorinsky model (2.1),

$$\begin{cases} \forall \mu \in \mathcal{D} \text{ and } t \in [0, T_f], \text{find } (\mathbf{u}_h(t), p_h(t)) = (\mathbf{u}_h(t; \mu), p_h(t; \mu)) \in Y_h \times M_h \text{ such that} \\ (\partial_t \mathbf{u}_h, \mathbf{v}_h)_\Omega + a(\mathbf{u}_h, \mathbf{v}_h; \mu) + b(\mathbf{v}_h, p_h; \mu) + a_S(\mathbf{w}_h; \mathbf{w}_h, \mathbf{v}_h; \mu) + c(\mathbf{u}_h, \mathbf{u}_h, \mathbf{v}_h; \mu) = \langle \mathbf{f}, \mathbf{v}_h \rangle & \forall \mathbf{v}_h \in Y_h, \\ b(\mathbf{u}_h, q_h; \mu) = 0 & \forall q_h \in M_h, \end{cases} \tag{2.4}$$

where the bilinear forms $a(\cdot, \cdot; \mu)$ and $b(\cdot, \cdot; \mu)$ are defined as

$$a(\mathbf{u}, \mathbf{v}; \mu) = \frac{1}{\mu} \int_\Omega \nabla \mathbf{u} : \nabla \mathbf{v} \, d\Omega, \qquad b(\mathbf{v}, q; \mu) = -\int_\Omega (\nabla \cdot \mathbf{v}) q \, d\Omega; \tag{2.5}$$

while the trilinear form $c(\cdot, \cdot, \cdot; \mu)$ is defined as

$$c(\mathbf{z}, \mathbf{u}, \mathbf{v}; \mu) = \frac{1}{2} \left[\int_\Omega (\mathbf{z} \cdot \nabla \mathbf{u}) \mathbf{v} \, d\Omega - \int_\Omega (\mathbf{z} \cdot \nabla \mathbf{v}) \mathbf{u} \, d\Omega \right]. \tag{2.6}$$

Moreover, the non-linear form $a_S(\cdot; \cdot, \cdot; \mu)$, is a Smagorinksy modelling for the eddy viscosity term, and it is given by

$$a_S(\mathbf{z}; \mathbf{u}, \mathbf{v}; \mu) = \int_\Omega \nu_T(\mathbf{z}) \, \nabla \mathbf{u} : \nabla \mathbf{v} \, d\Omega. \tag{2.7}$$

3. Smagorinsky Reduced basis problem

In this section we introduce the Reduced Basis (RB) model for problem (2.1). The RB problem reads

$$\begin{cases} \text{Find } (\mathbf{u}_N, p_N) \in Y_N \times M_N \text{ such that} \\ a(\mathbf{u}_N, \mathbf{v}_N; \mu) + b(\mathbf{v}_N, p_N; \mu) + a_S(\mathbf{w}_N; \mathbf{w}_N, \mathbf{v}_N; \mu) + c(\mathbf{u}_N, \mathbf{u}_N, \mathbf{v}_N; \mu) = \langle \mathbf{f}, \mathbf{v}_N \rangle & \forall \mathbf{v}_N \in Y_N, \\ b(\mathbf{u}_N, q_N; \mu) = 0 & \forall q_N \in M_N. \end{cases} \tag{3.1}$$

Here, we denote by Y_N the reduced velocity space, and by M_N the reduced pressure space. Their dimensions are intended to be much smaller than their finite element counterparts Y_h and M_h. The computation of the reduced spaces is done through a POD+Greedy approach [8]. We follow a POD strategy considering the time as a parameter, and the Greedy algorithm for the Reynolds number physical parameter. For the POD, we use a separate strategy in the sense that we apply the POD to a velocity correlation matrix $\mathbb{S}^u$, and also to a pressure correlation matrix $\mathbb{S}^p$. To build the correlation matrices, it is necessary to establish spatial norms for velocity and pressure, since the time should be considered as a parameter. In this case, we use the H^1-seminorm, and L^2-norm for the pressure, for the spaces Y_h and M_h, respectively.

In order to guarantee the inf-sup condition (2.3) for the reduced spaces, we need to introduce the inner pressure *supremizer*, defined by

$$\left(\nabla T_p^\mu q_h, \nabla \mathbf{v}_h\right)_\Omega = b(q_h, \mathbf{v}_h; \mu) \quad \forall \mathbf{v}_h \in Y_h. \tag{3.2}$$

In [13], G. Stabile and G. Rozza propose two different strategies to implement the *supremizer* when POD is applied over a parameter. We use here the *exact supremizer enrichment*, this is, we compute the *supremizer* basis from the pressure basis obtained in the POD procedure. We consider Algorithm 1 for the construction of reduced spaces.

Algorithm 1 POD+Greedy with *supremizer*

Set $\epsilon_{1,tol}, \epsilon_{2,tol} > 0$, $N_{max} \in \mathbb{N}$, $\mu^* \in \mathcal{D}_{train}$, $\mathbb{Z}^u = [\,]$, $\mathbb{Z}^p = [\,]$ and $S = \{\,\}$;
while $N < N_{\max}$ **do**
  $S = S \cup \{\mu^*\}$;
  Compute $U_h^n(\mu^*) = (\mathbf{u}_h^n(\mu^*), p_h^n(\mu^*))$ for $n = 1, \ldots, L$;
  Build $\mathbb{S}^u = [\underline{\mathbf{u}}_h^1(\mu^*), \underline{\mathbf{u}}_h^2(\mu^*), \ldots, \underline{\mathbf{u}}_h^L(\mu^*)]$, $\mathbb{S}^p = [\underline{p}_h^1(\mu^*), \underline{p}_h^2(\mu^*), \ldots, \underline{p}_h^L(\mu^*)]$;
  $[\underline{\xi}_1^u, \ldots, \underline{\xi}_{M^u}^u] = \text{POD}(\mathbb{S}^u, \epsilon_{1,tol})$;
  $[\underline{\xi}_1^p, \ldots, \underline{\xi}_{M^p}^p] = \text{POD}(\mathbb{S}^p, \epsilon_{1,tol})$;
  $\mathbb{Z}^u = [\mathbb{Z}^u, \underline{\xi}_1^u, \ldots, \underline{\xi}_{M^u}^u]$;
  $\mathbb{Z}^p = [\mathbb{Z}^p, \underline{\xi}_1^p, \ldots, \underline{\xi}_{M^p}^p]$;
  $[\underline{\varphi}_1^u, \ldots, \underline{\varphi}_{N^u}^u] = \text{POD}(\mathbb{Z}^u, \epsilon_{2,tol})$;
  $[\underline{\varphi}_1^p, \ldots, \underline{\varphi}_{N^p}^p] = \text{POD}(\mathbb{Z}^p, \epsilon_{2,tol})$;
  Compute $\varphi_{N^u+i}^u = T_p^\mu \varphi_i^p$ for $i = 1, \ldots, N^p$;
  $N = N^u + 2N^p$;
  $Y_N = \{\varphi_i^u\}_{i=1}^{N^u+N^p}$, $M_N = \{\varphi_i^p\}_{i=1}^{N^p}$,
  $\mu^* = \arg\max_{\mu \in \mathcal{D}train} \Delta_N(\mu, L)$;
  $\epsilon_N = \Delta_N(\mu^*, L)$;
  if $\epsilon_N \leq \epsilon_{tol}$ **then**
    $N_{max} = N$;
  end if
end while

We summarize in the following Algorithm 1:

1. For a given μ^*, we solve the Smagorinsky Model for any time t_n for $n = 1, \ldots, L$, and we save the result in the snapshot matrices $\mathbb{S}^u$ for velocity and $\mathbb{S}^p$ for pressure.
2. We apply the POD procedure separately for velocity and pressure, for a given tolerance $\varepsilon_{1,tol}$ and we add the results in the matrices $\mathbb{Z}^u$ and $\mathbb{Z}^p$.
3. Finally, we apply a second POD to $\mathbb{Z}^u$ and $\mathbb{Z}^p$ for a given tolerance $\varepsilon_{2,tol}$. This procedure avoid repetition in the basis.
4. We compute the *supremizer* for the pressure basis resulting above and we add it to the velocity basis, obtaining Y_N and M_N.

5. Lastly, we apply the Greedy Algorithm to the RB problem associated to the spaces Y_N and M_N, obtaining the new parameter μ^*. We use the estimate $\Delta_N(\mu)$ at the last time since it is assumable that the energy spectrum is well-developed at that time.

4. Approximation of eddy viscosity term and pressure stabilizing coefficient

In this section, we present the approximation of the non-linear terms with respect to the parameter, throughout the Empirical Interpolation Method [1, 7]. The Smagorinsky eddy-diffusion term defined in (2.2), $\nu_T(\nabla \mathbf{w}) := \nu_T(\mu)$, is a non-linear function of the parameter, and consequently needs to be linearised with the EIM to build the RB model, in order to reduce the on-line computation times.

For this purpose, we build a reduced-basis space $W_M^S = \{q_1^S(\mu), \dots, q_{M_1}^S(\mu)\}$ by a greedy procedure selection. In this case, we consider the time also as a parameter jointly with the Reynolds number. Thus, we approximate the non-linear Smagorinsky term by the following trilinear form:

$$a_S(\mathbf{w}_N; \mathbf{w}_N, \mathbf{v}_N; \mu) \approx \hat{a}_S(\mathbf{w}_N, \mathbf{v}_N; \mu), \tag{4.1}$$

where,

$$\hat{a}_S(\mathbf{w}_N, \mathbf{v}_N; \mu) = \sum_{k=1}^{M_1} \sigma_k^S(\mu) s(q_k^S, \mathbf{w}_N, \mathbf{v}_N), \tag{4.2}$$

with,

$$s(q_k^S, \mathbf{w}_N, \mathbf{v}_N) = \sum_{K \in \mathcal{T}_h} \left(q_k^S \nabla \mathbf{w}_N, \nabla \mathbf{v}_N\right)_K, \tag{4.3}$$

In practise, we solve problem (3.1) considering $\hat{a}_S(\cdot, \cdot; \mu)$ instead of $a_S(\cdot; \cdot, \cdot; \mu)$.

5. *A posteriori* error indicator based upon Kolmogorov's theory

The aim of this section is to introduce an *a posteriori* error indicator for the selection of the parameter in the Greedy procedure. This indicator is based upon the Kolmogorov turbulence theory, which introduces an expression for the energy cascade. The main idea of this indicator is that a trial solution is accurate if its energy spectrum is close to the theoretical $k^{-5/3}$ spectrum predicted by the Kolmogorov theory.

Andrei Kolmogorov stated that under certain hypothesis, there exists an inertial range $[k_1, k_2]$ where the energy spectrum $E(k)$ can be expressed by the wavenumber k and the turbulent dissipation ε, this is, $E(k) = C\varepsilon^{2/3}k^{-5/3}$, k_1 and k_2 two wavenumbers associated to the largest inertial scale of the flow and the smaller scale under which the viscosity effects take place, respectively.

By considering the Smagorinsky model (2.7), we are solving the scales in some inertial subrange $[k_1, k_c]$, with $k_c = \delta^{-1}$, where δ is the mesh size. The mesh $\mathcal{T}_h$ should be carefully chosen in order to solve a part of the inertial range, this is, $k_c \in [k_1, k_2]$. We then assume that within the inertial sub-range $[k_1, k_c]$, the energy spectrum has the expression

$$E(k, \mu) = \alpha(\mu) k^{-5/3}, \tag{5.1}$$

where $\alpha(\mu) > 0$ depends on the turbulent dissipation ε.

Thus, let $E_N(k; \mu)$ be the energy spectrum associated to $\mathbf{u}_N(\mu)$, solution of (3.1). We define an *a posteriori* error indicator as follows

$$\Delta_N(\mu) = \min_{\alpha} \left(\int_{k_1}^{k_c} |E_N(k; \mu) - \alpha(\mu) k^{-5/3} dk \right)^{(1/2)}. \tag{5.2}$$

This *a posteriori* indicator measures how close is a given solution (either ROM or FOM-obtained) to the theoretical Kolmogorov spectrum, in the range of inertial wavenumbers $[k_1, k_c]$ which is solved by the Smagorinsky model.

6. Numerical results

In this section, we solve the Smagorinsky model (2.4) for 2D flows, in the time interval $t \in [0, 30]$, over the unit square $\Omega = [-1/2, 1/2]^2$ with periodic boundary conditions. We do not consider any source, thus $\mathbf{f} = 0$. We select the Reynolds number, μ, as the parameter, ranging on $\mathcal{D} = [1000, 16000]$. We consider a structured mesh, where we divide each edge in $\mathcal{N} = 64$ intervals, obtaining a mesh with 8192 triangles and 4225 vertices. We consider the inf-sup stable Taylor-Hood Finite Element, i.e., $\mathbb{P}2 - \mathbb{P}1$ Finite Element for velocity-pressure discretization. We consider a Crank-Nicolson scheme for the time derivative discretization.

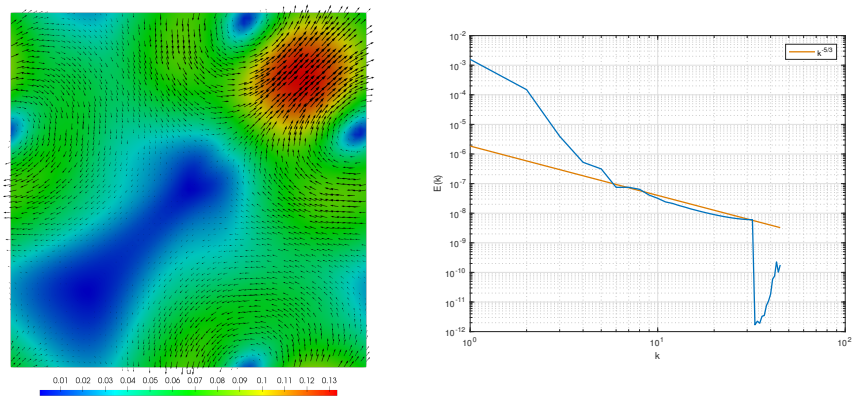
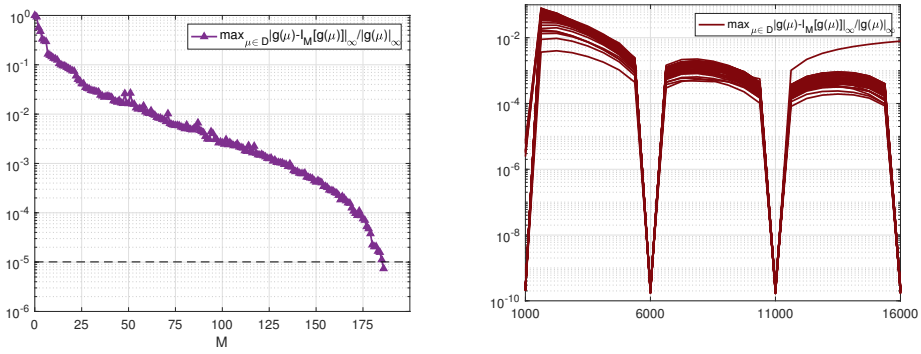


Fig. 1 Initial condition $\mathbf{u}_h^0$ and its energy spectrum.



(a) EIM history convergence error.

(b) EIM error, each line correspond to a time step t_n for $n = 1, \ldots, L$.

Fig. 2 EIM applied to unsteady Smagorinsky Model.

To determine the initial condition, we look for a velocity field with an inertial energy spectrum as in (5.1). We consider a velocity field $\mathbf{w}_h^0 = (v, v)$, where v is defined through its Fourier transform:

$$\hat{v}(k) = \begin{cases} k^{-(5/3+1)/2} & \text{if } 0 < k \leq \mathcal{N}/4, \\ 0 & \text{other case.} \end{cases} \tag{6.1}$$

To determine the initial condition for problem (2.4), we solve the Smagorinsky model taking $\mathbf{w}_h^0$ as the initial state for $\mu = 8500$, the intermediate Reynolds number, and we take as initial condition the velocity field for $t = 15$.

In Figure 1, we show the initial condition $\mathbf{u}_h^0$ and its energy spectrum. For wavenumbers between $k_1 = 5$ and $k_c = 32$, we obtain a good approximation of the inertial spectrum. For $k > 32$, we observe an abrupt decay of the energy. This is produced by the wavenumbers that are out of the circle of the largest radio inside the unit square and the viscous effects. As we can see, we start from a well-developed inertial energy spectrum.

As we mentioned in section 4, we need to linearise the Smagorinsky eddy viscosity term (2.2), with respect to the parameter. For this purpose, we compute the finite element solution $(\mathbf{u}_h^n(\mu), p_h^n(\mu))$ for all $n = 1, \ldots, L$, $\mu = \{1\,000, 6\,000, 11\,000, 16\,000\}$ and we apply the EIM to compute the approximation of the eddy viscosity function. We stop the algorithm on 186 basis functions when the error is below $\epsilon_{EIM} = 10^{-5}$. The convergence error is shown in Figure 2a, while in Figure 2b, we show the error, where each line represents a time step t_l for $n = 1, \ldots, 48$.

We compare the use of the estimate $\Delta_N(\mu)$ introduced in (5.2) versus the use of the exact error at the final time, $T_f = 30$

$$\varepsilon_N(\mu) = \|\mathbf{u}_h(\mu) - \mathbf{u}_N(\mu)\|_{0,2,\Omega}, \tag{6.2}$$

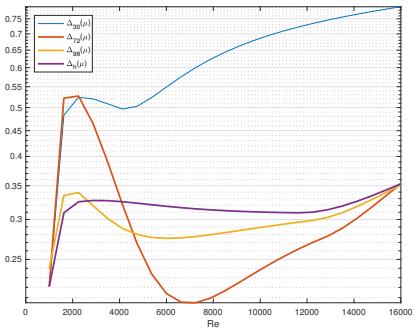
for the parameter selection.

In Table 1, we show the comparison using the indicator $\Delta_N(\mu)$ (left table) and the exact error $\varepsilon_N(\mu)$ (right table) for the selection of the parameter μ. Using $\Delta_N(\mu)$ for the stopping criteria, we stop the algorithm in the third iteration, since the next Reynolds number has been already selected and we remain with the same number of basis functions, $N = 98$ in Table 1. The exact error and the number of RB basis are similar if we use either the indicator

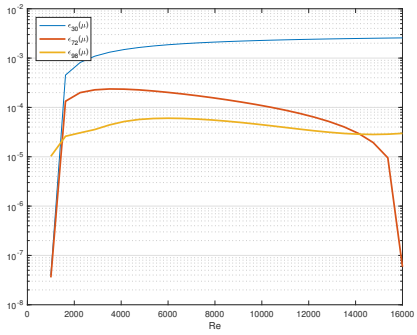
It.	μ	N	$\max_{\mu} \Delta_N(\mu)$	$\max_{\mu} \varepsilon_N(\mu)$
1	1 000	30	$7.92 \cdot 10^{-1}$	$2.57 \cdot 10^{-3}$
2	16 000	72	$5.27 \cdot 10^{-1}$	$2.37 \cdot 10^{-4}$
3	2 250	**98**	$3.51 \cdot 10^{-1}$	$\mathbf{6.04 \cdot 10^{-5}}$
4	16 000	98		

It.	μ	N	$\max_{\mu} \varepsilon_N(\mu)$
1	1 000	30	$2.57 \cdot 10^{-3}$
2	16 000	72	$2.37 \cdot 10^{-4}$
3	3 500	**100**	$\mathbf{3.52 \cdot 10^{-5}}$
4	7 250	119	$3.53 \cdot 10^{-5}$

Tab. 1 Step by step of the POD+Greedy algorithm, using $\Delta_N(\mu)$ (left table) and $\varepsilon_N(\mu)$ (right table) for the parameter selection.



(a) Estimate $\Delta_N(\mu)$ in each iteration and $\Delta_h(\mu)$.



(b) Error $\epsilon_N(\mu)$ in each iteration.

Fig. 3 Convergence of POD+Greedy algorithm.

$\Delta_N(\mu)$ or the exact error $\varepsilon_N(\mu)$, which supports that the use of the indicator $\Delta_N(\mu)$ is quasi-optimum. Notice that a relatively small decrease of the indicator Δ_N produces a much larger decrease of the error ε_N.

In Figure 3a, we show the comparison of the indicator $\Delta_N(\mu)$ at each iteration of the POD+Greedy algorithm described in Table 1 versus

$$\Delta_h(\mu) = \min_{\alpha} \left(\int_{k_1}^{k_c} |E_h(k;\mu) - \alpha(\mu) k^{-5/3}|^2 \, dk \right)^{1/2}$$

and $E_h(k;\mu)$ for $k \in (k, k_c)$ represents the energy spectrum of $\mathbf{u}_h(\mu)$.

Since the reduced solution is built from the FE approximation, we should not expect that $\Delta_N(\mu)$ tends to 0 when $N \to \infty$, it should rather converge to Δ_h. We observe in Figure 3a that indeed $\Delta_N(\mu)$ approaches Δ_h as N increases and Δ_h is not zero as the finite element solution is just an approximation of the physical flow, with some error with respect to the theoretical inertial spectrum in $k^{-5/3}$. The best that we can expect in the ROM is to reproduce the approximation to the theoretical inertial spectrum given by the FOM solution. In Figure 3b, we show the error $\varepsilon_N(\mu)$ for $\mu = \{1\,000, 1\,625, \ldots, 16\,000\}$ at each POD+Greedy algorithm iteration. In the last iteration, the error is smaller than 10^{-4}.

In Table 2, we show the computational time, the values of the estimates, and the error between the RB and FE solution at the final time $T_f = 15$.

	$\mu = 1\,825$	$\mu = 4\,804$	$\mu = 11\,757$	$\mu = 13\,605$	$\mu = 14\,027$
T_{FE}	55.63s	58.67s	58.3s	58.09s	57.94s
T_{RB}	2.95s	2.99s	2.83s	2.92s	3.06s
Speedup	19	20	21	20	19
$\Delta_N(\mu)$	$3.42 \cdot 10^{-1}$	$2.8 \cdot 10^{-1}$	$2.97 \cdot 10^{-1}$	$3.11 \cdot 10^{-1}$	$3.17 \cdot 10^{-1}$
$\Delta_h(\mu)$	$3.18 \cdot 10^{-1}$	$3.23 \cdot 10^{-1}$	$3.09 \cdot 10^{-1}$	$3.2 \cdot 10^{-1}$	$3.25 \cdot 10^{-1}$

Tab. 2 Validation of RB model.

We obtain speed-ups ratio close to 20, what is satisfying for an evolution turbulence model. We can observe that the value of the indicator for the full-order solution is near to the value of the indicator for the reduced-order solution. We already saw that considering $\epsilon_N(\mu)$ instead of $\Delta_N(\mu)$ does not significantly reduce the number of basis functions, thus, the indicator yields nearly optimal results.

7. Conclusions

In this work, we have presented an *a posteriori* error indicator, based upon the Kolmogorov turbulence theory for the unsteady Smagorinsky turbulence model, applied in the construction of a Reduced Basis Method by means of a POD+Greedy algorithm. We have validated this indicator with an academic numerical test, in which we have observed that considering the use of the indicator generates reduced spaces that do not substantially differ from those constructed using the exact error instead of the indicator. Actually, the number of basis functions when we consider either the *a posteriori* error indicator and the exact error are quite close.

Acknowledgements

This work has been funded by the Spanish Government Project PID2021-123153OB-C21.

References

[1] Maxime Barrault, Yvon Maday, Ngoc Cuong Nguyen, and Anthony T. Patera. An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations. *C.R. Acad. Sci. Paris Sér. I Math.*, 339:667–672, 2004.

[2] F. Brezzi, J. Rappaz, and P.A. Raviart. Finite dimensional approximation of nonlinear problems. *Numer. Maht.*, 36:1–25, 1980.

[3] Tomás Chacón Rebollo, Enrique Delgado Ávila, and Macarena Gómez Mármol. On a certified VMS-smagorinsky reduced basis model with LPS pressure stabilisation. *Applied Numerical Mathematics*, 185:365–385, 2023.

[4] Tomás Chacón Rebollo, Enrique Delgado Ávila, Macarena Gómez Mármol, Francesco Ballarin, and Gianluigi Rozza. On a certified smagorinsky reduced basis turbulence model. *SIAM Journal on Numerical Analysis*, 55(6):3047–3067, 2017.

[5] Tomás Chacón Rebollo and Roger Lewandowski. *Mathematical and Numerical Foundations of Turbulence Models and Applications*. Springer New York, 2014.

[6] Simone Deparis. Reduced basis error bound computation of parameter-dependent Navier-Stokes equations by the natural norm approach. *SIAM J. Sci. Comput.*, 46(4):2039–2067, 2008.

[7] Martin A. Grepl, Yvon Maday, Ngoc C. Nguyen, and Anthony T. Patera. Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 41(3):575–605, 2007.

[8] Bernard Haasdonk. Convergence rates of the POD–greedy method. *ESAIM: Mathematical Modelling and Numerical Analysis*, 47(3):859–873, apr 2013.

[9] J. S. Hesthaven, G. Rozza, and B. Stamm. *Certified Reduced Basis Methods for Parametrized Partial Differential Equations*. Springer, 2015.

[10] P. Holmes, J. L. Lumley, and G. Berkooz. *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge, 1996.

[11] Alfio Quarteroni, Andrea Manzoni, and Federico Negri. *Reduced Basis Methods for Partial Differential Equations: An Introduction*. Springer, 2015.

[12] L. Sirovich. Turbulence and the dynamics of coherent structures. parts i-iii. *Quart. Appl. Math.*, 45(3):561–590, 1987.

[13] Giovanni Stabile and Gianluigi Rozza. Finite volume POD-galerkin stabilised reduced order methods for the parametrised incompressible navier–stokes equations. *Computers & Fluids*, 173:273–284, sep 2018.

Hopf Bifurcation for a Functional Differential Equation (FDE) with respect to the delay

Juan Francisco Padial[1], **Alfonso Casal**[2]
1. jf.padial@upm.es *Universidad Politécnica de Madrid, Spain*
2. alfonso.casal@upm.es *Universidad Politécnica de Madrid, Spain*

Abstract

Motivated by some facts of the car traffic flow, such as oscillatory behavior in the car–following situations, we build and study a new traffic model. Since its mathematical expression is a Functional (Delayed) Differential Equation, we start with a presentation of this type of equations and the features we need to deal with our model, mainly the Hopf Bifurcation result for these equations.

Keywords Bifurcation, Delay Differential Equations, Oscillatory Solutions,Traffic Models.

1. Introduction

When modelling the evolution of many situations of the real world by using ordinary or partial differential equations, it is implicitly assumed that the future of the system is completely determined by its state in a particular moment, which can be (or taken as) the present. When dealing with many other situations, or refining previous models of some others, it becomes apparent that the model to study the future state of the system must include some of the past states of the system. For those cases, the right mathematical tool is that of the Functional (Delayed) Differential Equations (for several applications, see [9]). In Section 2 we give some of the main definitions related to these equations. We also present, in Section 3, a general theory of Hopf bifurcation for Functional (Delayed) Differential Equations (see Hale [7] and Hale and Lunel [8]). In Section 4, we introduce a traffic model built by Padial and Casal [11] involving both driver and mechanic reaction times. Looking for oscillatory behaviour of solutions of its solutions, it turns out that the delay time play an important role, as the suitable parameter to study the bifurcation. In Section 5, we adapt the Hopf Bifurcation theory introduce by Hale for the case where the bifurcation parameter is the delay time, and we apply this result to our car-following model.

2. Functional Differential Equations

One definition of the Differential Delay Equations (DDE) or Functional Differential Equations (FDE), in one of their proper settings can be found in Hale [7] or Lunel and Hale [8].

Let $\mathbb{R}^n$ the $n-$dimensional linear vector space over the real numbers with euclidean norm $|\cdot|$, $C([a,b]\,,\ \mathbb{R}^n)$ is the Banach space of continuous functions mapping the interval $[a,b]$ into $\mathbb{R}^n$ with the topology of uniform convergence. Consider a given number $\tau \in \mathbb{R}$, $\tau > 0$ and let $C := C([-\tau,0],\mathbb{R}^n)$ and the norm of an element $\phi \in C$ by $\|\phi\| = \sup_{-\tau\le\theta\le 0}\|\phi(\theta)\|$.

Given two real numbers $\sigma, A \ge 0$ and $\mathbf{x} \in C([\sigma-\tau,\sigma+A]\,,\mathbb{R}^n)$, then for any $t \in [\sigma,\sigma+A]$ we define $\mathbf{x}_t(\theta) = \mathbf{x}(t+\theta)$ with $\theta \in [-\tau,0]$. Given $\mathcal{D}$ a subset of $\mathbb{R}\times C$, we consider the functional $\mathbf{F} : \mathcal{D} \subset \mathbb{R}\times C([-\tau,0],\mathbb{R}^n) \to \mathbb{R}^n$ such that

$$\frac{d}{dt^+}\mathbf{x}\,(t) = \mathbf{F}(\tau,\mathbf{x}_t) \tag{2.1}$$

where $\frac{d}{dt^+}$ represents the right–hand derivative, and in the following we denote by $\mathbf{x}' = \frac{d}{dt^+}\mathbf{x}$. We say that the relation (2.1) is a *Retarded Functional Differential Equation* on $\mathcal{D}$, a RFDE associated with $\mathbf{F}$ (we denote by RFDE($\mathbf{F}$) if we need to emphasize the the equation is defined by $\mathbf{F}$).

Definition 2.1 Given $\tau > 0$ and $C := C^2([-\tau,0],\mathbb{R}^n)$, a function $\mathbf{x}$ is said to be a solution of the *Retarded Functional Differential Equation* (2.1) on $\mathcal{D} \subset \mathbb{R}\times C$ for a given functional $\mathbf{F} : \mathcal{D} \to \mathbb{R}^n$, if there are σ and $A \ge 0$, such that $\mathbf{x} \in C^2([\sigma-\tau,\sigma+A]\,,\mathbb{R}^n)$, $(t,\mathbf{x}_t) \in \mathcal{D}$ and $\mathbf{x}$ satisfies equation (2.1) for any $t \in [\sigma,\sigma+A]$.

In the same framework as in Definition 2.1, we introduce the

Definition 2.2 (IVP) For given $\sigma \ge 0$, $\phi \in C$, we say that $\mathbf{x}(\sigma,\phi,\mathbf{F})$ is a *solution of the initial value problem for* the *Retarded Functional Differential Equation* (2.1) on $\mathcal{D} \subset \mathbb{R}\times C$ for a given functional $\mathbf{F} : \mathcal{D} \to \mathbb{R}^n$, *with initial value ϕ at σ*, or a *solution through* (σ,ϕ), if there is a real number $A \ge 0$ such that $\mathbf{x}(\sigma,\phi,\mathbf{F})$ is a solution of equation (2.1) on $[\sigma-\tau,\ \sigma+A)$ and $\mathbf{x}_\sigma(\sigma,\phi,\mathbf{F}) = \phi$.

In this sense, solutions of FDE could be viewed as curves in a Banach space. In the above mentioned references [7] and [8], suitable definitions and results of the fundamental theory, such as those on existence, uniqueness, continuous dependence and differentiability on data and parameters, regularity with respect to initial conditions and continuation are given (in these equations is important to distinguish between forward and backwards continuation). Answers on these questions depend of the regularity of the above function.

3. Hopf bifurcation for FDE

To present the Hopf Bifurcation Theory, we will refer and state the Hopf Bifurcation Theorem as in Hale [7, Theorem 1.1, p. 246].

We recall the formulation of the Hopf bifurcation Theorem 1.1 of [7] for RFDE (2.1). Let $\mathbf{F}$ be of class $k \geq 2, \mathbf{F}(\tau, \mathbf{0}) = \mathbf{0}$ for all $\tau \in \mathbb{R}$, $C := C([-\tau, 0], \mathbb{R}^n)$ and $\mathbf{x}_t(\theta) = \mathbf{x}(t+\theta)$ with $\theta \in [-\tau, 0]$. Define $\mathbf{L} : \mathbb{R} \times C \to \mathbb{R}^n$ by

$$\mathbf{L}(\tau)\psi = \mathbf{F}_\phi(\tau, \mathbf{0})\psi \tag{3.1}$$

with $\psi \in C$, where $\mathbf{F}_\phi(\tau, \mathbf{0})$ is the derivative of $\mathbf{F}(\tau, \phi)$ with respect to $\phi \in C$ at $\phi = 0$ and define

$$f(\tau, \phi) = \mathbf{F}(\tau, \phi) - \mathbf{L}(\tau)\phi.$$

We have to consider also the following hypotheses:

(H1) The linear RFDE($\mathbf{L}(0)$) (that is, $\mathbf{x}' = \mathbf{F}_\phi(0, \mathbf{0})\mathbf{x}_t$) has a simple purely imaginary characteristic root $u_0 = y_0 i \neq 0$ and all characteristic roots $u (u = x + yi)$ are different of $u_0, \overline{u}_0$ (conjugate of u_0) and satisfy $u \neq hu_0$ for any integer h.

By Lemma 2.2 of Section 7.2 of [7], There exist $\tau_0 > 0$ and simple characteristic root $u(\tau)$ of the linear RFDE($\mathbf{L}(\tau)$) (that is, $\mathbf{x}' = \mathbf{F}_\phi(\tau, \mathbf{0})\mathbf{x}_t$) such that has a continuous derivative $u'(\tau)$ for $|\tau| < \tau_0$. Moreover, we assume that

(H2) $\mathrm{Re}(u'(0)) \neq 0$ (*transversality condition*).

We introduce the additional notation of [7], to make the statement of the result more specific. By taking τ_0 sufficiently small, we may assume $\mathrm{Im}u(\tau) \neq 0$ for $|\tau| < \tau_0$ and obtain a function $\phi_\tau \in C$ which is continuously differentiable in τ and allows to define a basis for the solutions of the RFDE($\mathbf{L}(\tau)$) corresponding to $u(\tau)$. The functions

$$\Phi_\tau := (\mathrm{Re}\phi_\tau, \mathrm{Im}\phi_\tau)$$

form a corresponding basis for the characteristic roots $u_0(\tau), \overline{u}_0(\tau)$. Similarly, a basis Ψ_τ for the adjoint equation can be obtained, with $\langle \Psi_\tau, \Phi_\tau \rangle = I$. Decomposing C by $(u_0(\tau), \overline{u}_0(\tau))$ as $C = \mathcal{P}_\tau \oplus Q_\tau$, then Φ_τ is a basis for $\mathcal{P}_\tau$. We know that

$$\Phi_\tau(\theta) = \Phi_\tau(0) \exp B(\tau)\theta, \quad -\tau \leq \theta \leq 0,$$

and the eigenvalues of the 2×2 matrix $B(\tau)$ are $u_0(\tau)$ and $\overline{u}_0(\tau)$. By a change of coordinates and maybe redefining the parameter τ we may assume that

$$B(\tau) = y_0 B_0 + \tau B_1(\tau)$$

with

$$B_0 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 1 & \gamma(\tau) \\ -\gamma(\tau) & 1 \end{pmatrix}$$

where $\gamma(\tau)$ is continuously differentiable on $0 \leq |\tau| < \tau_0$. We can now state the Hopf bifurcation theorem and we refer to the conclusions stated in this theorem as a *Hopf Bifurcation*.

Theorem 3.1 *Suppose $F(\tau, \phi)$ has continuous first derivatives with respect to τ, ϕ, $F(\tau, 0) = 0$ for all τ and Hypothesis (H1) and (H2) are satisfied. Then there are constants $a_0 > 0$, $\tau_0 > 0$, $\delta_0 > 0$, functions $\tau(a) \in \mathbb{R}$, $\omega(a) \in \mathbb{R}$, and an $\omega(a)$ −periodic function $x^*(a)$, with all functions being continuously differentiable in a for $|a| < a_0$, such that $x^*(a)$ is a solution of equation (2.1) with*

$$x_0^*(a)^{P_\tau} = \Phi_{\tau(a)} y^*(a), \qquad x_0^*(a)^{Q_\tau} = z_0^*(a)$$

where $y^*(a) = \mathrm{col}(a, 0) + o(|a|)$, $z_0^*(a) = o(|a|)$ as $|a| \longrightarrow 0$. Furthermore, for $|\tau| < \tau_0$, $|\omega - (2\pi/y_0)| < \delta_0$, every ω−periodic solution of equation (2.1) with $\|x_t\| < \delta_0$ must be of the above type except for a translation in phase.

4. A car–following model

Our investigation on the impact of the reaction delay enhances phenomenological insights into the emergence and evolution of traffic congestion. In traffic, the phenomenon of stop–and–go waves (known as a start–stop waves, or a 'phantom jam') has been empirically studied by many authors. Some studies have shown that a change in driver's sensitivity (for instance, a sudden acceleration or deceleration) can lead to such oscillatory behaviour.

When dealing with the delay equations arising in car–following problems, and in order to obtain more refined models, it seems convenient to consider the delay as the total effect of different causes (see [13] among others). One of them is the time between drivers' perception of changes or hazards ahead, becoming aware of them and acting accordingly. Another part of the reaction time would be the processing of the information, and the actions taken on the mechanisms of the car.

Concerning the driver, there has been an evolution of models characterized by increasingly sophisticated assumptions about the driver behaviour. There are further models taking into account several other features, some of them still open, as those considering the comparison of the driver's estimation of spacing with the actual spacing, or more realistic ones, where the delay may be dependent of the state of the system, such as relative positions or velocities In some other fields this situation has been considered, as in [1] and in the references therein, mainly [9].

By taking the delay as central in the analysis of the change of stability of the solutions, we can also analyse the time–delayed traffic dynamics in a control problem perspective. In several works, within a large family of problems, the delay is a tool for control (see e.g. [2], [3], [13]).

Some previous car–following models built from a follow–the–leader General Model (GM). Let us consider a car following another, under the following assumptions: the cars flow on a single lane, if t is the time (assuming that the initial time $t_0 = 0$), for the following car at the time t, $X_1(t)$ is its position, $X_1'(t)$ is its velocity, and $X_1''(t)$ is its acceleration. The driver of the following car adjust his speed with respect to the speed of the leading car, as before, at the time t, $X_0(t)$ is its position, $X_0'(t)$ is its velocity, and $X_0''(t)$ is its acceleration. There exist a time lag τ reaction of the following car to the actions of the leading car. As said above,

$$\tau = \tau_d + \tau_m$$

with τ_d is the reaction delay of the driver and τ_m is the mechanic delay of his car.

A first analysis is to consider that the acceleration (or deceleration) of following car is proportional to the perceived difference with respect to car leading car Mathematically,

$$X_1''(t+\tau) = \left(X_0'(t) - X_1'(t)\right) u.$$

where t is the time, $X_0(t)$ is the position of the leading car and $X_1(t)$ is that of the following car at the instant t. X_i' and X_i'' are the velocity and the acceleration respectively. We assume that the velocity of leading car (i. e. $X_0'(t)$) is a given positive constant v_0. This is the velocity that the following car want to reach, and keeping a safety distance m.

Now we can distinguish if u is a constant or a non-constant function depending, for instance, of the distance between cars, the relative velocity, etc. (e.g. $u = u\left(X_0, X_1, X_1'\right)$).

When u is a *constant function*, the car–following model is linear. In this case, the acceleration (response) is directly proportional to the relative velocity (stimulus) (see [4]).

When u is a *non-constant function*, the car–following model is non-linear. Several relationship can be considered, usually a certain power (to be determined) of the following car's speed. Moreover, Gazis, Herman and Potts [6] found that the above equation could not quite explain the traffic situation in higher density since, in it, the behaviour of driver that follows does not take into account the relative spacing between cars.

Considering the *separation* between the cars ($s(t) = X_0(t) - X_1(t) = v_0 t - X_1(t)$) and the corresponding *relative velocity* and the *relative acceleration* of the following car ($s'(t) = v_0 - X_1'(t)$ and $s''(t) = -X_1''(t)$), for the car-following model, we introduce the initial value problem

$$s''(t+\tau) = -g\left(s'(t), s(t)\right), \quad t \geq 0 \tag{4.1}$$

$$s(t) = s_0(t), \; s'(t)\,\phi(\theta) = \phi_0(\theta), \quad \theta \in [-\tau, 0] \tag{4.2}$$

where the real function $g(s, s')$ with $(s, s') \in \mathbb{R}^2$ satisfying some conditions, and the functions $\phi(\theta) := \begin{pmatrix} s \\ s' \end{pmatrix}(\theta)$ and $\phi_0(\theta) := \begin{pmatrix} s_0 \\ s_0' \end{pmatrix}(\theta)$ are regular enough (see Padial and Casal [11]).

The equation (4.1) of this model is a *Functional Differential Equation* (FDE).

The model needs that the real function $g(s, s')$ satisfies some conditions:

i) For any v_0 there exists an m, minimum recommended distance between cars. We assume that the car following the leader is in an *equilibrium state* when there is no speed difference with the leading car, that is $s'(t) = 0$; and when it follows it at the safe minimum distance $s(t) = m$, thus the velocity of the following car is constant and then the acceleration is zero, that is $g(m, 0) = 0$ at the *equilibrium point*.

ii) There is a maximum acceleration of the following car, $a > 0$, thus $g(s, s') \leq a$, and a maximum deceleration, $b < 0$, thus $b \leq g(s, s')$. So, $b \leq g(s, s') \leq a$ for all $(s, s') \in \mathbb{R}^2$.

iii) If the relative velocity $s'(t) = 0$, then $g(s, 0) > 0$ if $s > m$, $g(s, 0) = 0$ if $s = m$ and $g(s, 0) < 0$ if $s < m$. If the relative velocity $s'(t) > 0$ (increasing the distance between cars), the car that follows would accelerate even when the distance $s(t) < m$. If the relative velocity $s'(t) < 0$ (decreasing the distance between cars), the car that follows would decelerate even when the distance $s(t) > m$.

iv) Moreover, g is increasing with respect to s.

On the other hand, we want to include in the modelling the *temperament of the driver* and that *of the mechanic of its vehicle*, meaning a sort of intensity of the response d (in some references this type of parameters are called *aggressivity* both for the drivers and the cars [12]). We will also take into account a parameter k which determine the driver's intensity of the action according to the safe distance and the minimum relative velocity which the driver is able to perceive (which is not neither the real relative velocity nor the real safe distance, but the driver's perception of them). A suitable choice is to take $k = m/w$, where m is the safe minimum distance and w is the relative velocity that the driver can perceive.

As a convenient functions satisfying the above conditions, we consider a class of *sigmoidal functions*. In particular, we will take a function g of the distance between the cars s and of the speed difference s':

$$g(s, s') = a - \frac{(a+b)}{1 + \frac{b}{a} e^{d(s-m+ks')}}, \quad \forall (s, s') \in \mathbb{R}^2. \tag{4.3}$$

Under the regularity conditions, the results of existence, uniqueness and continuous dependence on the initial data and forward continuation are fulfilled [8] for the FDE problem (4.1) and (4.2).

5. Time delay as bifurcation parameter

Our particular car–following modelling has led us to a FDE in which the delay τ is a parameter including, namely, the different reaction times corresponding to the drivers, to the mechanic of the cars and to some others related to them. Our interest is to give an explanation on how, in this case, the structure of the solutions can change, from constant to oscillatory solutions, when the delay parameter varies. We will show that theses changes of structure can be described as a Hopf bifurcation phenomenon. To do that, we will write the second order delay differential equation (4.1) in the form of first order delay system (2.1). As usual, we introduce two functions $z_1(t) = s(t) - m$ and $z_2(t) = s'(t)$ (recall that m is the safety distance). We rescale the time by making $\bar{t} = t + \tau$, and we rename $\bar{t}$ as t obtaining the equivalent system

$$\begin{cases} z_1'(t) &= z_2(t) \\ z_2'(t) &= -g(z_1(t-\tau) + m, z_2(t-\tau)) \end{cases}$$

As before, let us consider the FDE with

$$\mathbf{F} : \mathcal{D} \subset \mathbb{R} \times C^2([-\tau, 0], \mathbb{R}^2) \rightarrow \mathbb{R}^2 \tag{5.1}$$

$$(\sigma, \mathbf{z}_t) \rightarrow \mathbf{F}(\sigma, \mathbf{z}_t) = \left(z_{2_t}, -g\left(z_{1_t} + m, z_{2_t}\right)\right) \tag{5.2}$$

with $\sigma \in \mathbb{R}$, $\mathbf{z} = (z_1, z_2) \in C^2([-\tau, 0], \mathbb{R}^2)$ and $z_{i_t}(t) = z_i(t + \theta)$, $\theta \in [-\sigma, 0]$, $i = 1, 2$.

By the definition of g, notice that $\mathbf{F}(\sigma, \phi)$ has a continuous first and second continuous derivatives in ϕ for all σ real and ϕ in $C := C^2([-\tau, 0], \mathbb{R}^2)$ and $\mathbf{F}(\sigma, \mathbf{0}) = \mathbf{0}$ for all $\sigma > 0$ (notice that $\mathbf{0} = (0, 0)$ is the equilibrium point for (5.2) and $(m, 0)$ is the equilibrium point for (4.1) and that $g(m, 0) = 0$). These properties on $\mathbf{F}$ and the initial condition (4.2) ensure affirmative and convenient answers to the basic fundamental theory [7, Chap. 2].

Theorem 5.1 *For* $\mathbf{F}$ *defined in* (5.2)*, there exists* $\tau_0 > 0$ *such that the problem RFDE* (2.1)*,* (4.2) *for this* τ_0 *has a periodic solution.*

For the proof of the theorem we will use the characteristic equation associated to the FDE equation (4.1).

Taking $x = \sigma - m + k\omega$ and $\tilde{g}(x) = g(\sigma, \omega)$, the McLaurin series in $x = 0$ of $\tilde{g}$ is $\tilde{g}(x) = d\frac{ab}{a+b}x + O\left(x^2\right)$. In particular

$$g(\sigma, \omega) = d\frac{ab}{a+b}(\sigma - m + k\omega) + \text{Higher Order Terms}$$

Considering the McLaurin series of $\tilde{g}(x)$ $(= g(\sigma, \omega))$ in (4.3)

$$\begin{aligned} s''(t+\tau) &= -\tilde{g}(s(t) - m + ks'(t)) + \text{Higher Order Terms} \\ &= -d\tfrac{ab}{a+b}(s(t) - m + ks'(t)) + \text{Higher Order Terms} \end{aligned}$$

Making the change of variables $S = s - m$ and renaming the coefficients, we obtain

$$S''(t+\tau) = -DKS'(t) - DS(t) \tag{5.3}$$

with $D = d\frac{ab}{a+b}$ and $K = k$. The quasi–characteristic equation for this delay equation is

$$w^2 e^{w\tau} + DKw + D = 0. \tag{5.4}$$

Finally, making $u = w\tau$ we obtain $u^2 e^u + DK\tau u + D\tau^2 = 0$. Taking $u = x + iy$, form the above transcendental equation, we obtain the curves

$$\begin{aligned} \left((x^2 - y^2)\cos(y) - 2xy\sin(y)\right)e^x + D\tau^2 + DK\tau x &= 0, \\ \left((x^2 - y^2)\sin(y) + 2xy\cos(y)\right)e^x + DK\tau y &= 0, \end{aligned} \tag{5.5}$$

corresponding to the real and imaginary parts respectively.

Proof (of the Theorem 5.1) For the proof of this result we use the Hopf bifurcation Theorem 3.1. To do that, we need to prove that the hypothesis *(H1)* and *(H2)* are fulfilled.

By the definition (3.1) of $\mathbf{L}$, the simple purely imaginary characteristic root of the RFDE($\mathbf{L}(\tau)$) can be identify with the simple purely imaginary characteristic root of system (5.5).

To prove *(H1)* we compute the roots $u = x + iy$ of the system (5.5). From the Implicit Function Theorem, there exist τ_0 such that $0 < \tau < \tau_0$ there exist solutions for (5.5). Now we obtain the purely imaginary roots $u_0 = y_0 i$, taking $x = 0$. From the last system we obtain that y_0 has to verify the following system

$$\begin{aligned} -y^2\sin(y) + DK\tau y &= 0, \\ -y^2\cos(y) + D\tau^2 &= 0. \end{aligned}$$

Solving in τ, we obtain that

$$y_{0_\pm}(\tau) = \pm\tau\frac{\sqrt{2}}{2}D^{1/2}\left(\sqrt{K^4D^2 + 4} + K^2D\right)^{1/2} \neq 0, \tau > 0.$$

Let $u_{0_\pm} = y_{0_\pm} i$. Thus *(H1)* is fulfilled.

We need to check that the transversality condition *(H2)* holds when $x = 0$. We derivate implicitly the quasi-characteristic equations (5.4). Let $J_1(x, y)$ be the real part of left hand side of (5.4) and $J_2(x, y)$ the imaginary part of left hand side of (5.4) denoting $w = x + iy$. Now the equation (5.4) is

$$J_1(x, y) + J_2(x, y)\, i = 0,$$

with

$$\begin{aligned} J_1(x, y) &= \left(x^2 - y^2\right)e^{x\tau}\cos(y\tau) - 2xye^{x\tau}\sin(y\tau) + DKx + D, \\ J_2(x, y) &= 2xye^{x\tau}\cos(y\tau) + \left(x^2 - y^2\right)e^{x\tau}\sin(y\tau) + DKy \end{aligned}$$

(see (5.5)). To obtain $x' := \frac{d}{d\tau}x(\tau)$ and $y' := \frac{d}{d\tau}y(\tau)$, we derivate the last equation:

$$\begin{aligned} \frac{\partial}{\partial x}J_1(x, y)\,x' + \frac{\partial}{\partial y}J_1(x, y)\,y' &= 0, \\ \frac{\partial}{\partial x}J_2(x, y)\,x' + \frac{\partial}{\partial y}J_2(x, y)\,y' &= 0. \end{aligned}$$

We are interested in non-trivial solutions for this linear system in (x', y'). Thus, we impose the condition that the range of the matrix of system is one. So we compute the determinant of the matrix of the system

$$\Delta(x, y) = \begin{vmatrix} \frac{\partial}{\partial x}J_1(x, y) & \frac{\partial}{\partial y}J_1(x, y) \\ \frac{\partial}{\partial x}J_2(x, y) & \frac{\partial}{\partial y}J_2(x, y) \end{vmatrix}.$$

The hypothesis of transversality (*H2*) is equivalent to verify that the derivative of the real part in the point $(0, y)$ is different from zero. So, we solve $\Delta(0, y) = 0$. Calculating the partial derivatives and substituting above and computing for $x = 0$, we obtain

$$\begin{aligned}
\frac{\partial}{\partial x} J_1(0, y) &= KD - 2y\sin(y\tau) - y^2\tau\cos(y\tau), \\
\frac{\partial}{\partial y} J_1(0, y) &= -2y\cos(y\tau) + y^2\tau\sin(y\tau), \\
\frac{\partial}{\partial x} J_2(0, y) &= 2y\cos(y\tau) - y^2\tau\sin(y\tau), \\
\frac{\partial}{\partial y} J_2(0, y) &= KD - 2y\sin(y\tau) - y^2\tau\cos(y\tau),
\end{aligned}$$

$$\Delta(0, y) = \begin{vmatrix} KD - 2y\sin(y\tau) - y^2\tau\cos(y\tau) & -2y\cos(y\tau) + y^2\tau\sin(y\tau), \\ 2y\cos(y\tau) - y^2\tau\sin(y\tau) & KD - 2y\sin(y\tau) - y^2\tau\cos(y\tau) \end{vmatrix},$$

$$\Delta(0, y) = \Big(KD - 2y\sin(y\tau) - y^2\tau\cos(y\tau)\Big)^2 + \Big(2y\cos(y\tau) - y^2\tau\sin(y\tau)\Big)^2$$

For the values such that Δ is zero the (*H2*) of the Hopf conditions is fulfilled and there are changes of the structure from constant to periodic solutions according with the numerical results. To look for solutions to equation $\Delta(0, y) = 0$, it is equivalent to look for solutions to the following non-linear system

$$\begin{aligned}
KD - 2y\sin(y\tau) - y^2\tau\cos(y\tau) &= 0 \\
2y\cos(y\tau) - y^2\tau\sin(y\tau) &= 0.
\end{aligned}$$

If y and τ verify the above system, we get that $y = +\sqrt{-2 + \sqrt{4 + \tau^2 K^2 D^2}}/\tau$ and calling $z = y\tau$, from the last equation we obtain that $z\tan(z) = 2$ and joint to the relation between y and τ we get a solutions for the system. Then, (*H2*) is fulfilled.

□

We show for some particular values of the parameters, the existence of periodic solution by solving the $\Delta(0, y) = 0$. These test values comes from experimental data and those which allow to improve the graphic

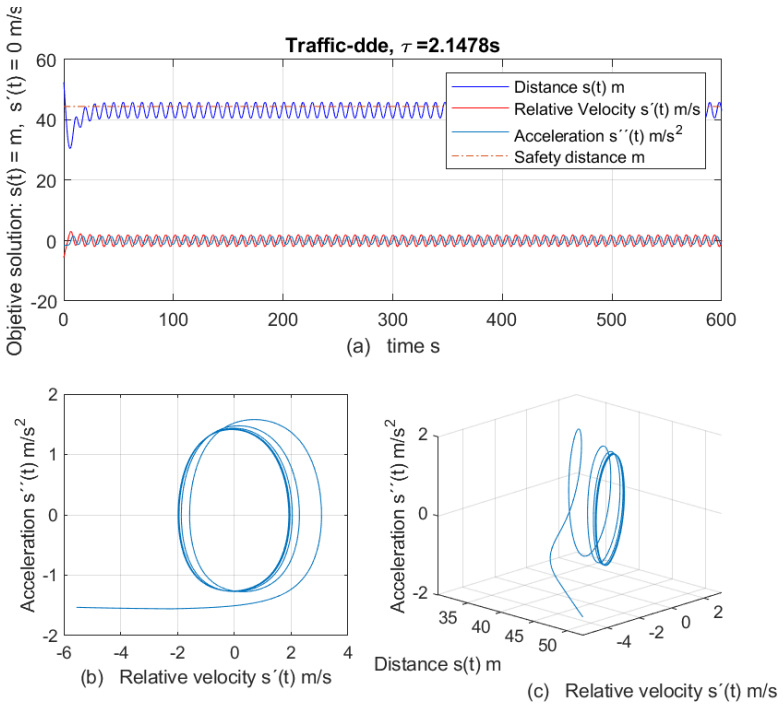


Fig. 1 Delay time $\tau = 2.147780$. A stable periodic solution appears. The follower doesn't get a constant velocity as the leader.

visibility of the behaviour of the solution (see e.g. [5] and its references). We will take the following parameters:

$a = 2.0576\text{m/s}^2$, $b = 1.5677\text{m/s}^2$, $d = 0.1124$, $m = 44.4444\text{m}$, $k = m/w = 11.3890\text{s}$ ($w = 3.9024\text{m/s}$). For these parameter, $K = k = 11.389$ and $D = d\frac{ab}{a+b} = 0.1$. We can obtain numerically that $\Delta(0, 0.501389) \approx 10^{-9}$ for $\tau = 2.147780$, given an oscillatory solution for the FDE.

For to show the numerical solution graphically, we also fix a constant velocity of the leader car, v_0, at 80 km/h = 22.2222 m/s and for the follower, its velocity is 100 km/h = 27.7778 m/s constant in $[-\tau, 0]$. Thus, the relative velocity $s'(\theta) = -5.5556$ m/s for any $\theta \in [-\tau, 0]$. On the other hand, the initial distance between vehicles it was taken at 20m plus the safety distance $m = 44.4444m$. So, the initial function distance will be $s(\theta) = (20 + m) + s'(\theta)\theta = 64.4444 - 5.5556\theta$ for all $\theta \in [-\tau, 0]$. That is

$$\phi_0(\theta) = \begin{pmatrix} 64.4444 - 5.5556\,\theta \\ -5.5556 \end{pmatrix}, \quad \theta \in [-\tau, 0].$$

For these parameters we solve numerically the problem (4.1)-(4.2) by using the code `dde23` of Matlab R2018a. Additionally this procedure allows us to present graphically the appearance (see Fig. 1) of the periodic solution for the problem (4.1)-(4.2) obtained for the bifurcation parameter $\tau = 2.147780$s.

In these phenomena we obtain a transition from a constant distance between two vehicles to another behaviour which is oscillatory and this change can be noticed in the real traffic. This change of the structure, a bifurcation, is due to a delay in the action of the driver and the vehicle.

The simulation studies show interesting effects on the dynamics of solutions of the system associated to the car–following model. In particular, there are changes of structure in its solutions as the delays vary. There are equilibrium states of the system, stable or unstable, and as the delay vary, a given equilibrium, a stable solution, may loss its stability and other equilibria may branch off. Very often this happens to constant solutions which lead to time periodic oscillations.

Acknowledgements

The authors were partially supported by the Spanish Project PID2020-112517GB-I00 of the Ministerio de Ciencia e Innovación (Spain); and Aula Universidad Empresa of the UPM, *BID-Group One on the Quality Culture*.

References

[1] Casal A., Corsi, L., de la Llave, R.: Expansions in the delay of quasi-periodic solutions for state dependent delay equations J. Phys. A, **53** 235202, https://doi.org/10.1088/1751-8121/ab7b9e (2020).

[2] Casal, A., Díaz, J. I.: On the complex Ginzburg–Landau equation with delayed feedback. Mathematical Models and Methods in Applied Sciences, Vol. **16**(1), 1–17 (2006).

[3] Casal, A., Díaz, J. I.: Feedback Delay as a Control Tool: The Complex Ginzburg–Landau Equation with Local and Non-local Delayed Perturbations. Chapter 17th in Recent Trends in Chaotic, Non-linear and Complex Dynamics, World Scientific Series on Non-linear Science Series B: Vol. **19**, https://doi.org/10.1142/9789811221903_0017 (2021).

[4] Chandler, R. E., Herman, R., Montroll, E. W.: Traffic dynamics: studies in car following. Operations Research, Vol. **6**(2), 165–184 (1958).

[5] Gasser, I., Seidel, T., Sirito, G., Werneber, B.: Bifurcation Analysis of a Class of car–following Traffic Models II: Variable Readtion Times and Aggressive Drivers. Bulletin of the Institute of Mathematics, Academia Sinica (New Series), Vol. **2** n. 2, 587-607 (2007).

[6] Gazis, D. C., Herman, R., Potts, R. B.: Car–following theory of steady state traffic flow. Operations Research, **7**(4), 499–595 (1959).

[7] Hale, J. K.: Theory of Functional Differential Equations. Springer (1977).

[8] Hale, J. K., Lunel, S. M. V.: Introduction to Functional Differential Equations. Springer-Verlag (1993).

[9] Hartung, F., Krisztin, T., Walther, H. O., Wu, J.: Handbook of Differential Equations Vol 3, Ordinary Differential Equations: Theory and Applications (Chapter 5, Functional Differential Equations with State-Dependent Delays), A. Cañada P. Drabek A. Fonda, Eds., North Holland (2006).

[10] Padial, J.F.: Uniqueness result of a two–lane car traffic flow model. AIP Conference Proceedings 1978, 350005, https://doi.org/10.1063/1.5043958 (2018).

[11] Padial, J.F., Casal, A. Bifurcation in car-following models with time delays and driver and mechanic sensitivities. Rev. Real Acad. Cienc. Exactas Fis. Nat. Ser. A-Mat. 116–180 (2022). https://doi.org/10.1007/s13398-022-01307-4

[12] Sipahi, R., Atay, F. M., Niculescu, S.-J.: Stability of traffic flow behavior with distributed delays modeling the memory effects of the drivers. Paper 10, http://hdl.handle.net/2047/d10009395 (2007).

[13] Sipahi, R., Niculescu, S-I.: Analytical stability study of a deterministic car following model under multiple delay interactions. 6th IFAC Workshop on Time Delay Systems, TDS 2006, Jul 2006, L'Aquila, Italy 187–192 (2006).

On the Motion of Two Point Masses inside a Homogeneous Cloud

Luis Floría [1]
Departamento de Física Teórica. Universidad de Zaragoza, Spain

Abstract

We consider the problem of motion of two material points within a homogeneous spherical cloud, under the hypotheses that the mutual interactions between these point masses and between the bodies and the particles of the cloud are described by their Newtonian gravitational attraction. The constant density of the cloud is supposed to be sufficiently small so that the resistance of the medium to the motion of those point masses can be neglected.

Under these hypotheses, the problem of relative motion of the said bodies can be recast as a perturbed Keplerian system (Radzievskij's two–body problem), in which the perturbing effects are formalized by a conservative central force. With the help of the first integrals of the angular momentum and the total energy, the standard solution procedure in terms of plane polar coordinates (r, φ) allows the reduction of the problem to a quadrature involving an integral that apparently is Abelian, and thanks to which the orbit equation can be obtained in finite terms.

In this paper some previous analytical investigations on this problem are reviewed, and then we explore other different analytical approaches to the study of this Radzievskij problem (namely Planetary Equations in Gaussian form, some methods of regularisation and linearisation of Keplerian systems, and Hamilton–Jacobi technique).

2020 Mathematics Subject Classification System: 70 F 15, 70 F 05, 70 M 20, 70 H 15, 70 H 20.

Keywords and expressions: homogeneous cloud, gravitational two–body problem, perturbed Keplerian systems, central force, Planetary Equations, regularisation and linearisation, Hamilton–Jacobi Theory.

1. Introduction

Radzievskij, [10], dealt with the problem of motion of two point masses, m_1 and m_2, within a *homogeneous spherical cloud*, assuming that the mutual interactions between these point bodies and between the bodies and the particles of the cloud are described by their *gravitational attraction* according to Newton's Law of Universal Gravitation. In addition to this, the constant *density of the cloud* is supposed to be sufficiently small so that the resistance of the medium to the motion of the bodies can be neglected.

The problem of *relative motion* of one of these particles with respect to the other one can be reformulated as a *perturbed Kepler problem* in which the perturbing force is also a *conservative central force*. Accordingly, this problem of relative motion can be treated as a conservative central–force one–body problem (Boccaletti and Pucacco, [2], Ch. 2, §2.1, pp. 126–131; Goldstein [5], Ch. 3, §3.1–§3.3, pp. 70–82). Consequently this system admits the first integral of the orbital angular momentum vector and that of the total mechanical energy.

On the basis of these considerations, Radzievskij concentrated on the motion within the (fixed) orbital plane, introduced polar coordinates (r, φ) in that plane, and (following the standard solution procedure in polar coordinates starting from the said first integrals) reduced the problem to a quadrature involving an integral that is *(apparently) Abelian*, and thanks to which the orbit equation can be obtained in finite terms.

In this respect we would like to point out that the quadrature that Radzievskij considered to be an Abelian integral is in fact an *elliptic integral* that can be treated according to formulae of Byrd and Friedman, [4].

Later on Mihailović, [7–9], taking different approaches, also studied this problem posed by Radzievskij.

In this paper we apply several standard techniques of Classical Analytical and Celestial Mechanics and Astrodynamics to look further into this same problem.

In particular we will resort to some methods of *regularisation and linearisation* of Keplerian systems, namely: *Binet's method* (Boccaletti and Pucacco, [2], Ch. 2, §2.1, pp. 134–135; Goldstein, [5], Ch. 3, §3.5, pp. 85–86), and the *Izsák–Sperling method* (Bond and Allman, [3], Ch. 9, §9.3, pp. 151–154; Izsák [6]; Sperling [11]), by means of which the equations of motion governing perturbed Keplerian systems can be brought into the form of second–order ordinary differential equations governing perturbed harmonic oscillators.

We also obtain the *Planetary Equations in Gaussian Form* (Abad, [1], Ch. 12, §12.3, pp. 195–197) corresponding to this Radzievskij problem, reformulate these equations with the eccentric anomaly of elliptic Keplerian motion as the independent variable, and integrate the resulting differential system over one revolution along the unperturbed orbit, which allows us to identify the secular and periodic terms in the variation of a set of elliptic Keplerian orbital elements over one period of that eccentric anomaly.

Finally, taking advantage of the *Theory of Canonical Transformations* within the framework of Hamiltonian Mechanics (Boccaletti and Pucacco, [2], Ch. 1, Part C, §1.12 – §1.16, pp. 76–106; Goldstein [5], Ch. 9 – 10, pp. 378–498), we will consider the *Hamilton–Jacobi method* (Boccaletti and Pucacco, [2], Ch. 1, Part C, §1.15 – §1.16,

pp. 90–106; Goldstein [5], Ch. 10, pp. 438–498) looking for a complete solution to the Hamilton–Jacobi partial differential equation attached to Radzievskij's problem formulated in the *Hill–Whittaker canonical variables*, also called *polar nodal variables* (Abad, [1], Ch. 9, §9.8, pp. 158–159), proceeding by separation of variables and quadratures (which will also involve elliptic integrals).

2. The Kepler Problem and Its First Integrals

The problem of motion in space of two material particles, with masses m_1 and m_2, under the effect of forces due to their mutual interactions ("internal forces" that satisfy the Law of Action and Reaction, or Third Law of Motion, established by Newton), can be reduced (Boccaletti and Pucacco, [2], Ch. 2, §2.1, pp. 126–128) to *two independent –or decoupled– (sub)problems*: the problem of *motion of their centre of mass* (which will be a uniform rectilinear motion), and the *problem of relative motion* (motion of one of the particles with respect to the other one).

In the special case in which the interactions between those particles are governed by their mutual gravitational attraction according to Newton's Law of Universal Gravitation, we are dealing with the so–called *gravitational two–body problem*, and the corresponding problem of relative motion is known as the *Kepler problem* (Abad, [1], Ch. 7, §7.5–7.6, pp. 115–118; Bond and Allman, [3], Ch. 2, §2.2, pp. 13–15).

The differential equation of motion of the Kepler problem in Newtonian formulation (Abad, [1], Ch. 7, §7.6, Eq (7.23), p. 116–117, Ch. 8, §8.1, pp. 123–124; Boccaletti and Pucacco, [2], Ch. 2, §2.1, Eq (2.16), p. 131; Bond and Allman, [3], Ch. 2, §2.2, Eq (2.8), p. 14) can be formulated as the second–order ordinary differential equation (for the vector unknown function $\mathbf{r}$ of the scalar independent variable t, which represents the physical time)

$$\ddot{\mathbf{r}} \;=\; -\,\frac{\mu}{r^3}\,\mathbf{r} \;=\; -\,\frac{\mu}{r^2}\,\frac{\mathbf{r}}{r} \;=\; -\,\frac{\mu}{r^2}\,\widehat{\mathbf{r}}\,, \quad \text{with} \quad r \,=\, \|\,\mathbf{r}\,\|\,, \quad \widehat{\mathbf{r}} \,=\, \frac{\mathbf{r}}{r}\,, \quad \mu \,=\, \mathcal{G}\,(\,m_1 + m_2\,)\,, \tag{2.1}$$

where at any instant of time $\mathbf{r}$ is the relative position vector of one of the particles with respect to the other one, the scalar r is the (Euclidean) distance separating the particles, $\widehat{\mathbf{r}}$ stands for the (unit) direction vector of their relative position, μ is the Keplerian coupling parameter (or gravitational coupling parameter) of the (two–body) system of particles with masses m_1 and m_2, and $\mathcal{G}$ denotes the Universal Gravitational Constant.

In addition to this we use *dot notation* for derivatives with respect to physical time t.

The differential problem posed by Eq. (2.1) *possesses* the (time–independent) vector first integrals of the orbital angular momentum $\mathbf{G}$ and the Laplace vector $\mathbf{A}$, and the (time–independent) scalar first integral of the Keplerian energy $\mathcal{E}_k$ (Abad, [1], Ch. 8, §8.2, p. 124–126; Boccaletti and Pucacco, [2], Ch. 2, §2.1, p.128–134; Bond and Allman, [3], Ch. 2, §2.4, pp. 19–26, Ch. 8, §8.4, p.126; Goldstein, [5], Ch. 3, §3.2, pp.71–74, §3.9, p.102–105):

$$\mathbf{G}\,(\,t\,,\,\mathbf{r}\,,\,\dot{\mathbf{r}}\,) \;=\; \mathbf{r}\,\times\,\dot{\mathbf{r}}\,, \qquad \mathbf{A}\,(\,t\,,\,\mathbf{r}\,,\,\dot{\mathbf{r}}\,) \;=\; \dot{\mathbf{r}}\,\times\,\mathbf{G} \,-\, \frac{\mu}{r}\,\mathbf{r}\,, \qquad \mathcal{E}_k\,(\,t\,,\,\mathbf{r}\,,\,\dot{\mathbf{r}}\,) \;=\; \frac{1}{2}\,\|\,\dot{\mathbf{r}}\,\|^2 \,-\, \frac{\mu}{r}\,. \tag{2.2}$$

3. Perturbed Kepler Problems and Variation of Keplerian First Integrals under Perturbations

A *perturbed Kepler problem* is a problem of Keplerian motion *slightly distorted* by the occurrence of some other effects or phenomena due to the presence of other bodies or other forces acting on the original Keplerian system.

The differential equation of perturbed Keplerian motion, when additional *perturbing forces* $\mathbf{P} \,=\, \mathbf{P}\,(\,t\,,\,\mathbf{r}\,,\,\dot{\mathbf{r}}\,)$ are considered, can be written in the Newtonian formulation as (Abad, [1], Ch. 12, §12.1, pp.191–192; Bond and Allman, [3], Ch. 8, §8.1, p. 117)

$$\ddot{\mathbf{r}} \;=\; \mathbf{F}_{total}\,(\,t\,,\,\mathbf{r}\,,\,\dot{\mathbf{r}}\,) \;=\; \mathbf{F}_{Kepler}\,(\,-\,,\,\mathbf{r}\,,\,-\,) \,+\, \mathbf{P}\,(\,t\,,\,\mathbf{r}\,,\,\dot{\mathbf{r}}\,) \;=\; -\,\frac{\mu}{r^3}\,\mathbf{r} \,+\, \mathbf{P}\,(\,t\,,\,\mathbf{r}\,,\,\dot{\mathbf{r}}\,)\,. \tag{3.1}$$

Time variations of the above Keplerian constants of motion (2.2) under the effect of perturbing forces $\mathbf{P}$ are described by the equations (Bond and Allman, [3], Ch. 8, §8.6, Formulae (8.36–8.37), (8.41), (8.54), pp. 127–134),

$$\frac{d\,\mathbf{G}}{d\,t} \;=\; \mathbf{r}\,\times\,\mathbf{P}\,, \qquad \frac{d\,\mathcal{E}_k}{d\,t} \;=\; \mathbf{P}\cdot\mathbf{r}\,, \qquad \frac{d\,\mathbf{A}}{d\,t} \;=\; 2\,(\,\mathbf{P}\cdot\dot{\mathbf{r}}\,)\,\mathbf{r} \,-\, (\,\mathbf{r}\cdot\dot{\mathbf{r}}\,)\,\mathbf{P} \,-\, (\,\mathbf{P}\cdot\mathbf{r}\,)\,\dot{\mathbf{r}}\,. \tag{3.2}$$

4. The Perturbed Keplerian System Considered by Radzievskij. Force Model and First Integrals

After some simplifications Radzievskij ([10], p. 1309) arrives at a perturbed Keplerian system whose *force model* can be written in the form

$$\mathbf{F}_{total}\,(\,t\,,\,\mathbf{r}\,,\,\dot{\mathbf{r}}\,) = \mathbf{F}_{total}\,(\,\mathbf{r}\,) = \mathbf{F}_{Kepler}\,(\,\mathbf{r}\,) + \mathbf{P}\,(\,\mathbf{r}\,) = -\,\frac{\mu}{r^3}\,\mathbf{r} \,-\, k\,\mathbf{r} = -\left(\frac{\mu}{r^2} \,+\, k\,r\right)\frac{\mathbf{r}}{r} \,=\, f\,(r)\,\widehat{\mathbf{r}}\,, \tag{4.1}$$

with $k \,=\, (4/3)\,\pi\,\mathcal{G}\,\delta$, where δ is the (constant) *density* of the homogeneous cloud. Consequently, $0 \,<\, k \,<<\, 1$, and so k can be taken as a *small perturbation parameter*. And in this way the differential equation of motion for the perturbed Kepler problem studied by Radzievskij ([10], Eqs. (2), p. 1309) can be written as

$$\ddot{\mathbf{r}} \;=\; f\,(r)\,\widehat{\mathbf{r}} \;=\; -\left(\frac{\mu}{r^2} \,+\, k\,r\right)\frac{\mathbf{r}}{r} \;=\; -\,\frac{\mu}{r^3}\,\mathbf{r} \,-\, k\,\mathbf{r}\,. \tag{4.2}$$

In view of the form, sign and functional dependence of function $f(r) < 0$, the total force acting in this problem is an *attractive, conservative central force*, which entails that Eq. (4.2) admits the first integrals of the *angular momentum* vector $\mathbf{G}$ and the *total mechanical energy* $\mathcal{E}$ of the system (see Formulae (4.3) below). Notice also that the perturbing force in Eqs. (4.1)–(4.2), $\mathbf{P} = \mathbf{P}(t, \mathbf{r}, \dot{\mathbf{r}}) = \mathbf{P}(-, \mathbf{r}, -) = -k\,\mathbf{r}$, can be derived from the *scalar perturbing potential* $V_{Perturb.}(r) = (1/2)\,k\,r^2$, while the total force derives from the *total scalar potencial* $V_{total}(r) = V_{Kepler}(r) + V_{Perturb.}(r) = -(\mu/r) + (1/2)\,k\,r^2$. The existence of this scalar potential allows us to specify an analytical expression for the total mechanical energy $\mathcal{E}$ of the system in Eq. (4.3).

As a conclusion, from the said constants of motion in this Radzievskij two–body problem we have ([10], Eqs. (3) and (4), p. 1310)

$$\mathbf{G}(t, \mathbf{r}, \dot{\mathbf{r}}) = \mathbf{r} \times \dot{\mathbf{r}}\ ; \quad \mathcal{E}(t, \mathbf{r}, \dot{\mathbf{r}}) = \frac{1}{2}\,||\,\dot{\mathbf{r}}\,||^2 - \frac{\mu}{r} + \frac{1}{2}\,k\,r^2 \implies v^2 = ||\,\dot{\mathbf{r}}\,||^2 = \frac{2\mu}{r} - k\,r^2 + 2\mathcal{E}\ . \tag{4.3}$$

But the Laplace vector of the Kepler problem, $\mathbf{A}$, is not a constant of motion for this system, since by (3.2)

$$\frac{d\,\mathbf{A}}{d\,t} = (-k\,\mathbf{r}\cdot\dot{\mathbf{r}})\,\mathbf{r} + \left(k\,r^2\right)\dot{\mathbf{r}} = (-k\,r\dot{r})\,\mathbf{r} + \left(k\,r^2\right)\dot{\mathbf{r}} \neq \mathbf{0}\ .$$

5. The Radzievskij First–Integral Approach to the Solution. Orbit Equation in Finite Terms (Inverted Form)

After Introducing plane polar coordinates (r, φ) in the orbital plane, from the above first integrals (4.3) one has

$$\mathbf{G} = \mathbf{r} \times \dot{\mathbf{r}} \implies \|\,\mathbf{G}\,\| = G = r^2\dot{\varphi} = \text{const.} \implies \dot{\varphi} = \frac{d\,\varphi}{d\,t} = \frac{G}{r^2}\ , \tag{5.1}$$

$$v^2 = \left(\frac{d\,\mathbf{r}}{d\,t}\right)^2 = \dot{r}^2 + r^2\,\dot{\varphi}^2 = \underline{\left(\frac{d\,r}{d\,\varphi}\right)^2 \frac{G^2}{r^4} + \frac{G^2}{r^2} = \frac{2\mu - k\,r^3 + 2\mathcal{E}r}{r}}\ . \tag{5.2}$$

Note that the last expression in (5.2), read as a function of r, possesses *(at least) one real root* r_0. From the underlined part of Formula (5.2) Radzievskij obtained the differential equation of the orbit as a first–order ordinary differential equation for the unknown function r of the independent variable φ; by separation of variables, and considering some initial conditions of the form $t = t_0$, $r(t_0) = r_0$, $\varphi(t_0) = \varphi_0$, he arrived at an equation of the *orbit in finite terms (in the inverted form)* $\varphi = \varphi(r; r_0, \varphi_0)$, instead of a parametric representation $r = r(\varphi; r_0, \varphi_0)$ in plane polar coordinates with the polar angle as the parameter ([10], Eq. (5), p. 1310)

$$\varphi = \varphi(r; r_0, \varphi_0) = \varphi_0 + G\int_{r_0}^{r} \frac{(1/r)\,dr}{\sqrt{2\mu r - k\,r^4 + 2\mathcal{E}r^2 - G^2}}\ . \tag{5.3}$$

At this stage of his developments he stated that the integral in this expression is an *Abelian* (or *hyperelliptic*) *integral*. However it may be treated as an elliptic integral (Byrd and Friedman, [4]).

6. Some Analytical Treatments of Mihailović

In several of his papers, [7–9], Mihailović investigated the problem of motion (4.2) posed by Radzievskij.

In [7] he essentially performed a *vector–element treatment* of this perturbed Keplerian system, in terms of the Milanković elements $\mathbf{G}$, $\mathbf{A}$, t_p, where t_p is the instant of a pericentre passage; the perturbation equations are presented in the forms proposed by Milanković and Popović, and a vector element introduced by Bilimović is used.

Another vector–element treatment, *à la Milanković–Popović–Bilimović*, in terms of the time derivative of the Laplace vector of Keplerian motion $\mathbf{A}$ (see Formulae (2.2) above), is proposed in [8].

In [9] he reformulated the differential equation (4.2) in the form of the equation of motion of a 3–dimensional *nonlinearly–forced harmonic oscillator*, with the *forcing term* equal to the force giving rise to the Kepler problem, that is, $\ddot{\mathbf{r}} + k\,\mathbf{r} = -\left(\mu/r^3\right)\mathbf{r}$. The solution of this problem is formally approached according to the *method of variation of the (vector) integration constants*, $\mathbf{c}_1$ and $\mathbf{c}_2$, involved in the general solution to the second–order homogeneous linear differential equation with constant coefficients corresponding to the unperturbed harmonic oscillator, namely $\mathbf{r}(t; \mathbf{c}_1, \mathbf{c}_2) = \mathbf{c}_1 \cos\left(\sqrt{k}\,t\right) + \mathbf{c}_2 \sin\left(\sqrt{k}\,t\right)$. A bound for the time changes of the said quantities is also given: $|\,\dot{\mathbf{c}}_j\,| \leq \left(1/\sqrt{k}\right)\left(\mu/r^2\right)$.

7. Orbital Elements of an Elliptic Keplerian Motion and Planetary Equations for their Time Changes

Keplerian orbital elements are certain constants or parameters that unambiguously characterize a conic–section solution to a Kepler problem. At any instant of time they give account of the geometric shape, size, and position

of the orbit in space, but also of the position of the moving particle along the Keplerian conic–section at issue (Abad, [1], Ch. 9, §9.1 – §9.2, pp.141–145; Boccaletti and Pucacco, [2], Ch. 2, §2.5, p. 156–157; Bond and Allman, [3], Ch. 4, §4.6, p. 49–55; Goldstein, [5], Ch.10 , §10.7, p. 478–479).

Consider an *elliptic Keplerian orbit* characterized by the following *orbital elements*:

- a = semi–major axis of the ellipse,
- e = its numerical eccentricity (and the related notation $\eta = \sqrt{1 - e^2}$),
- I = inclination of the orbital plane (with respect to the $x_1 x_2$ fundamental plane of the coordinate system),
- Ω = argument of longitude of the ascending node,
- ω = argument of the periastron (or argument of the pericentre),
- ℓ = mean anomaly,

and the *additional concepts and notations*,

- f = Keplerian true anomaly,
- E = eccentric anomaly of elliptic Keplerian motion.

Orbital elements can undergo *variations under perturbations*. To study these variations, sets of differential equations describing such changes in orbital elements due to the presence of some perturbing force are established.

- If the perturbing forces admit a scalar (perturbing) potential $V_{Perturb.}$, the time variations of orbital elements can be expressed in terms of partial derivatives of the perturbing potential with respect to the orbital elements, by means of the so–called *planetary equations in the Lagrange form* (Abad, [1], Ch. 12, §12.2, pp. 192–195).

- If the perturbing forces cannot be derived from a scalar (perturbing) potential, the time variations of orbital elements can be developed in terms of the *radial, transversal, and normal components of the perturbing force*, which gives rise to the *planetary equations in the Gauss form* (Abad, [1], Ch. 12, §12.3, pp. 195–197).

In what follows we will take this last approach. To this end, let the perturbing force $\mathbf{P} = \mathbf{P}(t, \mathbf{r}, \dot{\mathbf{r}})$ be decomposed into its components with respect to the *orbital reference frame*, or *Gaussian frame*, $\{\mathbf{u}, \mathbf{v}, \mathbf{n}\}$ (Abad, [1], Ch. 6, §6.4, pp. 101–103, and Ch. 9, §9.4, §§9.4.5, pp. 149–150), $\mathbf{P} = (P_u, P_v, P_n)$. Obviously, if $\mathbf{P}$ is a central force then its transversal and normal components are zero, and $\mathbf{P} = (P_u, 0, 0)$. The planetary equations in Gaussian form for the preceding orbital elements are (Abad, [1], §12.3, Formulae (12.21), p. 196)

$$\frac{da}{dt} = \frac{2e\sin f}{n\eta} P_u + \frac{2a\eta}{nr} P_v , \qquad \frac{de}{dt} = \frac{\eta \sin f}{an} P_u + \left(\frac{\eta^3}{enr} - \frac{\eta r}{a^2 en} \right) P_v ,$$

$$\frac{dI}{dt} = \frac{r\cos(\omega + f)}{a^2 n\eta} P_n , \qquad \frac{d\Omega}{dt} = \frac{r\sin(\omega + f)}{a^2 n\eta \sin I} P_n ,$$

$$\frac{d\omega}{dt} = \frac{\eta\cos f}{aen} P_u + \frac{r(2 + e\cos f)\sin f}{a^2 en\eta} P_v + \frac{r\sin(\omega + f)\cos I}{a^2 n\eta \sin I} P_n ,$$

$$\frac{d\ell}{dt} = n + \left(\frac{-2r}{a^2 n} + \frac{\eta^2 \cos f}{aen} \right) P_u - \frac{r(2 + e\cos f)\sin f}{a^2 en} P_v .$$

The right–hand sides of these equations become simpler in the case of central forces, for which $P_v = 0$ and $P_n = 0$. In particular, for the case of our Radzievskij problem (4.2) the preceding planetary equations reduce to

$$\frac{da}{dt} = -\frac{2ek}{n\eta} r\sin f , \quad \frac{de}{dt} = -\frac{\eta k}{an} r\sin f ,$$

$$\frac{dI}{dt} = 0 , \qquad \frac{d\Omega}{dt} = 0 ,$$

$$\frac{d\omega}{dt} = -\frac{\eta k}{aen} r\cos f , \quad \frac{d\ell}{dt} = n + \frac{2k}{a^2 n} r^2 - \frac{\eta^2 k}{aen} r\cos f .$$

In order to obtain these planetary equations in Gaussian form *with respect to the eccentric anomaly* of elliptic Keplerian motion, E , as the (new) independent variable, by virtue of the chain rule we have that for any orbital element σ (Abad, [1], Ch. 8, §8.5, §§8.5.3, pp. 134–136; Boccaletti and Pucacco, [2], Ch. 2, §2.4, pp. 150–151)

$$\frac{d\sigma}{dt} = \frac{d\sigma}{dE}\frac{dE}{dt} = \frac{na}{r(E)}\frac{d\sigma}{dE} \implies \frac{d\sigma}{dE} = \frac{r(E)}{na}\frac{d\sigma}{dt} , \quad \text{where} \quad r(E) = a(1 - e\cos E) . \tag{7.1}$$

We are also interested in having the right–hand sides of the preceding equations available in terms of E ; for this reason we express the circular functions of the true anomaly f in terms of circular functions of the eccentric

anomaly E, with the semi–minor axis $b = a\sqrt{1 - e^2} = a\eta$, namely (Bond and Allman [3], Ch. 4, §4.3, Formulae (4.10) and (4.14), pp. 44–45)

$$\cos f = \frac{\cos E - e}{1 - e\cos E} = \frac{a\,(\cos E - e)}{r\,(E)}\,, \qquad \sin f = \frac{\sqrt{1 - e^2}\,\sin E}{1 - e\cos E} = \frac{b\,\sin E}{r\,(E)}\,,$$

and then the right–hand sides of the above equations are rewritten as explicit functions of E in the form

$$\begin{aligned}
&\frac{da}{dt} = -\frac{2ekb}{n\eta}\sin E\,, \qquad \frac{de}{dt} = -\frac{\eta k b}{an}\sin E\,, \qquad \frac{dI}{dt} = 0\,,\\
&\frac{d\Omega}{dt} = 0\,, \qquad \frac{d\omega}{dt} = -\frac{\eta k}{en}\,(\cos E - e)\,,\\
&\frac{d\ell}{dt} = n + \frac{2k}{a^2 n}\,r^2 - \frac{\eta^2 k}{en}\,(\cos E - e)\,.
\end{aligned}$$

To describe the variations of the orbital elements with respect to the eccentric anomaly E we reformulate these Gauss planetary equations with the eccentric anomaly as the independent variable, according to Formulae (7.1), and obtain the system of first–order ordinary differential equations

$$\begin{aligned}
&\frac{da}{dE} = -\frac{2eak}{n^2}\,(1 - e\cos E)\sin E\,, \qquad \frac{de}{dE} = -\frac{\eta^2 k}{n^2}\,(1 - e\cos E)\sin E\,,\\
&\frac{dI}{dE} = 0\,, \qquad \frac{d\Omega}{dE} = 0\,,\\
&\frac{d\omega}{dE} = -\frac{\eta k}{en^2}\,(\cos E - e)(1 - e\cos E)\,,\\
&\frac{d\ell}{dE} = (1 - e\cos E) + \frac{2k}{n^2}\,(1 - e\cos E)^3\\
&\qquad\qquad - \frac{\eta^2 k}{en^2}\,(\cos E - e)(1 - e\cos E)\,.
\end{aligned}$$

Integrating these equations with respect to the eccentric anomaly over one revolution along the unperturbed Keplerian ellipse (which entails that a, e, η, and n are taken as constant quantities on the right–hand sides) leads to

$$\begin{aligned}
&\Delta a = \frac{2eak}{n^2}\cos E - \frac{e^2 ak}{2n^2}\cos 2E\,, \qquad \Delta e = \frac{\eta^2 k}{n^2}\cos E - \frac{\eta^2 ek}{4n^2}\cos 2E\,,\\
&\Delta I = 0\,, \qquad \Delta\Omega = 0\,,\\
&\Delta\omega = \frac{3\eta k}{2n^2}\,E - \frac{\eta k\left(1 + e^2\right)}{en^2}\sin E + \frac{\eta k}{4n^2}\sin 2E\,,\\
&\Delta\ell = \left[1 + \frac{3\eta^2 k}{2n^2} + \frac{2k}{n^2}\left(1 + \frac{3}{2}\,e^2\right)\right]E - \left[e + \frac{\eta^2 k}{en^2}\left(1 + e^2\right) + \frac{6k}{n^2}\left(e + \frac{1}{4}\,e^3\right)\right]\sin E\\
&\qquad + \left[\frac{\eta^2 k}{4n^2} + \frac{3ke^2}{2n^2}\right]\sin 2E - \frac{ke^3}{6n^2}\sin 3E\,.
\end{aligned}$$

These expressions give account of the *secular* and *periodic variations* of those orbital elements over one period of the eccentric anomaly. Note that *the elements I and Ω remain constant* (as expected, given that their changes are governed by the normal component P_n of the perturbing force, and for central forces that component is zero).

8. Binet's Regularisation Method. Application to Our Radzievskij Problem

Under the effect of a conservative central force, with a force law given by a scalar function $f(r)$ and scalar potential $V(r)$, the equation of motion in the radial direction in polar coordinates (r, φ) taken within the orbital plane reduces to (Goldstein, [5], Ch. 3, §3.2, Eq (3–12), p. 74)

$$\ddot{r} - \frac{G^2}{r^3} = f(r)\,, \qquad f(r) = -\frac{dV(r)}{dr}\,.$$

This second–order ordinary differential equation, after the *changes of the dependent and independent variables* (Boccaletti and Pucacco, [2], Ch. 2, §2.1, pp. 134–135; Goldstein, [5], Ch. 3, §3.5, pp. 85–86) given by

$$r \longrightarrow u: \; r = \frac{1}{u} \; , \qquad t \longrightarrow \varphi: \; dt = \frac{r^2}{G} \, d\varphi \; , \tag{8.1}$$

with *prime notation* for derivatives with respect to the polar angle φ, takes on the form (*Binet's Equation*),

$$u'' + u = - \frac{1}{G^2} \left\{ \frac{1}{u^2} \, f\left(\frac{1}{u} \right) \right\} , \tag{8.2}$$

which admits the *(energy–like) first integral*

$$(u')^2 + u^2 = - \frac{2}{G^2} \, V\left(\frac{1}{u} \right) + \widetilde{\mathcal{E}} \; , \quad \text{with} \quad \widetilde{\mathcal{E}} = \text{const.}$$

By solving for u', this first integral provides us with a first–order ordinary differential equation of the orbit.

In particular, Binet's Method applied to our Radzievskij problem leads to the second–order ordinary differential equation, for the unknown function u of the independent variable φ,

$$u'' + u = \frac{\mu}{G^2} + \frac{k}{G^2 u^3} \; , \quad \text{with the first integral} \quad (u')^2 + u^2 = \frac{2\mu}{G^2} \, u - \frac{k}{G^2 u^2} + \widetilde{\mathcal{E}} \; .$$

From this first integral, solving for u', proceeding by separation of variables, and choosing some initial conditions at an instant $t = t_0$, $r(t_0) = r_0$, $u(t_0) = u_0$, $\varphi(t_0) = \varphi_0$, we obtain the following *equation of the orbit in finite terms* (in inverted form) in terms of an *elliptic integral*,

$$\varphi = \varphi(u; u_0, \varphi_0) = \varphi_0 + G \int_{u_0}^{u} \frac{u \, du}{\sqrt{2\mu u^3 - k + G^2 \widetilde{\mathcal{E}} u^2 - G^2 u^4}} \; .$$

9. Izsák–Sperling Regularisation Technique

A *perturbed Keplerian system* (3.1), under the *differential transformation* of the independent variable

$$t \longrightarrow s: \quad dt = r \, ds \qquad \text{(Sundman transformation)} \; ,$$

and after *embedding* the Keplerian first integrals of the Keplerian energy $\mathcal{E}_k$ and the Laplace vector $\mathbf{A}$, becomes

$$\mathbf{r}'' + (-2\mathcal{E}_k) \, \mathbf{r} = - \mathbf{A} + r^2 \, \mathbf{P}(t, \mathbf{r}, \mathbf{r}') \; , \tag{9.1}$$

with *prime notation* for derivatives with respect to the new independent variable s (Bond and Allman, [3], Ch. 9, §9.3, pp. 151–154), from which the following *scalar differential equation for the distance* r can be derived,

$$r'' + (-2\mathcal{E}_k) \, r = \mu + [\mathbf{r} \cdot \mathbf{P}(t, \mathbf{r}, \mathbf{r}')] \, r \; . \tag{9.2}$$

Taking these steps, the *transformed equations* (9.1) and (9.2) corresponding to our Radzievskij problem (4.2) are

$$\mathbf{r}'' + (-2\mathcal{E}_k) \, \mathbf{r} = - \mathbf{A} - \left(k r^2 \right) \mathbf{r} \; , \qquad r'' + (-2\mathcal{E}_k) \, r = \mu - k r^3 \; .$$

This last scalar equation for the distamce r admits the *(energy–like) first integral*

$$\frac{1}{2} \, \| \mathbf{r}' \|^2 - \mathcal{E}_k \, r^2 = - \mathbf{A} \cdot \mathbf{r} - \frac{1}{2} \, k \, r^4 + \mathcal{E}^* \; , \quad \text{with} \quad \mathcal{E}^* = \text{const.} \; ,$$

from which (after considering some initial conditions at an instant $t = t_0$, $s(t_0) = s_0$, $r(t_0) = r_0$) we can obtain the solution (in inverted form), involving an elliptic integral,

$$s = s(r; r_0, s_0) = s_0 + \int_{r_0}^{r} \frac{dr}{\sqrt{2\mathcal{E}^* - 3G^2 + 2\mu r + 2\mathcal{E}_k r^2 - k r^4}} \; .$$

10. Hill–Whittaker Polar Nodal Canonical Variables. A Hamiltonian Formulation of the Radzieskij System

Instead of using the set of polar spherical variables we will resort to the canonical set of the *Hill–Whittaker* variables, also known as *polar nodal variables* (Abad, [1] , Ch. 9, §9.8, pp. 158–159). Our notations for these variables will be $(r, \theta, \nu; p_r, p_\theta, p_\nu)$, with the following meaning,

- $r = \| \mathbf{r} \| \geq 0$ denotes the Euclidean norm of the position vector $\mathbf{r}$, and represents the Euclidean *distance from the origin to the moving particle*;

• $\theta = \omega + f$ is the *argument of latitude of the moving mass* (angular distance measured from the ascending node to the moving point), and is defined modulo 2π;

• $\nu = \Omega$ represents the *argument of longitude of the asceding node*;

as for the canonically conjugate momenta corresponding to these coordinates,

• $p_r = \dot{r}$ is the *radial component of the velocity* of the moving mass;

• $p_\theta = G = \| \mathbf{G} \|$ designates the *norm of the orbital angular momentum* of the particle, and

• $p_\nu = G \cos I$ denotes the *polar (or vertical) component of the angular momentum vector*.

Remember the notations previously used for some Keplerian orbital elements and parameters: ω = argument of the pericentre, f = true anomaly, Ω = longitude of the ascending node, I = inclination of the orbital plane.

The *Hamiltonian function of a perturbed Kepler problem in polar nodal variables* will be a function

$$\begin{aligned} \mathcal{H} &\equiv \mathcal{H}(t, r, \theta, \nu, p_r, p_\theta, p_\nu) = \mathcal{H}_0(-, r, -, -, p_r, p_\theta, -) + V_P \\ &= \left\{ \frac{1}{2} \left[p_r^2 + \frac{p_\theta^2}{r^2} \right] - \frac{\mu}{r} \right\} + V_P(t, r, \theta, \nu, p_r, p_\theta, p_\nu) \,, \end{aligned} \tag{10.1}$$

where $\mathcal{H}_0$ is the Hamiltonian of a pure Keplerian system, and V_P represents the *potential of the perturbing forces*.

In particular, for our *Radzievskij problem* (4.2) *in polar nodal variables*, with $V_P = V_P(r) = (1/2)\, k r^2$,

$$\mathcal{H} \equiv \mathcal{H}(-, r, -, -, p_r, p_\theta, -) = \frac{1}{2} \left[p_r^2 + \frac{p_\theta^2}{r^2} \right] - \frac{\mu}{r} + \frac{1}{2} k r^2 \,(= \text{energy of the system} \equiv \underline{\underline{p_L}}) \,. \tag{10.2}$$

Since the canonical variables θ, ν and p_ν are *ignorable (cyclic)*, their canonically conjugate variables are constants of the motion: $p_\theta = \text{const.}$, $p_\nu = \text{const.}$, $\nu = \text{const.}$

With the aim of taking advantage of the conservation of these quantities along the motion, we will look for changes to new canonical variables, starting from the polar nodal set, that *preserve the couple of canonically conjugate variables ν and p_ν*, i. e., transformations that, on these variables, act as the *identity transformation*. More specifically, we look for canonical transformations $(r, \theta, \nu; p_r, p_\theta, p_\nu) \longrightarrow (q_L, q_G, q_H; p_L, p_G, p_H)$, introducing *new canonical variables* $(q_L, q_G, q_H; p_L, p_G, p_H)$, in such a way that the pair of canonically conjugate variables $(\nu; p_\nu)$ should remain unchanged, while the transformation actively operates on the variables $(r, \theta; p_r, p_\theta)$ and $(q_L, q_G; p_L, p_G)$.

For instance, such transformations can be defined by means of *generating functions $S \equiv S(r, \theta, \nu; p_L, p_G, p_H)$ of the second kind*, that is, depending on the *old coordinates* and the *new momenta* (Boccaletti and Pucacco, [2], Ch. 1, Part C, §1.13, pp. 82–84; Goldstein, [5], Ch. 9, §9.1, pp. 383–384, Ch. 10, §10.1, pp. 438–442, §10.3, pp. 445–449)

11. A Canonical Transformation Yielding Some Canonical Constants

We perform a completely canonical transformation $(r, \theta, \nu; p_r, p_\theta, p_\nu) \longrightarrow (q_L, q_G, q_H; p_L, p_G, p_H)$, mixing *old coordinates* and *new momenta*, derived from a generating function of the second kind

$$S \equiv S(r, \theta, \nu; p_L, p_G, p_H) = \theta\, p_G + \nu\, p_H + \int_{r_0}^{r} \sqrt{Q}\; dr \,, \tag{11.1}$$

$$Q \equiv Q(r, -, -; p_L, p_G, -) = 2 p_L + \frac{2\mu}{r} - k r^2 - \frac{p_G^2}{r^2} = \frac{2 p_L r^2 + 2\mu r - k r^4 - p_G^2}{r^2} \tag{11.2}$$

where r_0 is a zero of the equation (in r) $Q(r, -, -; p_L, p_G, -) = 0$. In practice, it is usually taken as *the least positive root of that equation*.

The implicit *transformation equations derived from S* are (Boccaletti and Pucacco, [2], Ch. 1, Part C, §1.13, Formulae (1.C.70), p. 83; Goldstein, [5], Ch. 9, §9.1, Formulae (9.17ab), p. 383)

$$\begin{aligned} p_r\ (= \dot{r}) &= \frac{\partial S}{\partial r} = \sqrt{Q} \,, \qquad p_\theta\ (= G) = \frac{\partial S}{\partial \theta} = p_G \,, \qquad p_\nu = \frac{\partial S}{\partial \nu} = p_H \,, \\ q_L &= \frac{\partial S}{\partial p_L} = \int_{r_0}^{r} \frac{dr}{\sqrt{Q}} \quad \longrightarrow \underline{\text{generalized Kepler equation}} \,, \\ q_G &= \frac{\partial S}{\partial p_G} = \theta - p_G \int_{r_0}^{r} \frac{dr}{r^2 \sqrt{Q}} \,, \qquad q_H = \frac{\partial S}{\partial p_H} = \nu \,. \end{aligned}$$

Note that here p_L is the total energy of the problem, while p_G is the magnitude of the orbital angular momentum.

Alternatively, the coordinates q_L and q_G of the new canonical set are given by the relations

$$q_L = \int_{r_0}^{r} \frac{r\,dr}{\sqrt{2p_L r^2 + 2\mu r - kr^4 - p_G^2}} ,$$
$$q_G = \theta - p_G \int_{r_0}^{r} \frac{dr}{r\sqrt{2p_L r^2 + 2\mu r - kr^4 - p_G^2}} .$$

Introducing the following notations and abbreviations,

$$\mathcal{R}(r) = 2p_L r^2 + 2\mu r - kr^4 - p_G^2 , \quad \mathcal{I}_1 = \int_{r_0}^{r} \frac{r\,dr}{\sqrt{\mathcal{R}(r)}} , \quad \mathcal{I}_2 = \int_{r_0}^{r} \frac{(1/r)\,dr}{\sqrt{\mathcal{R}(r)}} , \tag{11.3}$$

one has that $\underline{\underline{q_L = \mathcal{I}_1}}$, and $\underline{\underline{q_G = \theta - p_G\,\mathcal{I}_2}}$.

Note that the above integrals $\mathcal{I}_1$ and $\mathcal{I}_2$ *share the form*

$$\mathcal{I} = \int_{r_0}^{r} \frac{(\text{rational function of } r)\,dr}{\sqrt{\mathcal{R}(r)}} , \tag{11.4}$$

and *can be reduced to elliptic integrals*, [4].

12. The Radzievskij Problem in the New Canonical Variables

The completely canonical transformation derived from the generating function S given in Formulae (11.1)–(11.2) converts the Hamitonian (10.2) of the Radzievskij problem into a function $\mathcal{K}(-, q_L, q_G, q_H, p_L, p_G, p_H)$, of the new variables, that in this case reads

$$\mathcal{K} \equiv \mathcal{K}(-,-,-,-,p_L,-,-) = p_L . \tag{12.1}$$

From the canonical equations of motion derived from this new Hamiltonian $\mathcal{K}$ we obtain a *canonical solution* to this Radzievskij problem in the new canonical variables,

$$\frac{dq_L}{dt} = \frac{\partial \mathcal{K}(p_L)}{\partial p_L} = 1 \implies q_L(t) = t + \text{const.} , \tag{12.2}$$

while *the remaining new variables are constant*.

Note also that we have already established that $q_L = \mathcal{I}_1$, and so $t + \text{const.} = \mathcal{I}_1$, which can be interpreted as a *generalized Kepler equation*.

Acknowledgements

The author has been partially supported by Grant E24–20R (Government of Aragón, European Social Fund).

References

[1] Alberto Abad. *Astrodinámica*. Bubok Publishing, S.L., 2012. http://www.bubok.es//libro/detalles/219952/Astrodinamica

[2] Dino Boccaletti, Giuseppe Pucacco. *Theory of Orbits* (Vol. 1: *Integrable Systems and Non–perturbative Methods*). Astronomy and Astrophysics Library. Springer, 1996.

[3] Victor R. Bond, Mark C. Allman. *Modern Astrodynamics. Fundamentals and Perturbation Methods*. Princeton University Press, 1996.

[4] Paul F. Byrd, Morris D. Friedman. *Handbook of Elliptic Integrals for Engineers and Scientists* (Second Edition, Revised). Series: Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen, Vol. 67. Springer, 1971.

[5] Herbert Goldstein. *Classical Mechanics* (Second Edition). Addison–Wesley Series in Physics. World Student Series. Addison–Wesley, 1980.

[6] Imre Izsák. Zur Regularisierung des Einzentrumproblems. *Communications of the Konkoly Observatory (Mitteilungen der Sternwarte)* **39** (Vol. III, No. 18): 3–5, 1955.

[7] Dobrivoje Mihailović. An Application of the Perturbation Equations to Radzievsky's Three-Body Problem. *Bulletin de la Société des mathématiciens et physiciens de la R. P. de Serbie*, VII(3–4): 223–228, 1955. (In Serbo–Croatian, with a summary in German).

[8] Dobrivoje Mihailović. Characteristics of the Time Variations of the Perihelion Vector in a Special Problem of Celestial Mechanics. *Publications de la Faculté d'Électrotechnique de l'Université à Belgrade* (Série: Mathématiques et Physique, Paper N⁰ **200,**): 1–6, 1967a. (In Russian).

[9] Dobrivoje Mihailović. Eine Methode für Lösung des Radziewsky's Dreikörperproblems. *Publications de la Faculté d'Électrotechnique de l'Université à Belgrade* (Série: Mathématiques et Physique, Paper N^{o} **201,**): 7–9, 1967b. (In German).

[10] V. V. Radzievskij. General Solution of a Case of the Problem of Three Bodies. *Transactions (Doklady) of the USSR Academy of Sciences* (New Series), 91(6): 1309–1311, 1953 (in Russian).

[11] Hans Sperling. Computation of Keplerian Conic Sections. *American Rocket Society (ARS) Journal*, 31(5): 660–661, 1961.

Comparison of the topological derivative behavior on different scenarios: the time-harmonic heat equation and Maxwell electric field equations

Ana Carpio[1], **Manuel Pena**[2], **María Luisa Rapún**[2]
1. *Universidad Complutense de Madrid, Spain*
2. *Universidad Politécnica de Madrid, Spain*

Abstract

The topological derivative is a very useful tool for the solution of inverse problems. In this paper we compare its behaviour on a couple of problems viewed as particular cases of an abstract one. The abstract formulation is useful to derive a general formula valid for many different scenarios. However, depending on the physical phenomenon under consideration, we can expect qualitatively different results.

1. Introduction

The solution of inverse problems is a very active research field. It has many applications as structural health monitoring, non-destructive testing, tumor shape detection or oil and gas prospecting to name a few of them.

All these problems follow a common pattern. There is some object (a cavity, an oil deposit, a tumor, ...) whose shape we are looking for, which is made of a different material than the background medium. Based on the knowledge of the properties of these two materials, we choose a phenomenon, like acoustic wave propagation, heat transfer, electromagnetic scattering, etc where these two materials behave differently. With this in mind we design a measuring experiment in which this physical magnitude is measured. For example, in this paper we consider a problem where dielectric objects with unknown shape are irradiated by electromagnetic waves and another one where a metallic plate with possible cavities is heated by an infrared lamp. In the first one, the electric field is measured at a series of antennas whereas on the second one the temperature on one of the sides of the plate is inferred from an infrared thermogram.

The direct problem is considered to be the one consisting of computing these physical quantities (electric field/temperature) at any point in the domain when the shape and properties of the object and the kind of excitation (incident electromagnetic wave, thermal radiation, etc) are known. Conversely, the problem of deducing the shape and possibly the properties of the object given some amount of measurements of the electric field/temperature, is the inverse problem.

While it is usually the case that the direct problem is well posed in the Haddamard sense, i.e. it has a solution which is unique and depends continuously on the input data, most inverse problems are severely ill-posed. On the one hand, if too few measurements are performed there are infinite solutions and on the other hand, if there are enough measurements it is often the case that measuring errors prevent the problem from having a solution at all. Even if the problem is formulated in a minimization fashion, the solution will not depend continuously on the measured data, that is, very small amounts of experimental error will lead to completely different approximations of the scatterer/cavity, rendering the solution useless in practical terms.

For this reason, special techniques and algorithms where the inverse problem is regularized are needed. One of these methods is based in the computation of the topological derivative [6] of a shape functional measuring how 'far' is the measured data to the one simulated for a generic known shape.

The formula for the topological derivative for the two-dimensional electromagnetic scattering problem was obtained in [3] (see [7] for the formula of the full three-dimensional case) and the first studies of the topological derivative applied to thermal waves were performed in [4]. In this paper we compare the performance of the topological derivative in two specific applications related with processing experimental two-dimensional electromagnetic data (previously studied in [2]) and synthetic thermograms of thin plates (studied in [8, 9]).

The paper is organized as follows: in section 2 the general physical model for both problems is presented. Section 3 is devoted to describing the concept of the topological derivative, why it is useful for solving this kind of problems and how it can be computed in a fast way. After that, in sections 4 and 5 the results for the electromagnetic scattering and the thermographic inspection problems are shown. In this section the abstract model as well as the aforementioned formula for the topological derivative are particularized for each problem and the results are analyzed. Finally, in section 6 some conclusions and future work are presented.

2. Abstract problem

Both inverse problems we are going to study respond to the same abstract formulation. Let $\mathcal{R} \subset \mathbb{R}^2$ be a connected domain and $\mathcal{D} \subset \mathcal{R}$ a bounded subset of it. The domain $\mathcal{R}$ will represent the background medium while the domain

$\mathcal{D}$ will be the defect, cavity or object whose shape we are trying to detect. The material properties will be described by two piecewise constant functions defined on $\mathcal{R}$ which, in general, can attain complex values:

$$\mu(\mathbf{x}) = \begin{cases} \mu_e & \mathbf{x} \in \mathcal{R} \setminus \overline{\mathcal{D}} \\ \mu_i & \mathbf{x} \in \mathcal{D} \end{cases}, \quad \lambda(\mathbf{x}) = \begin{cases} \lambda_e & \mathbf{x} \in \mathcal{R} \setminus \overline{\mathcal{D}} \\ \lambda_i & \mathbf{x} \in \mathcal{D} \end{cases}.$$

They are related to the magnetic permeability and the electrical permittivity in electromagnetic scattering, whereas in the thermal case the properties involved are the thermal inertia and the thermal conductivity of the materials.

If we denote by $\varphi : \mathcal{R} \to \mathbb{C}$ the complex amplitude of the electric field or the temperature at each point of the domain, then φ is the solution to the problem:

$$\begin{cases} \mu_e \Delta\varphi + \lambda_e \varphi = 0 & \text{in } \mathcal{R} \setminus \overline{\mathcal{D}} \\ \mu_i \Delta\varphi + \lambda_i \varphi = 0 & \text{in } \mathcal{D} \\ \varphi^+ = \varphi^- & \text{on } \partial\mathcal{D} \\ \mu_e \nabla\varphi^+ \cdot \mathbf{n} = \mu_i \nabla\varphi^- \cdot \mathbf{n} & \text{on } \partial\mathcal{D} \\ B(\varphi) = g & \text{on } \partial\mathcal{R} \end{cases}, \tag{2.1}$$

where $\mathbf{n}$ is the outward normal vector and the superscripts $^+$ and $^-$ refer to the limits when approaching $\partial\mathcal{D}$ from outside and inside respectively.

The operator B denotes a boundary value condition which will be different in each case, g being the 'forcing' term for the equation, that is, the incident field in the electromagnetic scattering problem and the heat flux coming from the lamp in the thermographic inspection problem.

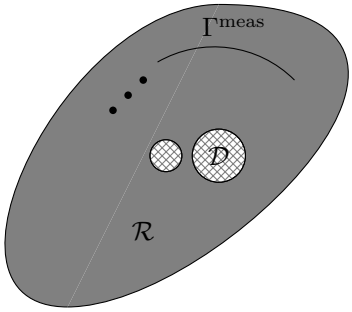


Fig. 1 Sketch of the setup for the abstract model. The set Γ^{meas} is not restricted to be connected and can consist of discrete points.

Given a set of points $\Gamma^{\text{meas}} \subset \overline{\mathcal{R}}$ where the variable φ is going to be measured (see Fig. 1 for a graphical illustration), we will denote by φ^{meas} the measured data. Note that, as neither the measuring process has absolute precision, nor the physical model is an exact representation of the real process, it will not be true, in general, that $\varphi^{\text{meas}} = \varphi_{|\Gamma^{\text{meas}}}$, that is, φ^{meas} does not coincide with the restriction of φ to Γ^{meas}.

In order to solve this kind of inverse problems, the aforementioned equality is usually relaxed to a minimization problem. If we define the functional:

$$j(\varphi) := \frac{1}{2} \int_{\Gamma^{\text{meas}}} |\varphi(\mathbf{x}) - \varphi^{\text{meas}}(\mathbf{x})|^2 \, d\gamma(\mathbf{x}), \tag{2.2}$$

and denote by φ_Ω the solution to problem (2.1) when setting $\mathcal{D} = \Omega$, then we can reformulate the original problem as a minimization problem where the aim is to find the set $\Omega \subset \mathcal{R}$ such that $j(\varphi_\Omega)$ is minimum, i.e.:

Find the domain $\Omega^* = \arg\min_{\Omega \subset \mathcal{R}} j(\varphi_\Omega)$ *such that* φ_Ω *is the solution to problem (2.1).*

Finding the minimizer Ω^* usually requires iterative methods which are computationally expensive. With our method we will find an approximation of the minimizer by a one step algorithm based on the concept of the topological derivative, which we describe in the following section.

3. Topological derivative

Given a shape functional $\mathcal{J}$, that is a rule that associates a real number to each shape, and an initial domain Ω the topological derivative, as presented in [10, 11] is a non-zero and finite function $\text{TD} : \Omega \to \mathbb{R}$ defined in such a way that the asymptotic expansion:

$$\mathcal{J}\left(\Omega \setminus \overline{\text{B}_\epsilon(\mathbf{x})}\right) = \mathcal{J}(\Omega) + \text{TD}(\mathbf{x}) f(\epsilon) + o(f(\epsilon)),$$

holds at every point $\mathbf{x}$ as $\epsilon \to 0^+$. The function f is a non-negative monotonically increasing function such that $f(0) = 0$, which acts as a measure of the size of the ball $\mathrm{B}_\epsilon(\mathbf{x})$.

That means that, at every point $\mathbf{x} \in \Omega$, the topological derivative is a measure of the sensitivity of the shape functional $\mathcal{J}$ to the inclusion of an infinitesimal hole. If $\mathrm{TD}(\mathbf{x})$ is large and negative it means that, at least to first order, if we consider a small perturbation centered at $\mathbf{x}$, the value of the functional evaluated in the perturbed domain would decrease with respect to the original one.

If we consider the shape functional:

$$\mathcal{J}\left(\mathcal{R} \setminus \overline{\mathrm{B}_\epsilon(\mathbf{x})}\right) \coloneqq j\left(\varphi_{\mathrm{B}_\epsilon(\mathbf{x})}\right), \tag{3.1}$$

where j is the functional defined at (2.2), then $\mathcal{J}$ will measure how far are the experimental measurements to the ones we would obtain if the scatterer/cavity were a ball of radius ϵ centered at the point $\mathbf{x}$.

By computing the topological derivative of the functional (3.1) and looking for the points where it attains its largest negative values, we will find points where, at least to first approximation, locating a small scatterer/cavity in the domain $\mathcal{R}$ makes the simulated measurements be closer to the experimental ones.

With this in mind, we will propose as reconstructed domains, the family of sets

$$\mathcal{D}_\lambda^{\mathrm{app}} = \left\{\mathbf{x} \in \mathcal{R} \,:\, \mathrm{TD}(\mathbf{x}) \leq \lambda \min_{\mathbf{y}\in\mathcal{R}} \mathrm{TD}(\mathbf{y}) \,, 0 < \lambda < 1\right\}, \tag{3.2}$$

where λ is a parameter that controls how conservative we are about considering a point as part of the object/cavity.

The topological derivative of this kind of shape functional can be computed in a very fast way. The abstract problem (2.1) when no object is present, admits a variational formulation: find $\varphi \in H^1(\mathcal{R})$ such that:

$$a(\varphi, \psi) = \ell(\psi) \quad \forall \psi \in H^1(\mathcal{R}),$$

where $a : H^1(\mathcal{R}) \times H^1(\mathcal{R}) \to \mathbb{C}$ is a bilinear form, $\ell : H^1(\mathcal{R}) \to \mathbb{C}$ is a linear form, and $H^1(\mathcal{R})$ is the Sobolev space of order 1. As shown in [3], the topological derivative can be expressed as:

$$\mathrm{TD}(\mathbf{x}) = \mathfrak{Re}\left(\left(2\mu_\mathrm{e}\frac{\mu_\mathrm{e} - \mu_\mathrm{i}}{\mu_\mathrm{e} + \mu_\mathrm{i}}\right)\nabla U(\mathbf{x}) \cdot \overline{\nabla V(\mathbf{x})} + (\lambda_\mathrm{i} - \lambda_\mathrm{e})\, U(\mathbf{x})\, \overline{V(\mathbf{x})}\right),$$

where U is the state field, which solves

$$a(U, \psi) = \ell(\psi) \quad \forall \psi \in H^1(\mathcal{R}),$$

that is, the state field corresponds to the solution when no scatterer or cavity is present. V is the adjoint field, and it is the solution to

$$a(\psi, V) = \delta j(U)(\psi) \quad \forall \psi \in H^1(\mathcal{R}),$$

where $\delta j(U)$ is the functional derivative of (2.2) at function U:

$$\delta j(U)(\psi) \coloneqq \mathfrak{Re}\left(\int_{\Gamma^{\mathrm{meas}}} (U(\mathbf{x}) - \varphi^{\mathrm{meas}}(\mathbf{x}))\, \overline{\psi(\mathbf{x})}\, \mathrm{d}\gamma(\mathbf{x})\right). \tag{3.3}$$

We can observe that the difference between the measured data and the one corresponding to a domain without any object or cavity acts as a forcing term in the adjoint problem.

In the following two sections we present the results obtained for the electromagnetic scattering problem as well as for thermographic inspection.

4. Electromagnetic scattering

The two-dimensional Fresnel database [1] contains experimental measurements of the electromagnetic scattering by different objects at several frequencies. It was developed by the Fresnel Institute of Marseille for a special issue of the journal Inverse Problems with the purpose of allowing different research groups to test or benchmark their inversion algorithms against common data.

In each experiment, objects are placed in the center of an an-echoic chamber and are irradiated by electromagnetic waves. The emitting antenna is located at $R_\mathrm{E} = 0.72\,\mathrm{m}$ from the center and we will denote its position by $\mathbf{x}^\mathrm{E}$. The total field is measured, both with the presence of the target and without it, at $N_\mathrm{R} = 49$ positions in a circumference of radius $R_\mathrm{meas} = 0.76\,\mathrm{m}$ and whose angle with respect to the emitting antenna are linearly spaced between 60° and 300°. We will look for the targets in the inspection zone $\{(x, y) \in \mathbb{R}^2 \;\; -L < x < L;\, -L < y < L\}$ with $L = 0.1\,\mathrm{m}$. In Fig. 2 an schematic of the experimental setup is shown compared with the size of the inspection zone, as well as one of the targets compared with the inspection zone.

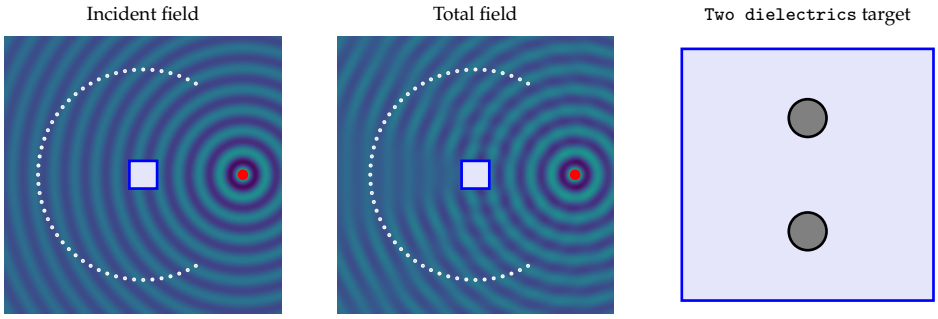


Fig. 2 Sketch of the experimental setup. The emitting antena is the red point, whereas the receiving antennas are the white ones. The inspection zone is the lavander square. The incident field (measured without the presence of the objects) and the total field (corresponding to measurements in presence of the objects) are shown in the first two plots. The true objects and their location inside the inspection zone are illustrated in the third plot.

For a fixed frequency the aforementioned measuring process is repeated $N_{\mathrm{E}} = 36$ times, rotating the target 10° each time. However, we will assume that the target is fixed, and it is the whole set of antennas which rotates around it, changing the emitting direction. After the target is rotated the full 360°, the process is repeated for the next frequency in the dataset.

In this paper we will only present the results obtained for a target consisting in two cylinders made of dielectric material (see figure 2 for a sketch of its cross section), however we have studied more cases in [2]. Taking into account that all the targets are vertical cylinders with constant section and that both the emitting antenna as well as the receiving ones are vertically polarized, the vertical component of the electric field $E : \mathbb{R}^2 \to \mathbb{C}$ must solve the problem:

$$\begin{cases} \Delta E + k^2 E = 0 & \text{in } \mathbb{R}^2 \setminus \overline{\mathcal{D}} \\ \Delta E + k^2 \varepsilon_{\mathrm{r}} E = 0 & \text{in } \mathcal{D} \\ E^+ = E^- & \text{on } \partial\mathcal{D} \\ \nabla E^+ \cdot \mathbf{n} = \nabla E^- \cdot \mathbf{n} & \text{on } \partial\mathcal{D} \\ \lim_{r\to\infty} \sqrt{r} \left(\frac{\partial}{\partial r} \left(E - E^{\mathrm{inc}} \right) - ik \left(E - E^{\mathrm{inc}} \right) \right) = 0 \end{cases}, \tag{4.1}$$

where E^{inc} denotes the incident field, that is, the one that would appear if no object were present (see left image of figure 2). The dielectric objects are modelled as having a different electrical permittivity ε_{r} than that of the surrounding air, but same magnetic permeability. The wavenumber in the background is $k = \frac{2\pi\nu}{\mathfrak{c}}$, ν being the frequency and $\mathfrak{c}$ the speed of light. Finally, as the experiments are performed in an an-echoic chamber, the problem can be posed in $\mathcal{R} = \mathbb{R}^2$, so the boundary conditions become a radiation condition at infinity, known as Sommerfeld's radiation condition.

As Γ^{meas} consists of a finite set of isolated points, the misfit functional we use is a sum rather than an integral:

$$\mathcal{J}\left(\mathbb{R}^2 \setminus \overline{\Omega}\right) = \frac{1}{2} \sum_{n=1}^{N^{\mathrm{meas}}} \left| E_\Omega\left(\mathbf{x}_n^{\mathrm{meas}}\right) - E_n^{\mathrm{meas}} \right|^2 . \tag{4.2}$$

If we compare this model with the abstract model (2.1) we observe that the topological derivative of (4.2) can be computed as:

$$\mathrm{TD}\left(\mathbf{x}\right) = \left(\varepsilon_{\mathrm{r}} - 1\right) \mathfrak{Re}\left(U\left(\mathbf{x}\right) \overline{V\left(\mathbf{x}\right)} \right) \tag{4.3}$$

where the state field $U : \mathbb{R}^2 \to \mathbb{C}$ is the solution to the problem:

$$\begin{cases} \Delta U + k^2 U = 0, & \text{in } \mathbb{R}^2 \\ \lim_{r\to\infty} \sqrt{r} \left(\frac{\partial (U - E_{\mathrm{inc}})}{\partial r} - ik \left(U - E_{\mathrm{inc}} \right) \right) = 0, \end{cases} \tag{4.4}$$

and the adjoint field $V : \mathbb{R}^2 \to \mathbb{C}$ is the solution to the problem:

$$\begin{cases} \Delta V + k^2 V = \sum\limits_{n=1}^{N_{\mathrm{meas}}} \left(U - E_n^{\mathrm{meas}} \right) \delta_{\mathbf{x}_n^{\mathrm{meas}}} & \text{in } \mathbb{R}^2 \\ \lim_{r\to\infty} \sqrt{r} \left(\frac{\partial V}{\partial r} + ikV \right) = 0. \end{cases} . \tag{4.5}$$

On one hand, the adjoint field can be expressed as a combination of fundamental solutions:

$$V(\mathbf{x}) = \sum_{r=1}^{N_{\text{meas}}} \left(U - E_n^{\text{meas}}\right) H_0^{(2)}\left(k\left|\mathbf{x} - \mathbf{x}_r^{\text{meas}}\right|\right),$$

where $H_0^{(2)}$ denotes the Hankel function of second kind and zero order. On the other hand, the state field is the incident field and we can approximate its values in the inspection zone by fitting a fundamental solution to the data measured in front of the emitting antenna:

$$U(\mathbf{x}) = \frac{E_{\text{front}}^{\text{inc}}}{H_0^{(1)}\left(k\left(R_{\text{E}} + R_{\text{meas}}\right)\right)} H_0^{(1)}\left(k\left|\mathbf{x} - \mathbf{x}_E\right|\right),$$

where $E_{\text{front}}^{\text{inc}}$ denotes the measure of the incident field took at the receiving antenna which is right in front of the emitting one.

In figure 3 we show the topological derivative field for different frequencies. We can observe that in general we obtain very good reconstructions. The only reconstruction with poor quality is the one at the lowest frequency in the dataset (1 GHz) in which we cannot distinguish the number of scatterers. It is important to note that at this frequency, the wavelength would be approximately as large as 4 times the side of the inspection zone. What is

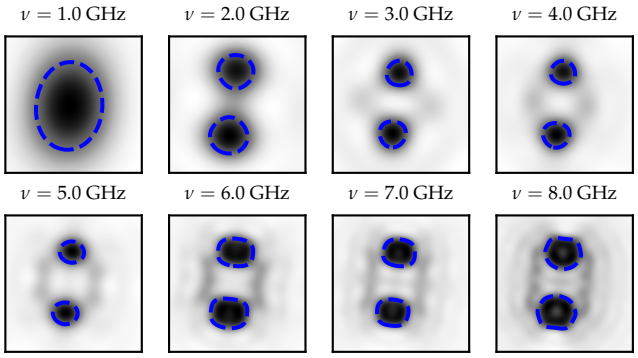


Fig. 3 Single frequency reconstructions. The topological derivative field is plotted in gray scale. The largest negative values correspond with dark colors. Superimposed is the contour of the approximated domain $\mathcal{D}_\lambda^{\text{app}}$ as defined in (3.2) corresponding to $\lambda = 0.7$.

perhaps even more impressive is that the reconstructions with fewer emitting directions are also very robust as can be seen in figure 4.

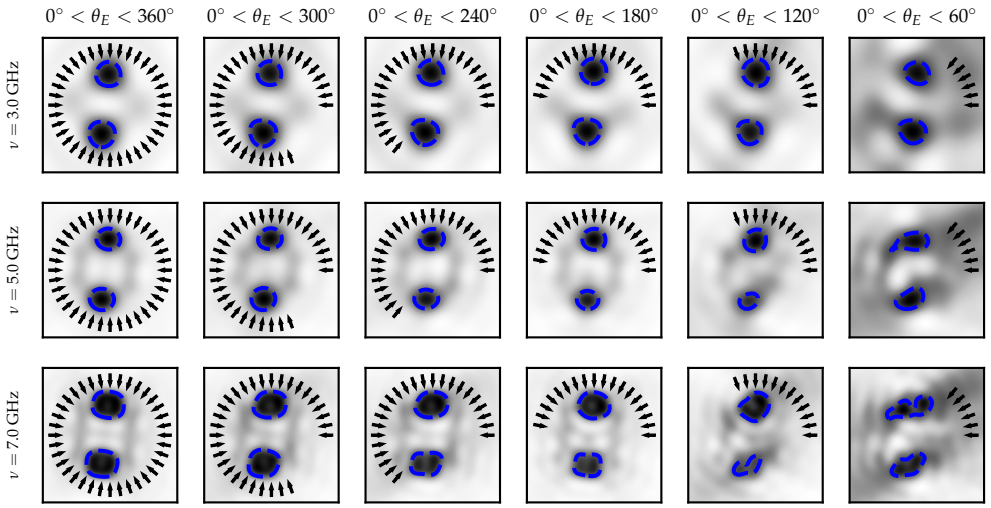


Fig. 4 Reconstructions with limited angle of emission for three different frequencies. The emitting directions are shown as black arrows.

5. Thermographic inspection

In [8,9] we tested a very similar algorithm in a problem consisting in the determination of the shape and number of defects or cracks in a metallic plate from a thermogram of one of its sides that was heated by an infrared lamp.

In this paper we will present a simplified model where the plate is considered to be two-dimensional. The region $\mathcal{R}$ occupied by the plate will be:

$$\mathcal{R} = \left\{(x, y) \in \mathbb{R}^2 \,:\, 0 < x < L_x,\, 0 < y < L_y\right\},$$

where $L_x = 0.01$ m and $L_y = 1$ m so it can be considered a thin plate. The boundary of the plate will be separated into three sets:

$$\Gamma_{\text{front}} \coloneqq \left\{(x, y) \in \mathbb{R}^2 \,:\, x = 0,\, 0 < y < L_y\right\}, \quad \Gamma_{\text{back}} \coloneqq \left\{(x, y) \in \mathbb{R}^2 \,:\, x = L_x,\, 0 < y < L_y\right\},$$

and

$$\Gamma_{\text{sides}} \coloneqq \left\{(x, y) \in \mathbb{R}^2 \,:\, 0 < x < L_x,\, y = 0\right\} \cup \left\{(x, y) \in \mathbb{R}^2 \,:\, 0 < x < L_x,\, y = L_y\right\}.$$

The plate will be heated from an infrared lamp situated at $\mathbf{x}_l$, on the same side as Γ_{front} see (Fig. 5) i.e. $\Gamma_{\text{meas}} = \Gamma_{\text{front}}$. We model the lamp as an isotropic emitter, so the heat absorbed by the plate would be:

$$q_{\text{lamp}}(\mathbf{x}, t) = \alpha \frac{I(t)}{2\pi} \frac{\mathbf{x} \cdot \mathbf{i}}{(\mathbf{x} - \mathbf{x}_l) \cdot (\mathbf{x} - \mathbf{x}_l)}, \quad \mathbf{x} \in \Gamma_{\text{front}},$$

where I is the intensity of the lamp, α is the surface absorptance of the plate and $\mathbf{i}$ denotes the unit vector along the x-axis.

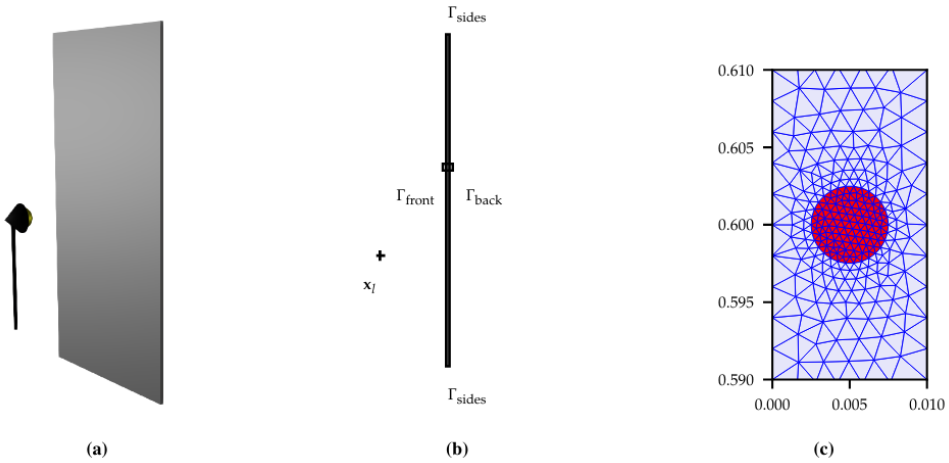


Fig. 5 Sketch of the thermal set up. (a) Three dimensional setup of the experiment. (b) Two-dimensional domain simplification. (c) Zoom of the computational mesh around the defect.

If the intensity of the lamp varies harmonically in time with frequency ω, then the temperature also varies harmonically at the same frequency, and its complex amplitude $T : \mathcal{R} \to \mathbb{C}$ must solve the problem:

$$\begin{cases} \kappa_{\text{e}} \Delta T + i\omega\rho_{\text{e}} c_{\text{e}} T = 0 & \text{in } \mathcal{R} \setminus \overline{\mathcal{D}} \\ \kappa_{\text{i}} \Delta T + i\omega\rho_{\text{i}} c_{\text{i}} T = 0 & \text{in } \mathcal{D} \\ T^+ = T^- & \text{on } \partial\mathcal{D} \\ -\kappa_{\text{e}} \nabla T^+ \cdot \mathbf{n} = -\kappa_{\text{i}} \nabla T^- \cdot \mathbf{n} & \text{on } \partial\mathcal{D} \\ -\kappa_{\text{e}} \nabla T \cdot \mathbf{n} = 0 & \text{on } \Gamma_{\text{sides}} \\ -\kappa_{\text{e}} \nabla T \cdot \mathbf{n} = h_{\text{eff}} T - Q_{\text{lamp}} & \text{on } \Gamma_{\text{front}} \\ -\kappa_{\text{e}} \nabla T \cdot \mathbf{n} = h_{\text{eff}} T & \text{on } \Gamma_{\text{back}} \end{cases}, \tag{5.1}$$

where Q_{lamp} denotes the complex amplitude of the absorbed heat. The coefficient h_{reff} takes into account the heat transfer between the plate and the surrounding air both via natural convection as well as infrared radiation.

A thermogram is a picture where each pixel color corresponds to the temperature of that point in the image. Hence, in this case the misfit functional can be expressed as:

$$\mathcal{J}\left(\mathcal{R} \setminus \overline{\Omega}\right) \coloneqq \frac{1}{2} \int_0^{L_y} |T_\Omega(\mathbf{x}) - T^{\text{meas}}(\mathbf{x})|^2 \, \mathrm{d}y. \tag{5.2}$$

Comparing with the abstract formulation (2.1) we observe that now both of the coefficients, namely the thermal conductivity κ, as well as the thermal inertia ρc, have different values inside the defect and in the rest of the plate.

The topological derivative of (5.2) can be expressed as

$$\mathrm{TD}\,(\mathbf{x}) = \mathfrak{Re}\left(2\kappa_\mathrm{e}\frac{\kappa_\mathrm{e}-\kappa_\mathrm{i}}{\kappa_\mathrm{e}+\kappa_\mathrm{i}}\nabla U\,(\mathbf{x})\cdot\overline{\nabla V\,(\mathbf{x})} + i\omega\left(\rho_\mathrm{i}c_\mathrm{i}-\rho_\mathrm{e}c_\mathrm{e}\right)U\,(\mathbf{x})\,\overline{V\,(\mathbf{x})}\right) \tag{5.3}$$

where the state field $U : \mathcal{R} \to \mathbb{C}$ is the solution to the problem:

$$\begin{cases} \kappa_\mathrm{e}\Delta U + i\omega\rho_\mathrm{e}c_\mathrm{e}U = 0 & \text{in } \mathcal{R} \\ -\kappa_\mathrm{e}\nabla U\cdot\mathbf{n} = 0 & \text{on } \Gamma_\mathrm{sides} \\ -\kappa_\mathrm{e}\nabla U\cdot\mathbf{n} = h_\mathrm{eff}U - Q_\mathrm{lamp} & \text{on } \Gamma_\mathrm{front} \\ -\kappa_\mathrm{e}\nabla U\cdot\mathbf{n} = h_\mathrm{eff}U & \text{on } \Gamma_\mathrm{back} \end{cases}$$

and the adjoint field $V : \mathcal{R} \to \mathbb{C}$ is the solution to the problem:

$$\begin{cases} \kappa_\mathrm{e}\Delta V - i\omega\rho_\mathrm{e}c_\mathrm{e}V = 0 & \text{in } \mathcal{R} \\ -\kappa_\mathrm{e}\nabla V\cdot\mathbf{n} = 0 & \text{on } \Gamma_\mathrm{sides} \\ -\kappa_\mathrm{e}\nabla V\cdot\mathbf{n} = h_\mathrm{eff}V - (U - T^\mathrm{meas}) & \text{on } \Gamma_\mathrm{front} \\ -\kappa_\mathrm{e}\nabla V\cdot\mathbf{n} = h_\mathrm{eff}V & \text{on } \Gamma_\mathrm{back} \end{cases}$$

In this case, we had no experimental data, and the forward problem was solved by a finite element method (FEM) approximation to generate numerical data simulating thermograms. The state and adjoint field problems were also solved by the FEM.

In Fig. 6 we show a plot of a thermogram as well as the topological derivative of the corresponding experiment, restricted to the side Γ_front. We can see that the topological derivative clearly marks the height (y-coordinate) at which the defect is located.

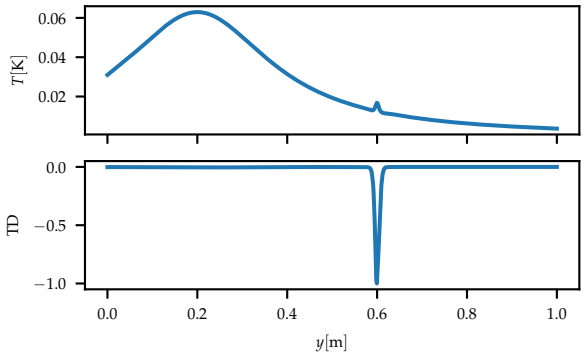


Fig. 6 Thermogram and topological derivative for a lamp at $\mathbf{x}_l = (-0.2, 0.2)$ and $\omega = 2.5$ GHz.

The topological derivative field in a region surrounding the defect is plotted in Fig. 7. As we can see, it is not able to show the correct depth of the defect (i.e. the x-coordinate of the center of the circle). This behaviour does not depend on the location or the frequency of the emitting lamp, and hence, cannot be solved by using a linear combination of topological derivatives corresponding to several experiments.

6. Conclusions and future work

The topological derivative has proven to be a very efficient and reliable indicator for object detection in different scenarios.

When tested against electromagnetic experimental data it was able to recover the shapes with great accuracy without a priori information on the number of objects or their size. Albeit being a one step method, the quality of the reconstructions was comparable to the iterative methods tested in the special session [1]. Furthermore, we have shown its accuracy against limited angle emissions, being able to approximate the shape with less than a quarter of the data.

In the thermal case it was able to recover the position of the defect in the y-direction and also gave a good estimation of the size of the defect, but has no information on its depth. In [9] it was shown that in the full three-dimensional case it is able to recover the projection of the shape on the measuring surface. This difference

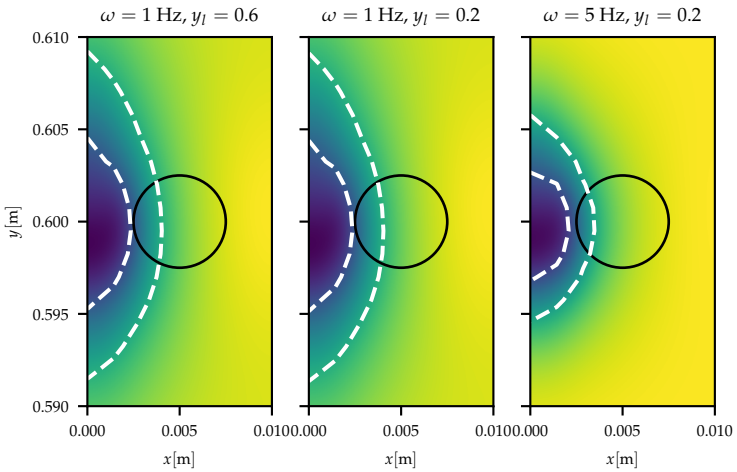


Fig. 7 Topological derivative field for different experiments.

in behaviour with respect to the electromagnetic case is due to the experimental thermogram being "closer" to the simulated thermogram one would obtain for a perturbation of radius $\epsilon \to 0$ centered at a point in the surface than for a thermogram corresponding to a perturbation of radius $\epsilon \to 0$ but centered at the correct depth.

If the topological derivative is going to be used for testing very thin plates, this is not a real problem, as we mostly need finding the y-coordinate of the defect. It would be interesting to study if the second order topological derivative [5] also promotes minima towards the surface of the plate or if it is able to recover the correct depth.

Acknowledgements

This research was founded by Spanish FEDER/MICINN-AEI grants MTM2014-56948-C2-1-P and MTM2017-84446-C2-1-R and the MCIN/AE/doi 10.13039/501100011033 grant PID2020-114173RB-I00.

References

[1] K. Belkebir and M. Saillard. Special section: Testing inversion algorithms against experimental data. *Inverse Problems*, 17(6):1565–1571, 2001.

[2] A. Carpio, M. Pena, and M. L. Rapún. Processing the 2D and 3D Fresnel experimental databases via topological derivative methods. *Inverse Problems*, 37(10):105012, 2021.

[3] A. Carpio and M.-L. Rapún. Topological Derivatives for Shape Reconstruction. In Luis L. Bonilla, editor, *Inverse Problems and Imaging: Lectures given at the C.I.M.E. Summer School held in Martina Franca, Italy September 15–21, 2002*, pages 85–133. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[4] A. Carpio and M.-L. Rapún. Domain reconstruction using photothermal techniques. *Journal of Computational Physics*, 227(17):8083–8106, 2008.

[5] J. R. de Faria, A. Novotny, R. Feijóo, E. Taroco, and C. Padra. Second order topological sensitivity analysis. *International Journal of Solids and Structures*, 44(14-15):4958–4977, 2007.

[6] G. R. Feijoo. A new method in inverse scattering based on the topological derivative. *Inverse Problems*, 20(6):1819–1840, 2004.

[7] M. Masmoudi, J. Pommier, and B. Samet. The topological asymptotic expansion for the Maxwell equations and some applications. *Inverse Problems*, 21(2):547–564, 2005.

[8] M. Pena and M.-L. Rapún. Detecting Damage in Thin Plates by Processing Infrared Thermographic Data with Topological Derivatives. *Advances in Mathematical Physics*, 2019:1–18, 2019.

[9] M. Pena and M.-L. Rapún. Application of the topological derivative to post-processing infrared time-harmonic thermograms for defect detection. *Journal of Mathematics in Industry*, 10(1):4, 2020.

[10] J. Sokolowski and A. Zochowski. On the Topological Derivative in Shape Optimization. *SIAM Journal on Control and Optimization*, 37(4):1251–1272, January 1999.

[11] J. Sokolowski and J.-P. Zolesio. Introduction to shape optimization. In *Introduction to Shape Optimization: Shape Sensitivity Analysis*, pages 5–12. Springer Berlin Heidelberg, Berlin, Heidelberg, 1992.

Data-driven Reduced Order Methods. Applications to transition spaces in buildings

Soledad Fernández-García[1], **Macarena Gómez-Mármol**[1], **Samuele Rubino**[1]
Dpto. Ecuaciones Diferenciales y Análisis Numérico. Universidad de Sevilla, Spain

Abstract

We develop a machine learning technique based on two different methods: Reduced-Order and interpolation methods. In particular, we use first Support Vector Regression method, and second an interpolation method on a data set of scattered data. The developped strategy is applied to predict thermal performances in courtyards. These predictions can help to improve the courtyard design, and subsequently, to get better energy performance of buildings.

1. Introduction

Data-driven mathematical methods are becoming a very important tool for modelling and understanding problems coming from all branches of the science and engineering. In general, the main goal of these methods is to capture with a small subset of the full and usually, high-dimensional state space, the most important properties of the evolution of a physical system.

More specifically, the main idea behind Reduced-Order Models (ROMs) is to consider time snapshots of the data (which could come, for instance, from a Partial Differential Equation solution), and build a lower dimensional system, which is able to catch the main features of the original model. A review of Data-driven methods for reduced order modelling can be consulted in [2].

On the other hand, Support Vector Machines (SVMs) were introduced in the 90s by Vapnik and his collaborators [3], in the framework of statistical learning theory. Although originally SVMs were thought to solve binary classification problems, they are currently used to solve various types of problems, for instance, regression problems [8], which we focus on, and can be considered a special case of ROMs methods.

The idea behind SVM methods, is to select a separation hyperplane, which is equidistant to the closest examples of each class, in order to achieve the so-called maximum margin on each side of the hyperplane. Furthermore, when defining the hyperplane, only those examples whose distance from the hyperplane is the margin distance are considered. These examples are called vectors support. More specifically, we focus on this work on Support Vector Regression (SVR) method, which is briefly presented in the folllowing section.

Although these methods could be used with data coming from very different fields, we have chosen to focus on prediction of temperature in courtyards. As it is well known, courtyards are an effective passive strategy for improving the energy performance of buildings. Here, we consider the prediction of courtyard thermal performances based on machine learning techniques, as we developped in [4]. In particular, in a first step, we use the SVR method based on monitored data in every hour of one week for different periods in various courtyards located in Badajoz, Córdoba and Seville (Spain), to predict the temperature inside the courtyard in the whole week. Subsequently, in a second step, based on the climate zoning and aspect ratios of a courtyard, we have predicted its courtyard thermal performances by interpolating the predicted data provided by the SVR method.

The rest of the chapter is outlined as follows. In Section 2, we present the SVR method. In Section 3, we predict the temperature of a courtyard by using its climate zone, period of the year and aspect ratios. Finally, Section 4 is devoted to the conclusions of this work.

2. Support Vector Regression method

We attempt to predict the value of a function $f\colon \mathbb{R}^N \to \mathbb{R}$ using some information related to it. In particular, to find the function f, we use the m-collection of experimental data associated to it. The idea of the SVR method [8] is to obtain a function such that for every sample (x_i, y_i), $i = 1, \dots, m$, it is satisfied that $|f(x_i) - y_i| \leq \varepsilon$, for some $\varepsilon > 0$ small. Concretely, given ε, γ and $C > 0$, the following optimization problem is considered,

$$\max \left\{ \frac{1}{2} \sum_{i,j=1}^{m} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \exp(-\gamma \|x_i - x_j\|^2) - \varepsilon \sum_{i=1}^{m} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{m} y_i (\alpha_i - \alpha_i^*) \right\},$$

$$\text{subject to } \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) = 0, \quad \text{for } \alpha_i, \alpha_i^* \in [0, C].$$

This problem can be solved, for instance, using the statistical software *R*. In particular, by using the *e1071* library [7], a software package designed to solve classification and regression problems using Support Vector Machines, which can be easily installed in *R*. The solution provides a possible candidate function,

$$f(x) = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) \exp(-\gamma \|x_i - x\|^2) + b,$$

where the constant $b \in \mathbb{R}$ can be computed by forcing the Karush-Kuhn-Tucker (KKT) condition [6]. The function $K(x, x') = \exp(-\gamma \|x - x'\|^2)$ is called radial basis kernel. It holds that $|f(x_i) - y_i| \leq \varepsilon$, for all $i = 1, \ldots, m$. The quality of function f depends on the choice of the parameters ε, γ and C.

3. Predicted temperature of a courtyard

As an instance of the Data Driven Reduced Order Methods, in this section we work on predictions of the value of the temperature inside a courtyard using its climate zone, period of the year and aspect ratios, as it was done more widely in [4]. To do that, we consider two stages.

In the first stage, we have worked with the data from 12 monitored courtyards in every hour of one week, from different periods of the year, in varios courtyard located in Badajoz, Córdoba and Seville (Spain), see Tab. 1.

Location	Courtyards
Badajoz	Juan Pablo II, Los Maestros, Manuel Gil, Suárez Somonte
Córdoba	Carlos Rubio, Pompeyos
Sevilla	Dr. Fleming, Hernán Cortés (2 and 3), IMUS, San Sebastián, UNED

Tab. 1 Location of monitored courtyards in this work.

For each courtyard and each week, we have used the SVR method to predict the value of the temperature inside the courtyard all along the week, by using some information related to it. More specifically, we attempt to predict the value of the temperature inside a courtyard considering the time, the outside courtyard temperature, the wind speed and direction. More specifically, we consider $x = (x^1, x^2, x^3, x^4)$, where

- x^1 = time (hours),
- x^2 = outside temperature (Celsius degrees),
- x^3 = wind speed (m/s),
- x^4 = wind direction (angle degrees),

and we look for a function $f : \mathbb{R}^4 \to \mathbb{R}$, such that $y = f(x)$ provides the temperature inside the courtyard.

In a second stage, using the library of predicted data obtained from the machine learning method, we have predicted the temperature inside a given courtyard, based on its climate zone, period of the year and aspect ratios.

We focus on each one of this two stages stage from now on. In the first stage, through the monitoring data and the SVR method, we have obtained a library of predicted temperatures inside various courtyards located in different cities of the south of Spain. In this second stage, by using this library, we predict the temperature inside a given courtyard.

We are going to use the concept Aspect Ratio, that we define subsequently. We consider two Aspect Ratios, the first one, ARI is defined as the relation between the width and the height, and the second one, $ARII$ is defined as the relation between the width and the length, as follows,

$$ARI = h_{max}/W \quad \text{and} \quad ARII = h_{max}/L,$$

where h_{max} is the maximum height, W is the width and L is the length of the courtyard.

Now that the concept of Aspect Ratio is defined, we proceed to predict the temperature inside a given courtyard in this way: We have classified the courtyards library into three different classes, depending on the range of temperatures of the patios and we have used an interpolation technique to predict the temperature inside a courtyard of the same class by using the aspect ratios data. In particular, we have classified the courtyards library into this three different classes:

C1. Temperature range: $(15°, 35°)$.

C2. Temperature range: $(20°, 40°)$.

C3. Temperature range: $(25°, 45°)$.

In Tab. 2 we classify the courtyards within these different classes. Note that some courtyards are in more than one class, because the temperature range in the courtyard has changed from one week to another. We denote by Wi the week number i of the measured data in the courtyard.

Temp. range class	Courtyards
C1	J. P. II, Los Maestros, M. Gil, S. Somonte
C2	C. Rubio (W3), H. Cortés (2 and 3), IMUS (W1,4), San Sebastián, UNED
C3	C. Rubio (W1,2), IMUS (W2,3), Pompeyos

Tab. 2 Classification of courtyards within temperature range class. We denote by Wi the week number i of the measured data in the courtyard.

Moreover, courtyard Dr. Fleming is between C2 and C3, so it has not been inlcluded in Tab. 2. Thus, for a given courtyard, we first estimate its range of temperature, classifying it into C1, C2 or C3.

Then, given the Aspect Ratios ARI and $ARII$ of some courtyards in the same class and their corresponding predicted temperatures through the SVR method, by an interpolation technique implemented in the scientific software MATLAB, we obtain a prediction of the temperature inside a courtyard of the same class. To do the interpolation, we have used the MATLAB function *scatteredInterpolant*, which perform interpolation on a 2-D data set of scattered data. In particular, it returns the interpolant F for the given data set, such that we can evaluate F at a set of query points in 2-D, to produce interpolated values $T_q = F(ARI_q, ARII_q)$, obtaining the temperature inside the courtyard T_q.

Also, a quantitative analysis has been carried out. On the one hand, we evaluated the relative error of the predicted temperature with respect to the monitored temperature in different discrete norms:

- $$L^1\,(\%) = \frac{\sum_{i=1}^{N} |T_{monit.} - T_{pred.}|(t_i)}{\sum_{i=1}^{N} T_{monit.}(t_i)} \cdot 100,$$
- $$L^2\,(\%) = \left[\frac{\sum_{i=1}^{N} (T_{monit.} - T_{pred.})^2(t_i)}{\sum_{i=1}^{N} T^2_{monit.}(t_i)} \right]^{1/2} \cdot 100,$$

where we have denoted by $T_{monit.}(t_i)$ (resp, $T_{pred.}(t_i)$) the monitored temperature (resp., the predicted temperature) at time $t_i, i = 1, \ldots, N$ (hours, $[h]$). Moreover, we evaluated the percentage in time for which the obtained absolute error within the predicted and the monitored temperature is less than or equal to a fixed tolerance $tol = 2^oC$. On the other hand, we computed the following statistical parameters: the correlation coefficient R, the Root Mean Square Error ($RMSE$) and the Mean Absolute Percentage Error ($MAPE$). The formulas for these parameters are:

- $$R = \frac{\sum_{i=1}^{N} (T_{monit.}(t_i) - \overline{T}_{monit.})(T_{pred.}(t_i) - \overline{T}_{pred.})}{\left[\sum_{i=1}^{N} (T_{monit.}(t_i) - \overline{T}_{monit.})^2 \sum_{i=1}^{N} (T_{pred.}(t_i) - \overline{T}_{pred.})^2 \right]^{1/2}},$$
- $$RMSE\,(^oC) = \left[\frac{\sum_{i=1}^{N} (T_{monit.} - T_{pred.})^2(t_i)}{N} \right]^{1/2},$$
- $$MAPE\,(\%) = \frac{1}{N} \sum_{i=1}^{N} \frac{|T_{monit.} - T_{pred.}|(t_i)}{T_{monit.}(t_i)} \cdot 100,$$

where in the formula for the correlation coefficient R we have denoted by $\overline{T}_{monit.}$ (resp, $\overline{T}_{pred.}$) the mean monitored temperature (resp., the mean predicted temperature). The values of the relative and absolute errors, and the statistical parameters are shown in tables 3 and 4 for the air temperature in each selected courtyard of each temperature range class.

Now, we show the predicted temperature in one courtyard of each temperature range class. We represent the predicted temperature in comparison to the monitored temperature inside the courtyard, as well as the outdoor temperature. For the class C1, we have considered the courtyard *Los Maestros*, located in Badajoz, (which has not been included in the courtyards C1 data base). The prediction is performed for the date 20th to 26th May 2018. The obtained results are represented in Fig. 1 and tables 3 and 4 (first row).

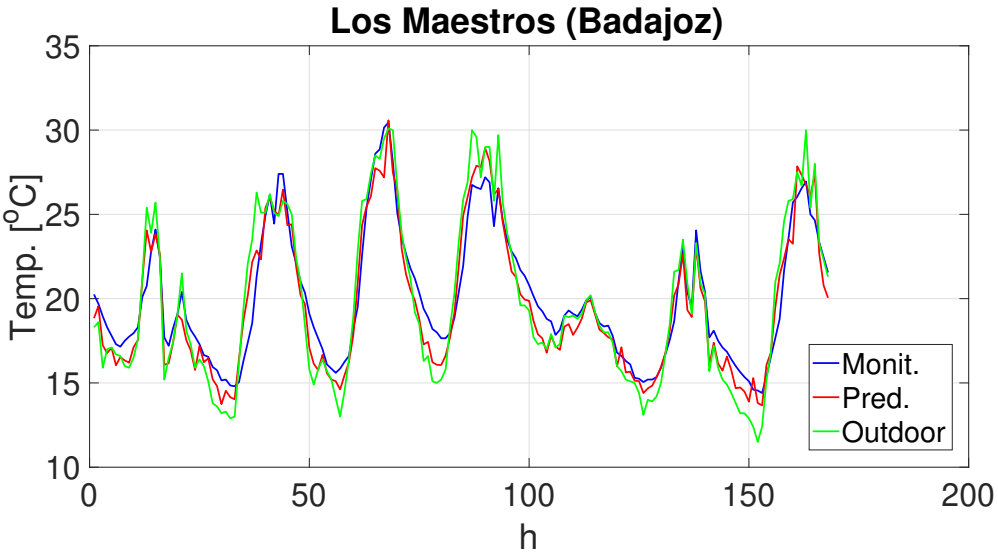


Fig. 1 Predicted temperature versus monitored and outdoor temperatures inside a C1 courtyard.

Temp. range class	L^1 (%)	L^2 (%)	Absolute error $\leq tol$ (%)
C1	5.09	6.10	91.67
C2	4.93	6.62	84.28
C3	3.50	4.31	89.88

Tab. 3 Relative and absolute errors for the air temperature in each selected courtyard of each temperature range class.

Temp. range class	R	$RMSE$ (^oC)	$MAPE$ (%)
C1	0.96	1.24	5.17
C2	0.88	1.62	4.82
C3	0.89	1.21	3.52

Tab. 4 Statistical parameters for the air temperature in each selected courtyard of each temperature range class.

Second, for the class C2 we have considered the courtyard *UNED*, located in Seville. The prediction is performed for the date 7th to 13th September 2018. The obtained results are represented in Fig. 2 and tables 3 and 4 (second row).

Finally, for the class C3 we have considered the courtyard *Carlos Rubio*, located in Córdoba. The prediction is performed for the date 26th July to 1st August 2017 (W1). The obtained results are represented in Fig. 3 and tables 3 and 4 (third row).

4. Conclusions

In this work, we have presented a prediction method based on machine learning techniques, and we have used it to predict courtyard thermal performances. The method is developped in two different stages. In a first step, we have used the SVR method based on provided data, to predict the temperature inside the courtyard in a whole week.

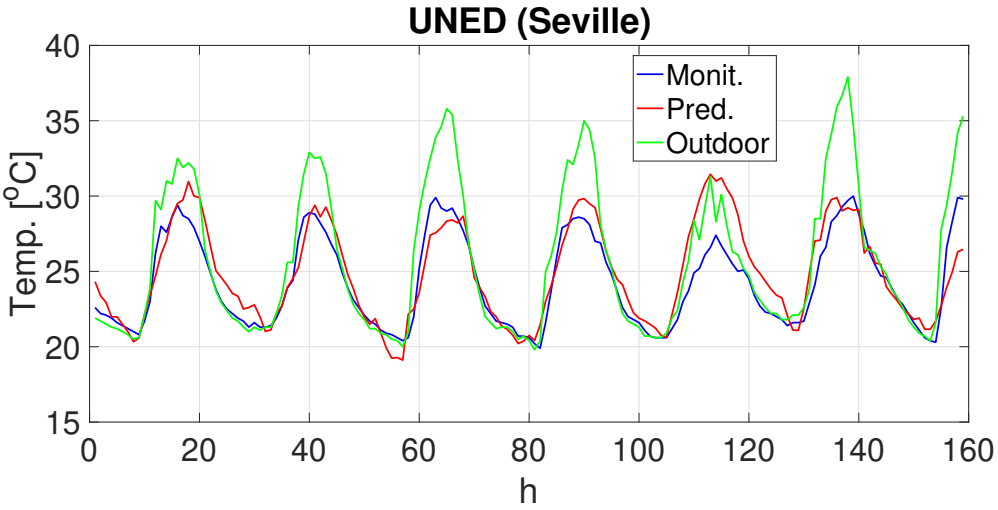


Fig. 2 Predicted temperature versus monitored and outdoor temperatures inside a C2 courtyard.

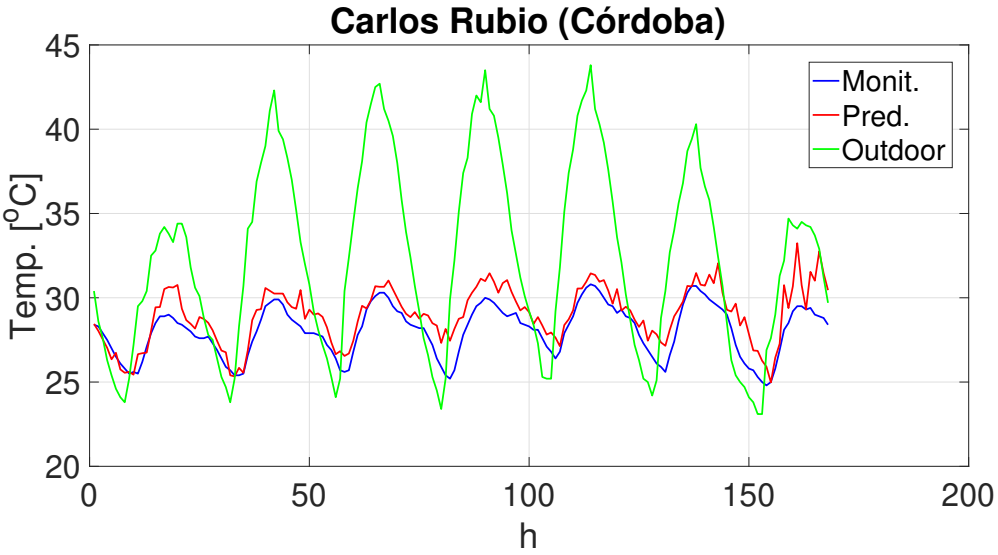


Fig. 3 Predicted temperature versus monitored and outdoor temperatures inside a C3 courtyard.

After that, based on the climate zoning and aspect ratios of a courtyard, we have predicted its courtyard thermal performances, by interpolating the predicted data provided by the SVR method. We proceed now to present briefly the quality of the obtained results.

On the one hand, the values for the relative errors in different discrete norms are around 5% and the percentage in time for which the obtained absolute error with restect to the monitored temperature is less than or equal to $tol = 2^oC$ is superior to 80%.

On the other hand, the values for the statistical parameters that indicate that the simulation is accurate are: $R \longrightarrow 1$, $RMSE \longrightarrow 0$, $MAPE \longrightarrow 0$, see [1,9]. The values of these parameters for the air temperature in the present courtyards for each simulation confirm that the used strategy is rather accurate. In particular, the correlation coefficient R is superior to 0.85 for all range classes. The $RMSE$ values are around $1.5^o\,C$ and the $MAPE$ values are around 5%.

We can see that the values of the computed statistical parameters are in a similar range than those obtained in [5] for a similar problem. There, the authors performed a very accurate courtyard thermal simulation based upon a Computational Fluid Dynamics (CFD) FreeFEM 3D model, which is much more computationally expensive than the machine learning technique SVR used in this work. In particular, the computation of one week temperature through the SVR method takes around one minute, while the CFD method takes around four minutes per one day of simulation. Thus, we have checked in this context that ROMs plus interpolation techniques are a powerfull tool to make predictions.

Acknowledgements

This research has been partially funded by Plan Estatal 2021-2023- Proyectos Investigación Orientada Grant PID2021-123153OB-C21. The authors want to thank Eduardo Díaz-Mellado, Carmen Galán-Marín and Carlos Rivera-Gómez, for the temperature data provided for this work.

References

[1] J. S. Armstrong and F. Collopy, *Error measures for generalizing about forecasting methods: Empirical comparisons*, Int. J. Forecast. **8**(1), 69–80, 1992.

[2] S. L. Brunton and J. N. Kutz, *Data-driven methods for reduced-order modeling*, Chapter 7, Volume 2- Snapshot-Based Methods and Algorithms, 2021.

[3] C. Cortes and V. Vapnik, *Support-vector networks*, Machine learning, **20**(3), 1995.

[4] E. Díaz-Mellado, S. Rubino, S. Fernández-García, M. Gómez-Mármol, C. Rivera-Gómez and C. Galán-Marín, *Applied Machine Learning Algorithms for Courtyards Thermal Patterns Accurate Prediction*, Mathematics, **9** 1142, 2021.

[5] V. P. López Cabeza, F. J. Carmona Molero, S. Rubino, C. Rivera Gómez, E. D. Fernández Nieto, C. Galán Marín and T. Chacón Rebollo, *Modelling of surface and inner wall temperatures in the analysis of courtyard thermal performances in Mediterranean climates*, Journal of Building Performance Simulation, **14**(2), 181–202, 2021.

[6] D. G. Luenberger and M. López Mateos, *Programación lineal y no lineal.* Number 90C05 LUEp. Addison-Wesley Iberoamericana, 1989.

[7] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel and F. Leisch, *e1071: Misc Functions of the Department of Statistics Probability Theory Group (Formerly: E1071)*, TU Wien. ***R*** package version 1.6-7, 2015.

[8] V. Vapnik, S. E. Golowich and A. Smola, *Support vector method for function approximation, regression estimation, and signal processing*, Advances in neural information processing systems. **9**, 1996.

[9] C. J. Willmott, *Some Comments on the Evaluation of Model Performance*, Bull. Am. Meteorol. Soc. **63**(11), 1309–1313, 1982.